

Women in signal processing

Edited by

Hagit Messer, Monica Bugallo, Maria Sabrina Greco
and Fauzia Ahmad

Published in

Frontiers in Signal Processing



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-83250-050-7
DOI 10.3389/978-2-83250-050-7

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Women in signal processing

Topic editors

Hagit Messer — Tel Aviv University, Israel

Monica Bugallo — Stony Brook University, United States

Maria Sabrina Greco — University of Pisa, Italy

Fauzia Ahmad — Temple University, United States

Citation

Messer, H., Bugallo, M., Greco, M. S., Ahmad, F., eds. (2022). *Women in signal processing*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83250-050-7

Table of contents

04	Editorial: Women in signal processing Hagit Messer
07	Precoded Cluster Hopping for Multibeam GEO Satellite Communication Systems Eva Lagunas, Mirza Golam Kibria, Hayder Al-Hraishawi, Nicola Maturo and Symeon Chatzinotas
18	Analysis of a 2D Representation for CPS Anomaly Detection in a Context-Based Security Framework Sara Baldoni, Marco Carli and Federica Battisti
29	Beamspace ESPRIT for mmWave Channel Sensing: Performance Analysis and Beamformer Design Sina Shahsavari, Pulak Sarangi and Piya Pal
45	VIPDA: A Visually Driven Point Cloud Denoising Algorithm Based on Anisotropic Point Cloud Filtering Tiziana Cattai, Alessandro Delfino, Gaetano Scarano and Stefania Colonnese
58	Multivariate Lipschitz Analysis of the Stability of Neural Networks Kavya Gupta, Fateh Kaakai, Beatrice Pesquet-Popescu, Jean-Christophe Pesquet and Fragkiskos D. Malliaros
77	Optimization of Network Throughput of Joint Radar Communication System Using Stochastic Geometry Shobha Sundar Ram, Shubhi Singhal and Gourab Ghatak
91	Small Object Detection and Tracking in Satellite Videos With Motion Informed-CNN and GM-PHD Filter Camilo Aguilar, Mathias Ortner and Josiane Zerubia
106	How Scalable Are Clade-Specific Marker K-Mer Based Hash Methods for Metagenomic Taxonomic Classification? Melissa Gray, Zhengqiao Zhao and Gail L. Rosen
118	Adaptive Discrete Motion Control for Mobile Relay Networks Spilios Evmorfos, Dionysios Kalogerias and Athina Petropulu
131	Bayesian Nonparametric Learning and Knowledge Transfer for Object Tracking Under Unknown Time-Varying Conditions Omar Alotaibi and Antonia Papandreou-Suppappola
144	Blind visual quality assessment of light field images based on distortion maps Sana Alamgeer and Mylène C. Q. Farias



OPEN ACCESS

EDITED AND REVIEWED BY
Fulvio Gini,
University of Pisa, Italy

*CORRESPONDENCE
Hagit Messer,
messer@eng.tau.ac.il

SPECIALTY SECTION
This article was submitted to Radar
Signal Processing,
a section of the journal
Frontiers in Signal Processing

RECEIVED 24 June 2022
ACCEPTED 18 July 2022
PUBLISHED 16 August 2022

CITATION
Messer H (2022), Editorial: Women in
signal processing.
Front. Sig. Proc. 2:977475.
doi: 10.3389/frsip.2022.977475

COPYRIGHT
© 2022 Messer. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Editorial: Women in signal processing

Hagit Messer*

School of Electrical Engineering, Tel Aviv University, Tel Aviv, Israel

KEYWORDS

women in science and engineering, signal processing, women in academia, science, engineering

Editorial on the Research Topic women in signal processing

One of the turn-points in my life was in the mid-90th, during the yearly major conference of the Signal Processing community, the IEEE international conference on acoustic, speech and signal processing (ICASSP). Women were always minorities in these meetings, and if one of them joined a chat in a social gathering, she were naturally considered as the wife of one of the men around. Being young and naïve then, I never saw it as an issue. However, at that specific meeting on 1995 I decided to join, for the first time, a social event, entitled “lunch for women in signal processing.” I found there a small but very diverse group of about 50 women from all around the world, and when each introduced herself, I had a very strong emotional reaction of a sisterhood. For the first time I felt at home in my professional community, and at that very specific moment I became active in the advancement of women in science and engineering, and in particular in my field, i.e., signal processing.

An essential question rises is about the quantity and the visibility of women in signal processing today. Such data is hard to trace, but fortunately, the IEEE keeps and publishes statistical records¹. These records show that while the overall share of women in the IEEE (including students) is still around 10%, in the signal processing society it is a bit but not much better, about 2,300 out of 19,000 (~12%). However, [Figure 1](#) shows a promising trend over the last decade: while the total number of women (non-students) in the IEEE signal processing society has increased by 45%, the number of women in higher-level grades (senior member and fellow) has doubled. Moreover, women take leadership positions in the IEEE signal processing society² with the current president Athina P. Petropulu and 11 out of its 23 board members being women³.

1 <https://mga.ieee.org/resources/annual-statistics>.

2 <https://signalprocessingsociety.org/our-story/board-governors>.

3 Note, however that in the EURASIP only 1 out of 8 board members is a women (<https://eurasip.org/organization/>).

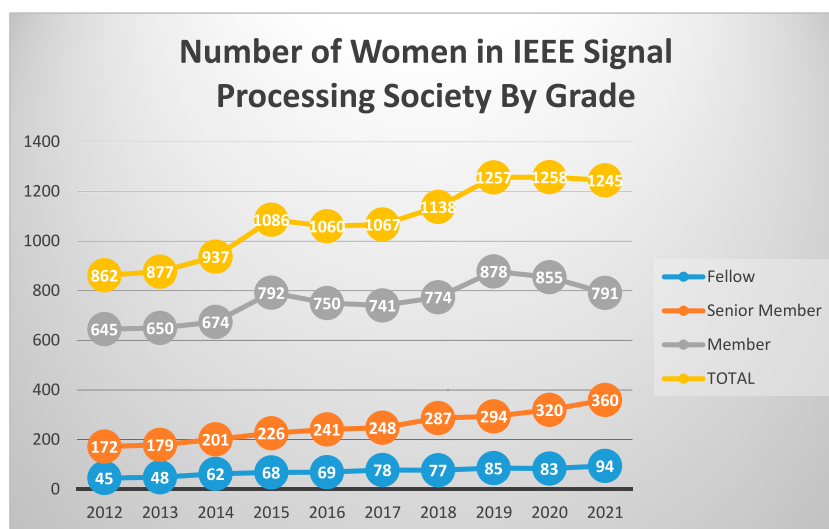


FIGURE 1

IEEE Signal Processing Society women members in all non-student grades over the decade of 2012–2021. Based on data published in the annual IEEE statistical reports.

With that in the background, I was happy to accept the role of an editor of the topic “*Women in signal processing*” for the journal *Frontiers in Signal Processing*. The first challenging task was to identify potential contributors, trying to open a gate for women outside the natural circle of collaborators and colleagues. Luckily, with the help of my co-editors [Monica Bugallo](#), [Maria Sabrina Greco](#) and [Fauzia Ahmad](#), we have identified 82 potential female contributors whom were directly approached, and managed to deliver this unique Research Topic of *Frontiers in Signal Processing*.

The 10 papers published under the topic of “*Women in signal processing*” came from senior or junior female researchers from North America, Europe and Asia. They cover a variety of topics in signal and image processing, involving learning techniques, integrating sensing and communication, satellite communication, security and more, presenting original research and methods:

- [Eva Lagunas et al.](#) from the University of Luxemburg contribute to GEO satellite communication and propose an efficient time–space illumination pattern design, where they determine the set of clusters that shall be illuminated simultaneously at each hopping event along with the dwelling time.
- [Sara Baldoni et al.](#) from Italy deal with security and propose a flexible context-based security framework by exploring two types of context: distributed and local.
- [Sina Shahsavari et al.](#) from University of California, San Diego present an error analysis for Angle of Arrival (AoA)

estimation in mmWave channel, a Research Topic relevant to 5G technologies and beyond.

- [Tiziana Cattai et al.](#) from Rome, Italy present an original Visually Driven Point cloud Denoising Algorithm (VIPDA) contributing for better digital representation of 3D surfaces;
- [Kavya Gupta et al.](#) from France present a stability analysis of fully connected neural networks allowing one to capture the influence of each input or group of inputs on the neural network stability.
- [Alotaibi and Suppappola](#) from Italy propose two methods for dealing with a primary source tracking a moving object under time-varying and unknown noise conditions;
- [Shobha Sundar Ram et al.](#) from Indraprastha Institute of Information Technology Delhi, India present contribution to the optimization of JRC—joint radar communication system;
- [Josiane Zerubia et al.](#) from France propose a track-by-detection approach to detect and track small moving targets by using a convolutional neural network and a Bayesian tracker;
- [Melissa Gray et al.](#) from Drexel University, United States study methods for metagenomic taxonomic classification, contributing to accurately identifying which microbes are present in a biological sample;
- [Athina Petropulu et al.](#) from Rutgers and Yale universities, United States, present methods for dealing with the problem of joint beamforming and discrete motion control for mobile relaying networks in dynamic channel environments;

These articles attract increasing attention from the relevant community all around the world, as indicated by *Women in signal processing* Frontiers Research Topic (frontiersin.org), and hopefully contribute to the visibility of women in signal processing.

Finally, I deeply thank my co-editors Monica Bugallo, Maria Sabrina Greco and Fauzia Ahmad and the frontiers staff for making it happens.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



Precoded Cluster Hopping for Multibeam GEO Satellite Communication Systems

Eva Lagunas*, Mirza Golam Kibria, Hayder Al-Hraishawi, Nicola Maturo and Symeon Chatzinotas

Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg, Luxembourg

OPEN ACCESS

Edited by:

Maria Sabrina Greco,
University of Pisa, Italy

Reviewed by:

Bokamoso Basutli,
Botswana International University of
Science and Technology, Botswana
Alessandro Guidotti,
University of Bologna, Italy

*Correspondence:

Eva Lagunas
eva.lagunas@uni.lu

Specialty section:

This article was submitted to
Signal Processing for
Communications,
a section of the journal
Frontiers in Signal Processing

Received: 07 June 2021

Accepted: 23 August 2021

Published: 04 October 2021

Citation:

Lagunas E, Kibria MG, Al-Hraishawi H,
Maturo N and Chatzinotas S (2021)
Precoded Cluster Hopping for
Multibeam GEO Satellite
Communication Systems.
Front. Sig. Proc. 1:721682.
doi: 10.3389/frsip.2021.721682

Beam hopping (BH) and precoding are two trending technologies for high-throughput satellite (HTS) systems. While BH enables the flexible adaptation of the offered capacity to the heterogeneous demand, precoding aims at boosting the spectral efficiency. In this study, we consider an HTS system that employs BH in conjunction with precoding in an attempt to bring the benefits of both in one. In particular, we propose the concept of cluster hopping (CH), where a set of adjacent beams are simultaneously illuminated with the same frequency resource. On this line, we propose an efficient time–space illumination pattern design, where we determine the set of clusters that shall be illuminated simultaneously at each hopping event along with the dwelling time. The CH time–space illumination pattern design formulation is shown to be theoretically intractable due to the combinatorial nature of the problem and the impact of the actual illumination design on the resulting interference. For this, we make some design decisions on the beam–cluster design that open the door to a less complex still well-performing solution. Supporting results based on numerical simulations are provided which validate the effectiveness of the proposed CH concept and a time–space illumination pattern design with respect to benchmark schemes.

Keywords: satellite communications, multibeam satellite, beam hopping, precoding, demand-matching

1 INTRODUCTION

The first generation of broadband multibeam satellites was launched in the 2000s, with the main objective to deliver internet services to people who had no access to faster forms of internet connectivity (ViaSat Inc., 2018). Driven by the success of the first generation of broadband satellites, new advanced satellite systems were set up during the 2010s with spot beams. Viasat-1 is a clear example of such next generation of satellites, which is able to serve 72 spot beams and reach a total capacity of 140 Gbps. Clearly, these accomplishments established the birth of the so-called generation of high-throughput satellite (HTS) systems (Cola et al., 2015). While wireline and wireless terrestrial broadband service lack the ability to leap across continents, oceans, and difficult-to-access areas, the inherent large coverage footprint of satellite communication networks make them the most suitable solution to expand networks over the world. Therefore, satellites can complement the terrestrial networks and offer important socioeconomic benefits, while increasing the satellite competitiveness.

From frequency/bandwidth to power allocation and coverage, the forthcoming generation of commercial satellite communication payloads offer enhanced flexibility to dynamically satisfy the customers' demands (Kisseleff et al., 2020; NetWorld 2020, 2019). Such reconfigurable satellite systems are clamored by operators and manufacturers to be one of the most groundbreaking

evolutions of satellite communications with an impact on lowering mission costs and enabling satellite systems to become more agile and responsive to market needs (SES, 2020; AIRBUS, 2021). These future satellite architectures are expected to offer terabit per second in-orbit capacity when and where needed. Such throughput enhancements can only be achieved by pushing forward the multibeam architecture with a reduced beam size, taking advantage of frequency reuse and reconfiguring the satellite capacity according to the heterogeneous traffic demands.

In response to the combination of ever-growing data demand with the inherent satellite spectrum scarcity (Kodheli et al., 2020), an intelligent allocation of satellite resources considering the new degrees of flexibility shall be conceived, particularly considering both the actual users' position as well as their traffic demand. This study focuses on two of the most promising disruptive techniques to tackle these specific challenges: linear precoding and time-flexible beam hopping.

1.1 Linear Precoded for Satellite Systems

While conventional satellite systems are designed to operate using an interference avoidance approach through a proper reuse of the available spectrum among beams, more recent paradigms have been proposed and studied which go in the opposite direction through the management and the exploitation of the interference among beams. The objective is clearly to maximize the use of the user link available spectrum (in terms of spectral efficiency), which represents a limited resource of the system. In this context, Vazquez et al. (2016) summarize multiuser multiple input single output (MU-MISO) digital signal processing techniques, such as linear precoding, that can be applied in the user link of a multibeam satellite system operating in full frequency reuse. While the concept of MU-MISO in satellite networks have been mostly theoretical, an actual live-based demonstration supported by the European Space Agency (ESA) has been carried out in ESA project LiveSatPreDem (2020), validating the feasibility of such technique considering the recently amended DVB-S2X specifications to support it. It is worth to remark that precoding is embedded at the gateway, thus keeping the complexity of the payload and user terminal (UT) infrastructure low.

In general, one of the main challenges faced by HTS systems (particularly for precoded systems) is the feeder link congestion, that is, the congestion on the bidirectional communication link between the gateway and the satellite. The increase in the capacity of the user link requires a corresponding increase in the capacity of the feeder link, which is currently limited by few GHz of available bandwidth (Kyrgiazos et al., 2014). In principle, the exploitation of higher frequency bands (e.g., Q/V) by this wireless link could address this issue. However, often this approach is not feasible in practice due to weather impairments at high frequencies (Zhang et al., 2017). A common alternative is the deployment of multiple gateways, where each gateway conveys the signals to be transmitted to a cluster of spot beams. This concept of beam-clustering would be relevant to this study and will be addressed in the next section.

1.2 Time-Flexible Beam Hopping

Beam hopping (BH) was originally proposed to deal with large multibeam coverage areas, by focusing the satellite resources to certain subset of beams, which is active for some portion of time, dwelling just long enough to satisfy the requested demands (Freedman et al., 2015). In doing so, BH is able to increase useable capacity and reduce unmet traffic demands, particularly in the presence of heterogeneous traffic demand.

The conventional BH illumination pattern is illustrated in **Figure 1A**, where the active spot beams are designed to have a border area formed by inactive beams such that a degree of isolation exists between each active beam. Note that the set of illuminated beams changes in each time slot based on a time-space transmission pattern that is periodically repeated. The time axis is divided in windows of duration T_H , which repeat following a regular pattern. Each BH window is segmented in N_s time slots and in each time slot a different set of beams is illuminated. By modulating the period and duration that each of the beams is illuminated, different offered capacity values can be achieved in different beams.

The BH procedure on the one hand allows higher frequency reuse schemes by placing inactive beams as barriers for the co-channel interference, and on the other hand allows the use of a reduced number of onboard power amplifiers, with a consequent reduction of payload mass. BH benefits have been well demonstrated, for example, ESA project BEAMHOP (2016), and the satellite standard DVB-S2X has recently included guidelines to enable beam hopping operation.

On the downside, we noticed that in certain scenarios where more than one adjacent beam is requesting high demand, the performance of BH is affected by the limitation of not being able to simultaneously activate neighboring beams with the same spectrum resource. The latter motivates the contribution of this study.

In summary, BH provides the means to flexibly adapt the offered capacity to the time and geographic variations of the traffic demands, while precoding exploits the multiplexing feature enabled by the use of multiple antenna feeds at the transmitter side to boost the spectral efficiency. These two effective strategies can create unique opportunities if they are properly combined.

1.3 Contribution: Precoded Cluster Hopping

The contributions of this study are summarized as follows:

- 1) Cluster hopping concept: we propose the novel cluster hopping (CH) concept as a natural combination of BH with precoding. In CH, multiple set of adjacent beams are illuminated at the same time with the same frequency resource. We define a cluster as the set of adjacent active beams that are served by a single gateway so that the whole coverage area can be served through multiple clusters/gateways. An example of the proposed CH is shown in **Figures 1B,C**, which requires the use of precoding to deal with the resulting interference as no separation line of inactive beams is considered within a beam cluster. CH was first introduced by the authors in Kibria et al. (2019). Herein,

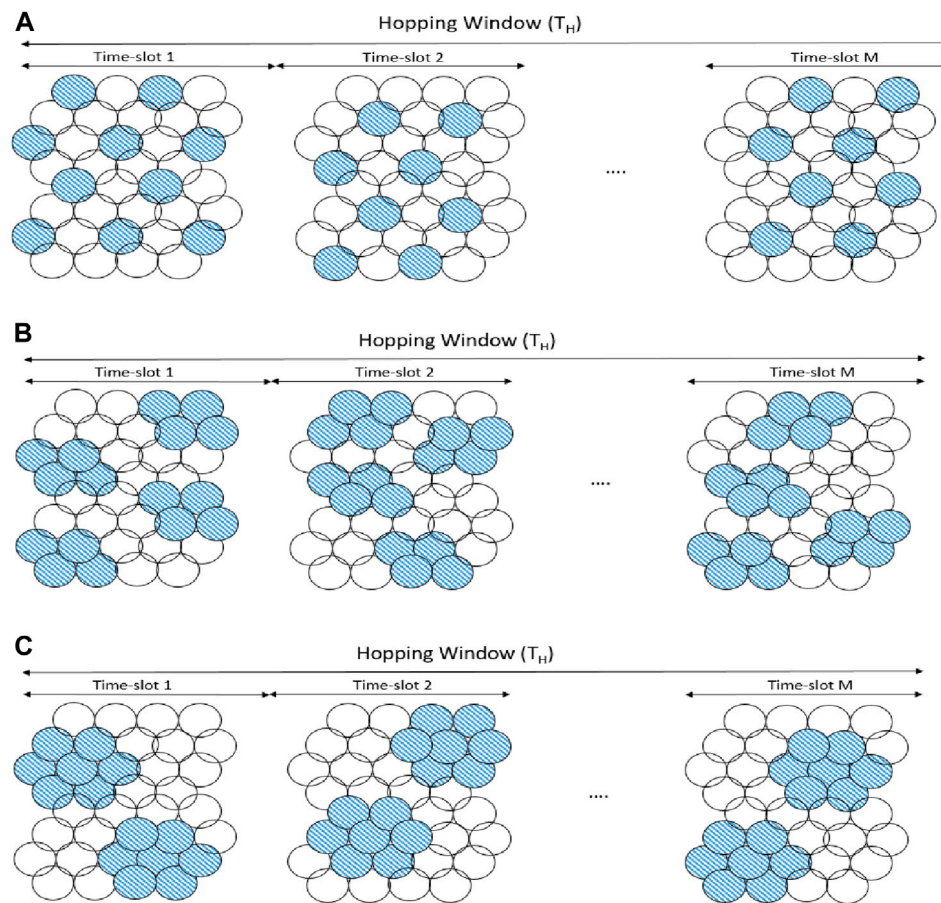


FIGURE 1 | Beam hopping illumination pattern: **(A)** conventional beam hopping; **(B)** proposed cluster hopping with four-beam clusters; and **(C)** proposed cluster hopping with seven-beam clusters.

we expand Kibria et al. (2019) with more technical details, expanding the numerical results.

- 2) Illumination pattern design: the illumination pattern design for conventional BH systems has been studied in Alegre-Godoy et al. (2012), Angeletti et al. (2012), Anzalchi et al. (2010), Cocco et al. (2018), and Lei et al. (2020). While Alegre-Godoy et al. (2012), and Anzalchi et al. (2010) focused on heuristic iterative suboptimal algorithms, Angeletti et al. (2012) and Cocco et al. (2018) considered genetic and simulated annealing algorithms, respectively, targeting global optimal solutions at the expenses of increased computational complexity. Finally, Lei et al. (2020) proposed to integrate deep learning into the optimization procedure in order to accelerate the optimization procedure. Herein, we propose an illumination pattern design for CH (and therefore considering precoding the corresponding clusters) under a fair beam demand satisfaction objective. In particular, we formulate the illumination pattern design as a max-min of the offered vs. demanded capacity subject to a set of practical constraints. The presence of binary assignment variable as well as nonlinearity caused by the interference as a function of such binary assignment variable makes the

problem non-convex and difficult to solve. To tackle this, we propose to limit the clustering to specific forms that allow us to 1) simplify the relationship between a specific beam illumination instance and the resulting interference and 2) reduce the search space of the feasible solutions and, therefore, obtain a low-complex solution. Although optimality cannot be guaranteed, this solution is shown to reach satisfactory results with affordable complexity.

- 3) Numerical evaluation: finally, we present supporting results based on numerical simulations using a software tool (SnT University of Luxembourg, 2020). We evaluate the beam demand satisfaction under different beam-clustering designs and different number of simultaneously activated beams, for different demand instances. We also compare the proposed CH with respect to the conventional BH technique and with respect to Ginesi et al. (2017). The latter represents a preliminary study carried out by ESA, where precoding was first combined with BH and a pragmatic, iterative, and heuristic approach was proposed for the illumination pattern design.

The rest of this article is organized as follows. **Section 2** introduces the system model and relevant nomenclatures. **Section 3** presents the proposed cluster hopping concept considered in this study and addresses the illumination pattern design. Supporting simulation results are presented in **Section 4**, and finally, concluding remarks as well as future research directions are provided in **Section 6**.

2 SYSTEM MODEL

Let us consider a high-throughput multi-beam satellite system with a total of N_b beams, from which only a subset of Q beams, $Q < N_b$, can be simultaneously activated at a particular time instance. We define the illumination ratio as Q/N_b , for example, a 1/4 illumination ratio means that 25% of the total number of beams is illuminated. We assume that the beams that are illuminated employ full-frequency reuse, meaning that all of them operate over the same spectrum B_w . For the sake of clarity, the feeder links (connection between gateways and satellite) is considered ideal, that is, noiseless and without channel impairments.

In this study, we use the following terminologies:

- **Cluster:** a group of adjacent beams simultaneously illuminated with the same spectrum B_w . To cope with the resulting interference, clusters are precoded.
- **Snapshot:** a particular arrangement of illuminated and non-illuminated clusters. For illustration purposes, **Figures 1B,C** show three and three snapshots, respectively. There can be as many as 2^{N_c} possible snapshots, N_c being the number of considered clusters. Of course, not all snapshots are valid in the sense that only a given number of nonadjacent clusters can be illuminated simultaneously because of payload limitations, that is, the illumination ratio.
- **Time slot:** a time slot or time instance defines the time granularity of the hopping operation, that is, the minimum illumination period for a selected snapshot. The hopping window, T_H , is equally divided into N_s time slots. Therefore, $T_H = N_s \times T_s$, where T_s denotes the duration of the time slot.
- **Hopping window:** as anticipated, the hopping window consists of N_s time slots and has a total duration of T_H . It also represents the maximal time period allowed to provide service to all the users in the coverage area.

Let us focus on a particular snapshot and on a particular cluster within that snapshot. The signal vector received by the N_c active beams within the i th cluster is denoted as $\mathbf{y}_i \in \mathbb{C}^{N_c \times 1}$ and further expressed as:

$$\mathbf{y}_i = \mathbf{H}_i \mathbf{x} + \mathbf{n}_i, \quad (1)$$

where $\mathbf{x} \in \mathbb{C}^{Q \times 1}$ denotes the transmitted symbols with $\mathbb{E}[\mathbf{x}\mathbf{x}^H] = \mathbf{I}_Q$ and $\mathbf{H}_i \in \mathbb{C}^{N_c \times Q}$ refers to the channel matrix of cluster i , which includes the components of all active beams and is assumed to be perfectly known at the transmitter, and $\mathbf{n}_i \in \mathbb{C}^{N_c \times 1}$ denotes the

additive Gaussian zero-mean unit-variance noise, that is, $\mathbb{E}[\mathbf{n}_i \mathbf{n}_i^H] = \mathbf{I}_{N_c}$.

For the sake of clarity, we drop the cluster subindex i throughout the following, which applies to any cluster. The entry at the k th row and q th column of the downlink channel matrix \mathbf{H} in (1) between the multibeam satellite and the N_c beams of the cluster is modeled as:

$$[\mathbf{H}]_{k,q} = \frac{\sqrt{G_{Rx} G_{k,q}}}{4\pi \frac{d_k}{\lambda} \sqrt{\kappa T_{Rx} B_w}} \quad (2)$$

where

- G_{Rx} is the receiver antenna gain (assumed to be the same for all UT for simplicity),
- $G_{k,q}$ is the gain from the q th beam seen at the k th beam,
- d_k is the distance between the satellite and k th beam,
- λ denotes the wavelength,
- κ denotes the Boltzmann constant,
- T_{Rx} is the clear sky noise temperature of the receiver.

Consequently, the received signal of the k th beam can be written as:

$$y_k = \underbrace{\mathbf{h}_k^T \mathbf{x}_k}_{\text{desired}} + \underbrace{\sum_{\substack{j \in \mathcal{C} \\ j \neq k}} \mathbf{h}_k^T \mathbf{x}_j}_{\text{intra-cluster interf.}} + \underbrace{\sum_{u \notin \mathcal{C}} \mathbf{h}_k^T \mathbf{x}_u}_{\text{inter-cluster interf.}} + n_k, \quad (3)$$

where \mathbf{h}_k^T denotes the k th row of matrix \mathbf{H} , and we distinguish two types of interference: (i) intra-cluster interference, with \mathcal{C} denoting the set of beams belonging to the same cluster as beam k and (ii) inter-cluster interference, which considers all the transmission signals not intended to cluster \mathcal{C} .

3 CLUSTER HOPPING DESIGN

Both interference components in (3) can be mitigated by considering precoding over all Q active beams. Although this is the best approach in terms of achievable capacity, its implementation is limited by the feeder link congestion. For such number of active beams, multiple and coordinated gateway stations are required, which is considered unlikely in practice due to the synchronization accuracy needed for coordinated precoding (Arapoglou et al., 2016).

Therefore, our first design decision is to mitigate the intra-cluster interference only by precoding clusters independently. Regarding the inter-cluster interference, its effect will be minimized by considering avoiding adjacent clusters to be simultaneously activated. These two assumptions, shall allow us to 1) upload the signals using multiple noncooperative gateways, 2) pursue a design under the assumption of negligible inter-cluster interference, and 3) reduce the number of possible snapshots and, thus, the search space of the cluster hopping design.

Taking into account the aforementioned assumptions, the offered capacity to beam k belonging to cluster \mathcal{C} can be expressed as:

$$c_k = B_w f_{DVB} \left(\frac{P_{\text{beam}} |\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum_{j \in \mathcal{C}, j \neq k} P_{\text{beam}} |\mathbf{h}_k^H \mathbf{w}_j|^2 + 1} \right) [\text{bits/sec}], \quad (4)$$

where f_{DVB} denotes the signal-to-interference and noise ratio (SINR) vs. the spectral efficiency (SE) mapping function according to the adaptive coding and modulation (ACM) scheme considered by the digital video broadcasting (DVB) standard (DVB-S2X, 2014). The transmit power per beam is assumed to be fixed and equally distributed across beams and it is denoted as P_{beam} . It is out of the scope of this study to optimize the transmit power.

As a consequence, the cluster capacity can be obtained by adding all the capacity of the beams belonging to that cluster: $C_i = \sum_{k \in \mathcal{C}_i} c_k$, where we have again reintroduced the cluster subindex i .

3.1 Objective

The objective is to obtain the optimal illumination pattern, that is, set of snapshots and their dwelling time, such that the demands of the beams/clusters are fairly satisfied. In other words, the optimal illumination pattern would be such that achieves $c_i \approx d_i$, $i = 1, \dots, N_b$, $C_i \approx D_i$, and $i = 1, \dots, N_c$, where d_i and D_i denote the demand of i th beam and i th cluster, respectively. Note that this study focuses on the demand-matching at the beam level. The task to distribute the beam capacity to the different end-users of that beam is known as user scheduling (Guidotti and Vanelli-Coralli, 2020; Honnaiah et al., 2021).

3.2 Proposed Illumination Design

Let us define our design variable with a set of binary vectors \mathbf{x}_t of dimension $N_c \times 1$, with components $x_t[i]$ being equal to one when cluster i is active at the time slot t .

Since the optimization of the illumination design is performed at the hopping window level, we scale down the cluster demand at the hopping-window level as $\hat{D}_i = T_H D_i$ [bits/hopping window], and the offered cluster capacity at time slot level as $\hat{C}_i = T_s C_i$ [bits/time-slot]. With these definitions, we can state that the actual offered capacity at the hopping window level can be computed as $\hat{R}_i = \sum_{t=1}^{N_s} x_t[i] \hat{C}_i$ [bits/hopping window], where the cluster offered capacity \hat{C}_i can be easily precomputed and stored.

As discussed, without making any assumption on the snapshot design, the number of possible binary arrangements in \mathbf{x}_t is 2^{N_c} , which might become untractable for realistic multibeam patterns. However, not all are valid snapshots for our problem as we have a couple of constraints, namely, maximum number of active beams per time slot (i.e., $\sum_{i=1}^{N_c} x_t[i] = Q'$), Q' denoting the number of active clusters, and activation of adjacent clusters shall be avoided. The latter constraint can be expressed as:

$$\mathbf{x}_t^T \mathbf{A} \mathbf{x}_t = 0, \quad (5)$$

where matrix $\mathbf{A} \in \{0, 1\}^{N_c \times N_c}$ represents the binary adjacency matrix of the clusters. It is a square symmetric matrix, that is, $\mathbf{A}_{i,j} = \mathbf{A}_{j,i}$ and $\mathbf{A}_{i,j} = 1$ when cluster i is adjacent to cluster j .

With all these in mind, the proposed cluster illumination pattern design is formulated in the following equation:

$$\begin{aligned} \max_{\{\mathbf{x}_t, t=1, \dots, N_s\}} \quad & \left(\frac{\hat{R}_1}{\hat{D}_1}, \dots, \frac{\hat{R}_{N_c}}{\hat{D}_{N_c}} \right) \\ \text{s.t.} \quad & \sum_{i=1}^{N_c} x_t[i] = Q', \\ & \mathbf{x}_t^T \mathbf{A} \mathbf{x}_t = 0, \quad t = 1, \dots, N_s \\ & x_t[i] \text{ binary}, \quad t = 1, \dots, N_s, \quad i = 1, \dots, N_c \end{aligned} \quad (6)$$

We can simplify the max-min optimization problem in (6) by turning it into a maximization problem with the help of an additional slack variable γ along with a new constraint $\frac{\hat{R}_i}{\hat{D}_i} \geq \gamma \triangleq \hat{R}_i \geq \hat{D}_i \gamma$:

$$\begin{aligned} \max_{\{\mathbf{x}_t, t=1, \dots, N_s\}} \quad & \gamma \\ \text{s.t.} \quad & \sum_{i=1}^{N_c} x_t[i] = Q', \\ & \mathbf{x}_t^T \mathbf{A} \mathbf{x}_t = 0, \quad m = 1, \dots, N_s \\ & x_t[i] \text{ binary}, \quad t = 1, \dots, N_s, \quad i = 1, \dots, N_c \\ & \hat{R}_i \geq \hat{D}_i \gamma, \quad i = 1, \dots, N_c \end{aligned} \quad (7)$$

One can observe that problem (7) is a linear programming (LP) problem involving a binary assignment variable. Although the inherent combinatorial problem remains, with the proposed constraints and a careful beam-clustering design, one can reduce significantly the search space. The beam-clustering aspects are discussed in the following section, while some numbers about the search space of problem (7) are provided in **Section 4**.

For solving (7), in this study, we rely on the optimization software Gurobi (GUROBI, 2021), which is convenient to solve mixed integer linear programming (MILP) problems such as the one in (7).

3.3 Clustering Definition

The offered capacity per cluster, that is, \hat{R}_i , strongly depends on the cluster shape and size. Deriving optimal clustering optimization would require an exhaustive search over all possible combinations of clustering options, including an irregular cluster size and overlapping clusters, rendering a huge search space. Moreover, the cluster definition also impacts on the complexity of the system as the number of possible snapshots is a function of the number of clusters. For example, a cluster with a small size will yield to a bigger search space for the problem in (7), while clusters with a big size will reduce the search space but provide less flexibility in the CH operation. To keep the complexity of (7) within tractable limits, we opt to have compact-shaped, nonoverlapping, and equal size cluster due to the following reasons:

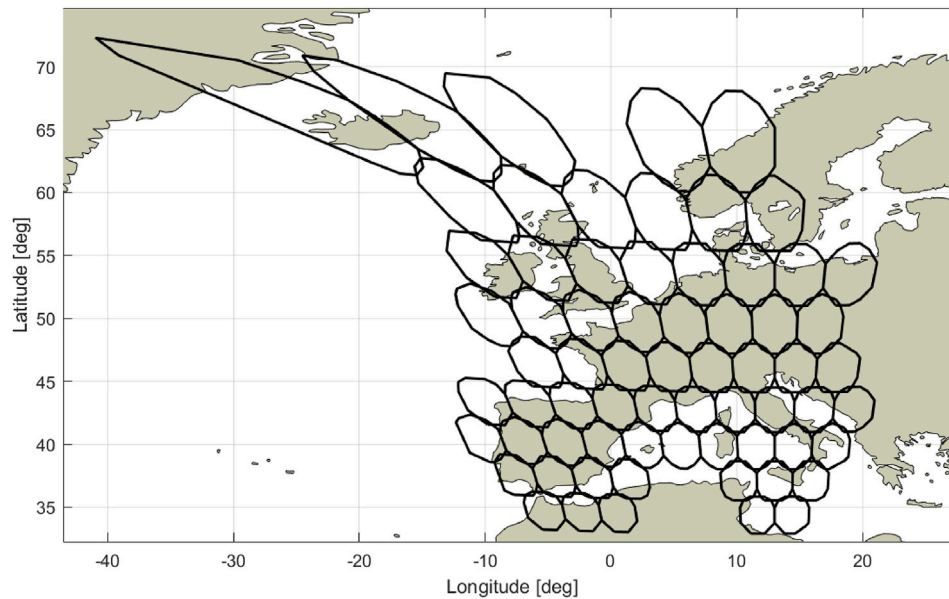


FIGURE 2 | Considered beam pattern with $N_b = 67$ beams.

TABLE 1 | Simulation parameters.

Satellite longitude	13°E (GEO)
Satellite height	35,786 km
Number of beams, N_b	67
Beam radiation pattern	Provided by ESA
Max. beam radiation pattern gain	52 dBi
Downlink carrier frequency	19.5 GHz
Satellite total power, P_{total}	6,000 W
User link bandwidth	500 MHz
Roll-off factor	20%
Effective user link bandwidth, B_w	417 MHz
Roll-off factor	20%
Illumination ratio, (Q/N_b)	1/4, 1/6, and 1/8
Duration of a time slot, T_s	1.3 ms
Hopping window, T_H	256 T_s
User terminal antenna gain, G_{Rk}	39.55 dBi
Noise power, $(\kappa T_{Rk} B_w)$	-118.42 dBW

- In the case of overlapping clusters, there will be a very large number of possible clusters making a huge search space for the proposed problem.
- Compact-shaped clusters are preferred vs. linear- or quasi-linear-shaped clusters in order to exploit the precoding benefits.
- The size of the clusters, as discussed before, brings a complexity-performance trade-off. Different cluster sizes will be evaluated in **Section 4**.

4 SIMULATION RESULTS

The simulation setup for evaluating the performance of the proposed precoded CH in an HTS system is as follows. The 67-beam GEO satellite beam pattern shown in **Figure 2** is

considered. The pattern has been generously provided by ESA in the context of the project FlexPreDem (ESA Project FlexPreDem, 2020). The transmit power per beam P_{beam} is a function of the illumination ratio as it is calculated as $P_{\text{beam}} = P_{\text{total}}/Q$. In other words, the total power P_{total} is equally distributed across the active beams. The rest of the simulation parameters are provided in **Table 1**.

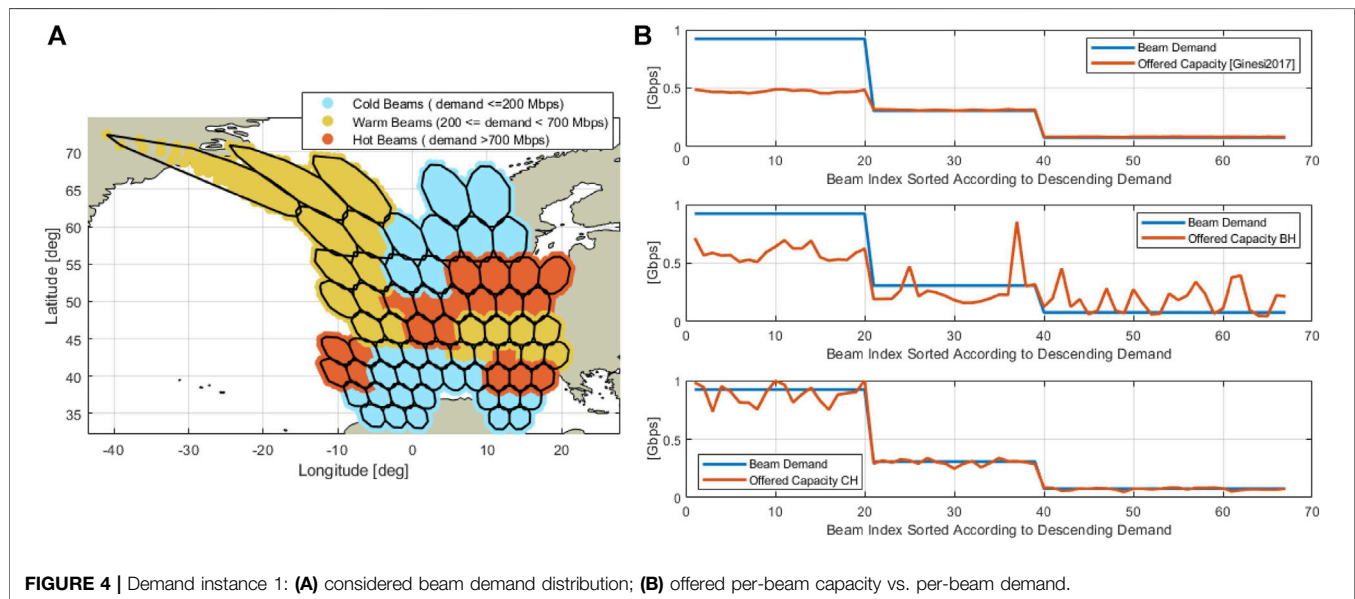
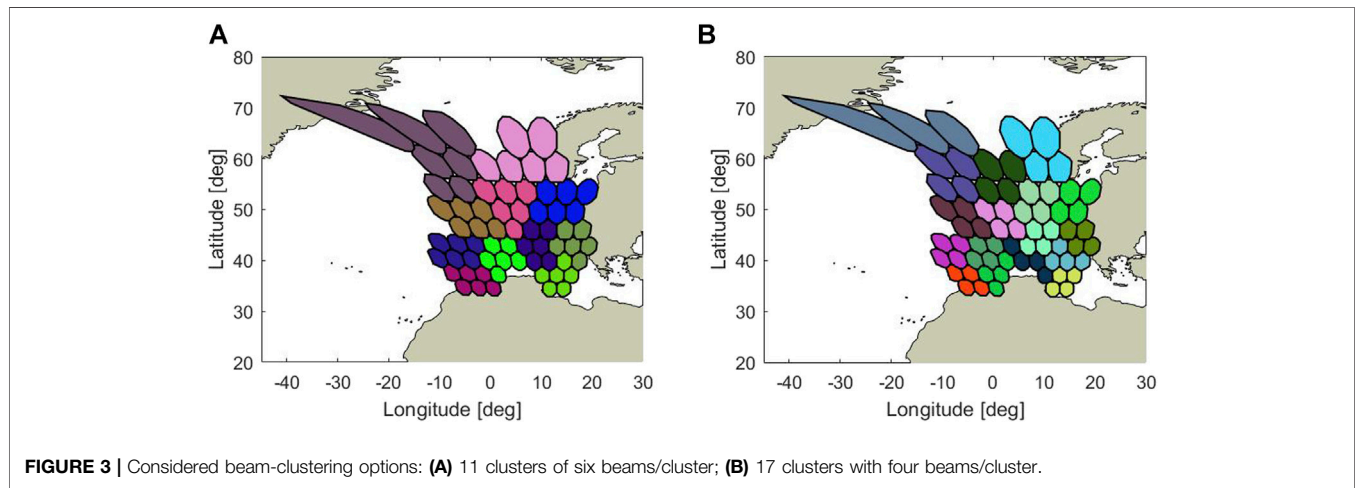
First of all, we provide some numbers in terms of the complexity scalability with the clustering definition. As shown in **Table 2**, assuming a cluster size equal to six beams for all clusters will result in a total of 21,211 clusters if no further assumptions are made. On the other hand, assuming a cluster size equal to four beams for all clusters will result in a total of 1,675 clusters. However, if we make the assumptions proposed in **Section 3.3**, these numbers can be reduced to 17 and 11, respectively, resulting in a more manageable number. As a consequence, we evaluate the performance of the CH concept under these later clustering options, both shown in **Figure 3**. The actual complexity of problem (7) is dictated by the number of snapshots resulting from the combination of the clustering definition and the illumination ratio. In other words, the binary combinations within \mathbf{x}_i are constrained by the number of clusters that can be simultaneously activated (Q') and the adjacent cluster avoidance. Considering these constraint, the number of snapshots N_p for different illumination ratios is given in **Table 2**. As expected, higher illumination ratios allow activating higher number of clusters per snapshot, therefore, resulting in higher number of possible snapshots. Still, the numbers shown in **Table 2** are tractable allowing to find a solution to problem (7) in a matter of seconds with conventional personal computers.

The proposed precoded CH scheme is evaluated in terms of unmet capacity and unused capacity. Both are figures of merits

TABLE 2 | Clustering impact on the combinatorial problem complexity.

Size of cluster	Number of clusters (compact, non-compact, overlapping, nonoverlapping)	Number of clusters (compact, overlapping, nonoverlapping)	Number of clusters (compact, nonoverlapping)
4 Beams	21,211	483	17
6 Beams	1,675	132	11

Assuming compact and nonoverlapping clusters			
Size of Cluster	Number of snapshots (Illum. Ratio 1/4)	Number of snapshots (Illum. Ratio 1/6)	Number of snapshots (Illum. Ratio 1/8)
4 Beams	304	263	101
6 Beams	36	35	11

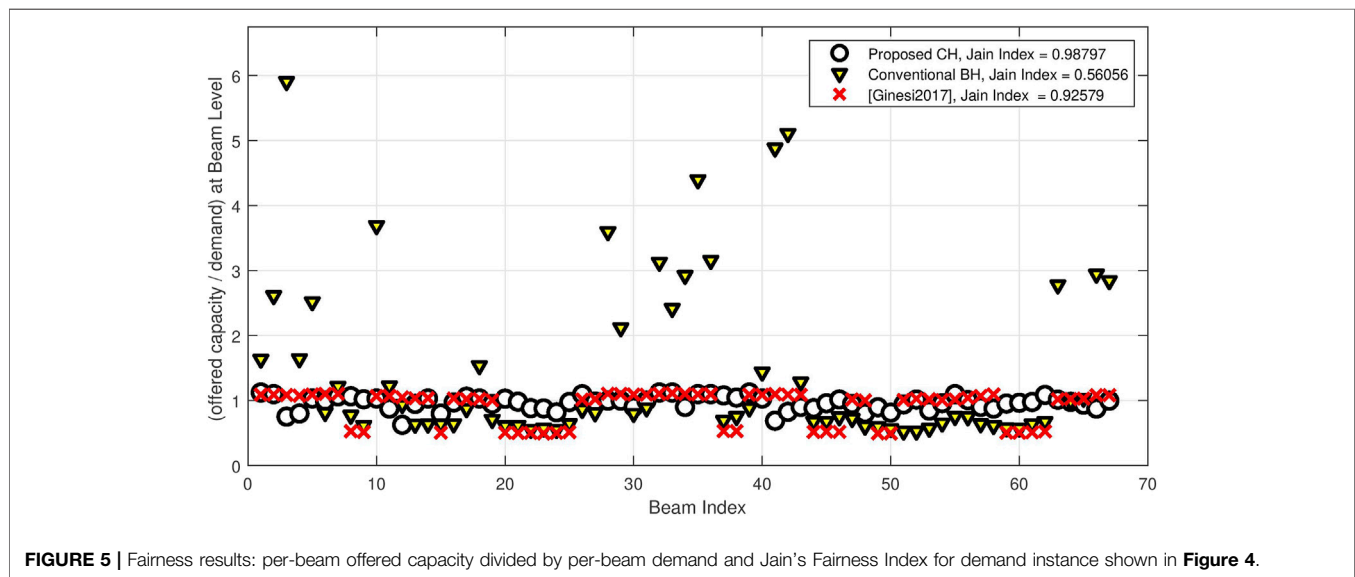


widely used for resource allocation in satellite communications. The first corresponds to the amount of demanded capacity that cannot be satisfied with the actual offered capacity and is defined as $C_{\text{unmet}} = \sum_{i=1}^{N_b} (d_i - c_i)^+$, where $(x)^+ = \max(0, x)$. The

second corresponds to the amount of offered capacity which exceeds the demanded capacity, and it is given by $C_{\text{unused}} = \sum_{i=1}^{N_b} (c_i - d_i)^+$. Ideally, both unmet and unused capacity should be zero.

TABLE 3 | Unmet and unused system capacity results for demand in **Figure 4**. Total demand is 26.46 Gbps.

Technique	Illum. Ratio	Offered capacity (Gbps)	Unmet capacity (Gbps)	Unused capacity (Gbps)	Satisfaction% (%)
Ginesi et al. (2017)	Not applicable	17.75	9.03	0.32	85.40
Conventional BH	1/4	27.17	8.99	9.71	79.89
	1/6	21.98	8.26	3.78	79.99
	1/8	16.64	9.83	0	64.61
Proposed CH	1/4	30.95	2.58	7.07	93.66
	1/6	25.46	1.50	0.50	94.00
	1/8	18.89	7.57	0	72.79

**FIGURE 5** | Fairness results: per-beam offered capacity divided by per-beam demand and Jain's Fairness Index for demand instance shown in **Figure 4**.

Next, we present the performance evaluation of the proposed precoded CH system, which is compared with a conventional BH scheme and the study in Ginesi et al. (2017). The results presented herein have been obtained with the MATLAB-based software tool (SnT University of Luxembourg, 2020). For the conventional BH scheme, we use to solve the same problem as in (7) but assuming single beam clusters and, as a consequence, without precoding. All our results include the inter-cluster interference.

Figure 4A shows a particular demand distribution composed of three types of beams represented with different colors depending on their demand: high demand, medium demand, and low demand. **Figure 4B** shows the per-beam demand vs. the offered per-beam capacity for the three techniques under evaluation. The clustering option of four beams/clusters has been considered for the CH solution in this case. We can observe that Ginesi et al. (2017) satisfies well the low-demand beams while it struggles in meeting the demand of high-demand beams. Similarly, the conventional BH also shows difficulties in matching the demand of high-demand beams, while it shows some mismatch as well for the rest of the beams. Finally, the proposed CH is shown to properly follow the demand of any type of beam. **Table 3** summarizes the system's unmet and unused

capacity results, that is, C_{unmet} and C_{unused} , for the demand in **Figure 4**. **Table 3** also shows the total offered capacity and the satisfaction percentage, which represents the amount of beams that are satisfied. Note that the benchmark (Ginesi et al., 2017) does not apply a specific illumination ratio as the number of active beams per time slot change over time. The first observation is that the proposed CH technique with an illumination ratio of 1/6 is providing the best unmet unused capacity trade-off, with both close to zero. Furthermore, CH is showing better demand satisfaction percentage too. The best results are achieved with an illumination ratio of 1/6 because this provides an overall system offered capacity of 25.46 Gbps, which closely matches the overall requested demand of 26.46 Gbps. From the results in **Table 3**, we conclude that CH combined with a proper illumination ratio outperforms the benchmark schemes.

To complement the previous results, we now evaluate the fairness of the proposed solution in **Figure 5**, where the ratio of the per-beam demand vs. the achieved per-beam offered capacity is shown, as well as the resulting Jain's fairness index proposed in Jain et al. (1984). In this study, we use the Jain's fairness metric as a measure of how the offered capacity matches the demand at a beam level. For this, we define ζ_i as the ratio between the offered

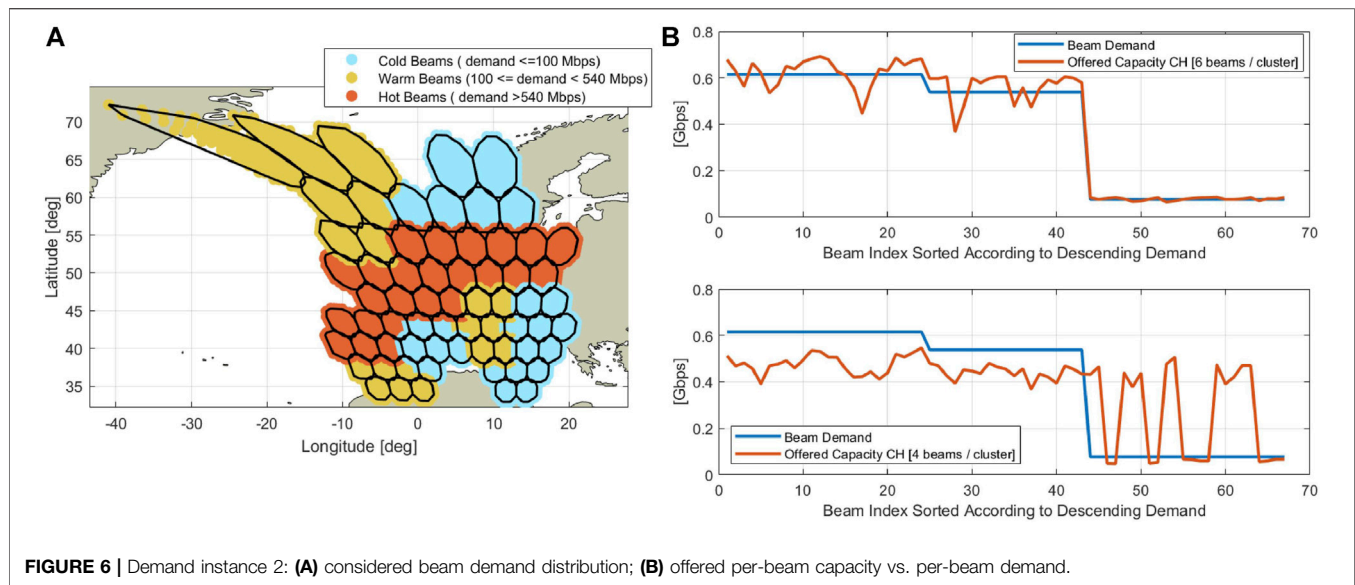


TABLE 4 | Unmet and unused system capacity results for demand in **Figure 6**. Total demand is 26.85 Gbps.

Technique	Illum. Ratio	Offered capacity (Gbps)	Unmet capacity (Gbps)	Unused capacity (Gbps)	Satisfaction (%)
Proposed CH	1/4	36.82	0.10	10.07	99.69
Six beams/cluster	1/6	27.58	0.86	1.59	97.08
	1/8	15.05	11.80	0	58.03
Proposed CH	1/4	30.25	2.27	5.67	94.06
Four beams/cluster	1/6	25.90	5.38	4.44	82.58
	1/8	18.73	11.07	2.95	65.50

capacity C_i and the demanded/ideal capacity D_i , that is, $\zeta_i = \frac{C_i}{D_i}$, $i = 1, \dots, N_b$. In this context, the Jain's fairness index is defined as:

$$J_{FI} = \frac{\left(\sum_{i=1}^{N_b} \zeta_i\right)^2}{N_b \sum_{i=1}^{N_b} \zeta_i^2} \in \left[\frac{1}{N_b}, 1\right]. \quad (8)$$

From **Figure 5**, it can be observed that the proposed CH outperforms the benchmark schemes in terms of fair per-beam demand satisfaction, as the values of ζ_i are closer to the idea value of 1 for all beams. The fairness of the proposed approach is confirmed by the Jain's index shown in the legend of **Figure 5**, where the proposed CH reaches a Jain's index of 0.99 (superior to that of the benchmarks).

Let us test now another demand instance with bigger demand areas, like the one shown in **Figure 6A**. For such big areas of demand, we expect the clustering of six beams/cluster to be a better fit. **Table 4** shows the results achieved with the proposed CH for different clustering options and different illumination ratios. The best match is given by the six beam/cluster option with 1/6 illumination ratio, where the unmet and unused capacity are equal to 0.86 Gbps and 1.59 Gbps, respectively, with a satisfaction percentage of 97%. The results shown in **Table 4** provide evidence on the fact that not only it is important to select an accurate illumination ratio but also a clustering definition adapted to the expected demands. Finally, to confirm the superiority of the six

beam/cluster for such type of demand distributions, **Figure 6B** provides the per-beam details of demand vs. offered capacity. It can be observe that the four beam/clustering option not only has problems in satisfying high-demands but also presents some mismatches for the low-demand beams.

5 PRACTICAL CONSIDERATIONS

5.1 Channel State Information Acquisition

Besides the synchronization aspects natural from beam-hopped transmission (Freedman et al., 2015), the main challenge of the proposed cluster hopping concept is the need of channel state information (CSI) at the gateway side. The most challenging problem in beam-hopped and precoded satellite systems is how often a ground terminal can measure its CSI vector (meaning the channel coefficient w.r.t. the satellite antennas). While the CSI estimation procedure can be based on already existing methods, the cluster hopping scheme requires some ad hoc adaptations due to time-variant nature of the cluster hopping procedure. In fact, since the set of illuminated beams changes over time, each user terminal is able to estimate a subset of coefficients of the complete CSI vector, which depends on the particular cluster than is being illuminated at that particular time instant. The latter can potentially lead to situations in which the gateway does not

have all the needed coefficients to compute the precoding matrix. When the channel is relatively stable, the use of previous CSI estimates may solve the problem, otherwise joint processing of previous and new CSI coefficients would be required as well as prediction methods.

Furthermore, the general problem in beam-hopped satellite systems is how often a ground terminal can measure its CSI vector (meaning the channel coefficient w.r.t. the satellite antennas). Clearly, a ground terminal can only perform measurements when it is being illuminated, and the number of measured CSI components depends on the particular cluster than it is being illuminated at that particular time instant. We should distinguished two cases:

- Illumination pattern composed of nonoverlapping clusters: this is the case assumed in this study. In this case, each ground terminal only needs the knowledge of the CSI components related to the satellite antennas that are active in the cluster that it belongs to. Therefore, we propose to rely on the CSI gathered in the previous time instant that this particular cluster was illuminated (which of course will imply some additional delay depending on the illumination period).
- Illumination pattern composed of overlapping clusters: this would be the case when dealing with high traffic demand areas that need to be illuminateds most of the time. For the sake of clarity, let us assume an example composed of three beams and two non-orthogonal clusters, the first cluster composed of beam 1 and beam 2, and the second cluster composed of beam 2 and beam 3. Let us focus on the terminals belonging to the beam 2 coverage area. Note that beam 2 is always active but once together with beam 1 and once together with beam 3. Therefore, this configuration implies that high-demand beams that are more often illuminated (i.e., beam 2) will have more accurate CSI that low-demand beams (i.e., beam 1 and beam 3), which are less often activated.

Generally speaking, we do not foresee the outdated CSI to have a strong impact. This is because a single DVB-S2(X) super-frame is enough to obtain a good channel estimation and, therefore, the outdated CSI will only be needed for the initial (single) super-frame.

5.2 Future Payload Antenna Systems

The results presented in this study have been obtained assuming a Direct Radiating Array (DRA)-based footprint pattern, which have been generated with internal software by ESA in 20 GHz, with 750 elements spaced five lambda, able to provide 67 beams within the desired coverage area. The trends in the satellite communications industry are evolving towards more advanced antenna architectures, for example, (defocused) phased array fed reflector (PAFR), whose phase response may differ from conventional single-feed-per-beam architecture or the DRA considered in this study. The PAFR may offer some benefits such as lower cost, high beam resolution, and smaller array size.

6 CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

In this study, we have proposed the combination of precoding and time-flexible payloads with BH capabilities. Focusing on the convergence of both techniques, we have proposed the so-called cluster hopping (CH) concept, which seamlessly combines these two paradigms and utilizes the strong points of each one. Supporting results based on numerical simulations are provided which validate the effectiveness of the proposed system in comparison with conventional BH and other works available in the literature. Particularly, CH shows great promise when dealing with high demands that cover large portionals of the Earth, thus spanning multiple satellite beams. The results of this study have highlighted the importance of an appropriate clustering design together with an appropriate illumination ratio, both based on the expected demand distribution. The latter opens opportunities for future research in this subject, namely the optimal clustering definition based on demand distribution input and the appropriate portion of beams that needs to be activated at a time. Furthermore, this study has considered the transmit power out of the scope for the sake of clarity but the transmit power represents another degree of freedom that can be considered within the optimization problem.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

EL led this manuscript in both technical contribution, production of results, and writing. MK contributed to the technical ideas and help in producing the numerical results. HA-H and NM contributed to the developed techniques. SC contributed in conceiving the main idea and supervised the findings of this study. All authors contributed to the manuscript and approved the submitted version.

FUNDING

This study has been partially supported by the Luxembourg National Research Fund (FNR) under the project FlexSAT (C19/IS/13696663) and by the European Space Agency (ESA) under the activity FlexPreDem: Demonstrator of Precoding Techniques for Flexible Broadband Systems. Please note that the views of the authors of this study do not necessarily reflect the views of ESA.

ACKNOWLEDGMENTS

The authors would like to thank Stefano Andrenacci and Joel Grotz from SES and Daniel P. Arapoglou from ESA for their support and valuable discussions and suggestions during the executions of this work.

REFERENCES

- AIRBUS (2021). Flexible Payloads. Available at: <https://www.airbus.com/space/telecommunications-satellites/flexible-payloads.html> (Accessed May 24, 2021).
- Alegre-Godoy, R., Alagha, N., and Vázquez-Castro, M. A. (2012). "Offered Capacity Optimization Mechanisms for Multi-Beam Satellite Systems," in 2012 IEEE International Conference on Communications (ICC) (IEEE), 3180–3184. doi:10.1109/ICC.2012.6364414
- Angeletti, P., Prim, D. F., and Rinaldo, R. (2012). "Beam Hopping in Multi-Beam Broadband Satellite Systems: System Performance and Payload Architecture Analysis," in 24th AIAA International Communications Satellite Systems Conference (San Diego, California: AIAA). doi:10.2514/6.2006-5376
- Anzalchi, J., Couchman, A., Gabellini, P., Gallinaro, G., D'Agostina, L., Alagha, N., et al. (2010). "Beam Hopping in Multi-Beam Broadband Satellite Systems: System Simulation and Performance Comparison with Non-hopped Systems," in Adv. Satellite Multimedia Systems Conf./Signal Process. For Space Commun. Workshop (Cagliari, Italy: ASMS/SPSC). doi:10.1109/ASMS-SPSC.2010.5586860
- Arapoglou, P.-D., Ginesi, A., Cioni, S., Erl, S., Clazzer, F., Andrenacci, S., et al. (2016). Dvb-s2x-enabled Precoding for High Throughput Satellite Systems. *Int. J. Satell. Commun. Netw.* 34, 439–455. doi:10.1002/sat.1122
- Cocco, G., de Cola, T., Angelone, M., Katona, Z., and Erl, S. (2018). Radio Resource Management Optimization of Flexible Satellite Payloads for DVB-S2 Systems. *IEEE Trans. Broadcast.* 64, 266–280. doi:10.1109/TBC.2017.2755263
- Cola, T. D., Tarchi, D., and Vanelli-Coralli, A. (2015). Future Trends in Broadband Satellite Communications: Information Centric Networks and Enabling Technologies. *Int. J. Satellite Commun. Networking* 33, 5. doi:10.1002/sat.1101
- DVB-S2X (2014). Second Generation Framing Structure, Channel Coding and Modulation Systems for Broadcasting, Interactive Services, News Gathering and Other Broadband Satellite Applications; Part 2: DVB-S2 Extensions (DVB-S2x), Document ETSI EN 302 307-2. Available at: https://www.etsi.org/deliver/etsi_en/302300_302399/30230702/01.01.01_20/en_30230702v010101a.pdf.
- ESA project BEAMHOP (2016). *Beam Hopping Techniques in Multi-Beam Satellite Systems*. Noordwijk, Netherlands: European Space Agency. Available at: <https://artes.esa.int/projects/beam-hopping-techniques-multi-beam-satellite-systems-eads-astrium>.
- ESA Project FlexPreDem (2020). Demonstrator of Precoding Techniques for Flexible Broadband Satellite Systems. Available at: <https://artes.esa.int/projects/flexpredem>.
- ESA project LiveSatPreDem (2020). Live Satellite Precoding Demonstration. Available at: <https://artes.esa.int/projects/livesatpredem>.
- Freedman, A., Rainish, D., and Gat, Y. (2015). "Beam Hopping – How to Make it Possible," in Proc. Of Ka and Broadband Communication Conference (IEEE).
- Ginesi, A., Re, E., and Arapoglou, P. (2017). "Joint Beam Hopping and Precoding in HTS Systems," in 9th Int. Conf. On Wireless and Satellite Systems (WiSATS) (IEEE).
- Guidotti, A., and Vanelli-Coralli, A. (2020). Clustering Strategies for Multicast Precoding in Multibeam Satellite Systems. *Int. J. Satell. Commun. Netw.* 38, 85–104. doi:10.1002/sat.1312
- GUROBI (2021). Mathematical Programming Solver. Available at: <https://www.gurobi.com/>.
- Honnaiah, P. J., Lagunas, E., Spano, D., Maturo, N., and Chatzinotas, S. (2021). "Demand-based Scheduling for Precoded Multibeam High-Throughput Satellite Systems," in IEEE Wireless Communications and Networking Conference (WCNC) (IEEE). doi:10.1109/wcnc49053.2021.9417300
- Jain, R., Chiu, D., and Hawe, W. (1984). *A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer System*. Hudson, MA: DEC Technical Report, 301.
- Kibria, M. G., Lagunas, E., Maturo, N., Spano, D., and Chatzinotas, S. (2019). "Precoded Cluster Hopping in Multi-Beam High Throughput Satellite Systems," in 2019 IEEE Global Communications Conference (GLOBECOM) (IEEE), 1–6. doi:10.1109/globecom38437.2019.9013589
- Kisseleff, S., Lagunas, E., Abdu, T., Chatzinotas, S., and Ottersten, B. (2020). Radio Resource Management Techniques for Multibeam Satellite Systems. *IEEE Commun. Lett.* 25, 2448–2452. doi:10.1109/LCOMM.2020.3033357
- Kodheli, O., Lagunas, E., Maturo, N., Sharma, S. K., Shankar, B., Mendoza, J. F., et al. (2020). "Satellite Communications in the New Space Era: A Survey and Future Challenges," in IEEE Communications Surveys Tutorials (ArXiv: 2002.08811).
- Kyrgiazos, A., Evans, B. G., and Thompson, P. (2014). On the Gateway Diversity for High Throughput Broadband Satellite Systems. *IEEE Trans. Wireless Commun.* 13, 5411–5426. doi:10.1109/TWC.2014.2339217
- Lei, L., Lagunas, E., Yuan, Y., Kibria, M. G., Chatzinotas, S., and Ottersten, B. (2020). Beam Illumination Pattern Design in Satellite Networks: Learning and Optimization for Efficient Beam Hopping. *IEEE Access* 8, 136655–136667. doi:10.1109/access.2020.3011746
- NetWorld 2020 (2019). White Paper: SatCom Resources for Smart and Sustainable Networks and Services. Available at: <https://www.networld-europe.eu/whitepaper-satcom-resources-for-smart-and-sustainable-networks-and-services/> (Accessed May 24, 2021).
- SES (2020). Unlocking Agility of SES Satellites with Adaptive Resource Control. Available at: <https://www.ses.com/blog/unlocking-agility-ses-satellites-adaptive-resource-control>.
- SnT University of Luxembourg (2020). Demonstrator of Precoding Techniques for Flexible Broadband Satellite Systems. Available at: https://www.wfr.uni.lu/snt/research/sigcom/sw_simulators/flexpredem.
- Vazquez, M. A., Perez-Neira, A., Christopoulos, D., Chatzinotas, S., Ottersten, B., Arapoglou, P.-D., et al. (2016). Precoding in Multibeam Satellite Communications: Present and Future Challenges. *IEEE Wireless Commun.* 23, 88–95. doi:10.1109/mwc.2016.1500047wc
- ViaSat Inc (2018). Satellite Internet Services. Available at: <https://www.satelliteinternet.com/> (Accessed May 24, 2021).
- Zhang, F., Zuo, Y., and Sun, H. (2017). "Techniques of Rain Fade Countermeasures in Ka-Band Satellite Communication on Ships," in 2017 16th International Conference on Optical Communications and Networks (ICOON) (IEEE), 1–3. doi:10.1109/icoon.2017.8121382

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Lagunas, Kibria, Al-Hraishawi, Maturo and Chatzinotas. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Analysis of a 2D Representation for CPS Anomaly Detection in a Context-Based Security Framework

Sara Baldoni¹, Marco Carli¹ and Federica Battisti^{2*}

¹Department of Industrial, Electronic and Mechanical Engineering, Roma Tre University, Rome, Italy, ²Department of Information Engineering, University of Padova, Padova, Italy

OPEN ACCESS

Edited by:

Monica Bugallo,
Stony Brook University, United States

Reviewed by:

Nasharuddin Zainal,
National University of Malaysia,
Malaysia
Joilson Rego,
Federal University of Rio Grande do
Norte, Brazil

*Correspondence:

Federica Battisti
federica.battisti@unipd.it

Specialty section:

This article was submitted to
Signal Processing for
Communications,
a section of the journal
Frontiers in Signal Processing

Received: 12 November 2021

Accepted: 24 December 2021

Published: 21 January 2022

Citation:

Baldoni S, Carli M and Battisti F (2022)
Analysis of a 2D Representation for
CPS Anomaly Detection in a Context-
Based Security Framework.
Front. Sig. Proc. 1:814129.
doi: 10.3389/frsip.2021.814129

In this contribution, a flexible context-based security framework is proposed by exploring two types of context: distributed and local. While the former consists in processing information from a set of spatially distributed sources, the second accounts for the local environment surrounding the monitored system. The joint processing of these two types of information allows the identification of the anomaly cause, differentiating between natural and attack-related events, and the suggestion of the best mitigation strategy. In this work, the proposed framework is applied the Cyber Physical Systems scenario. More in detail, we focus on the distributed context analysis investigating the definition of a 2D representation of network traffic data. The suitability of four representation variables has been evaluated, and the variable selection has been performed.

Keywords: security, context, anomaly detection, cyber-physical systems, network traffic, 2D representation

1 INTRODUCTION

During the last years, we witnessed a rapid spread of connected devices. This phenomenon involved several market segments from mass market to critical infrastructures (e.g., healthcare, transportation, energy and industrial systems) thus leading to a huge expansion of the attack surface. In addition, due to the use of connected devices in safety-critical applications, attacks may potentially result in the denial of pivotal services to the society or in life losses. To address this issue, the monitoring of connected systems and the identification of anomalous behaviors becomes of paramount importance.

The anomaly detection methods available in the literature can be mainly classified into two categories: signature-based and profile-based. The techniques belonging to the first class detect known anomalies by exploiting the a-priori knowledge of their features. The approaches residing in the second category, on the other hand, exploit the history of the nominal system behavior to define its normal profile. Then, an anomaly is defined as a system behavior that is significantly different from the modeled one. This may be due both to malicious actions and to genuine but unusual activities (Fernandes et al., 2015). Both categories show advantages and disadvantages. Profile-based techniques do not require a model for the anomalous behaviors thus allowing the detection of new and unforeseen anomalies. Signature-based approaches, on the contrary, are able to detect only previously known anomalous behaviors. However, since only well-known anomalies are identified, the false alarm rate can be reduced.

Anomaly detection systems can be also differentiated based on the approach adopted for setting the threshold. More in detail, two classes can be identified: the first foresees the manual setting of alert thresholds that are monitored by experts, whereas the second relies on an automatic approach that can also be based on artificial intelligence. The latter systems have proven to be more effective,

adaptive with respect to traffic variations (without the need to manually re-calibrate the thresholds) and require a reduced human intervention. In this work, the selection of the approach has been made considering that anomalies show the following features (Pang et al., 2021):

- *unknownness*: anomalies are linked to many unknowns (e.g., behavior, data structure, distribution);
- *heterogeneity*: anomalies are irregular and have different properties;
- *rarity*: anomalies are rare, so that it is difficult to gather a large amount of labeled anomalous samples.

For these reasons, in this work we propose a profile-based context-aware security framework which relies on deep learning approaches to detect anomalies. The core idea of the presented research is to satisfy the security requirements through the exploitation of a set of sources and the joint processing of the information flow. The proposed approach involves both local and non-local data. Local data account for information that is dependent on the environment in which they are collected whereas non-local data are monitored on a larger area typically through the deployment of sensors. In the following, non-local data will be indicated as distributed context, since the information sources are assumed to be spatially distributed, whereas local data will be referred to as the local context, since they allow to gather information about what happens in the proximity of the monitored system. The idea of context-aware security has been introduced in (Wang et al., 2010) where the context has been defined as *the set of environmental states and settings that either determine an application's behavior or in which an application event occurs*. Although this concept has been previously addressed in the literature, a general framework which jointly handles both distributed and local contexts is missing. Therefore, the combined processing of local and non-local data paves the way for the definition of a new approach for context-based security.

In this work, we apply the proposed framework to the Cyber Physical System (CPS) scenario. A CPS can be defined as the integration of computing, communication, and control capabilities for monitoring and managing physical objects. The connectivity allowed by the use of the Internet, on one hand has extended the ability of sharing information and on the other one has made these systems prone to vulnerabilities that did not exist before. The interaction between the Internet and CPSs requires greater efforts to ensure the security of connected systems. In fact, the security of CPSs concerns several aspects such as data collection, information transmission, and processing and control subsystems. The exploitation of both local and distributed contexts for estimating the anomaly origin in CPSs allows to evaluate the anomaly impact on the system, with respect to the available reaction and mitigation strategies, also considering the associated costs. In this work, we focus on the distributed-context analysis by adopting a 2D representation of network traffic to design the anomaly detection system. The use of a 2D representation allows to characterize the multi-input information collected from the sensors in a compact form.

The reminder of the paper is organized as follows. In **Section 2** the literature concerning context-based security and network anomaly detection systems based on multi-dimensional representations of data is reviewed. Then, **Section 3** describes the proposed framework and its application to the CPS scenario. Moreover, **Section 4** describes the dataset selected for the performed study, and the 2D traffic representation issue. At last, some preliminary results concerning the 2D data representation are provided in **Section 5**, and in **Section 6** the conclusions are drawn.

2 RELATED WORKS

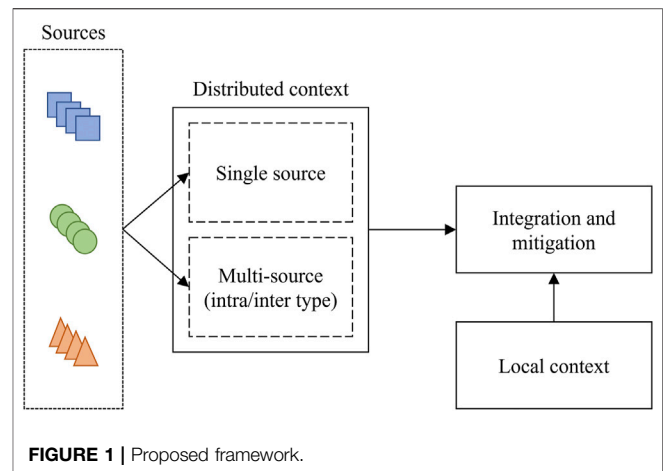
In this section, the related works concerning both context-based security approaches and anomaly detection methods exploiting a multi-dimensional representation of data are detailed.

2.1 Context-Based Security Approaches

In previous works, context-related information has been adopted for improving the safety and security of CPSs, (Ivanov et al., 2018). More specifically, context data has been exploited both for inferring information about the system state and for preventing wrong detections due to the presence of non reliable data. In (Sylla et al., 2019), for instance, a context-aware security architecture for Internet of Things (IoT) is proposed. More in detail, the authors suggest to select the security and privacy mechanisms based on the user contextual information (e.g., mobility). In (Dsouza et al., 2019), a context-aware biometric security framework fusing real-time data with contextual information such as the client setting area, lighting, and time is presented. Context is considered also in the security infrastructure for the IoT systems proposed in (Roukounaki et al., 2019). The environmental impact has been exploited in (Sharaf Dabbagh and Saad, 2019), where device fingerprinting for IoT authentication is analyzed. In this case the approach relies on the assumption that an attacker will not be able to imitate the environmental changes experienced by the legitimate device thus failing in reproducing an environment-based fingerprint. Moreover, an attribute-based encryption method which automatically learns the attributes thanks to a context-aware module is proposed in (Ghosh et al., 2021). Similarly, in (Alagar et al., 2018), a context-sensitive role-based access control technique for healthcare IoT is presented. Furthermore, in (Park et al., 2020), the concept of context-aware intrusion detection systems is realized by including networking conditions (e.g., source and destination address/port, access frequency, data traffic), and systematic operation conditions (e.g., idle CPUs or memory load) in the analyzed data. At last, in (Ehsani-Besheli and Zarandi, 2018), the context is exploited for detecting anomalies in embedded systems communications. The key innovation of our framework is the joint exploitation of local and distributed contexts. As previously mentioned, in fact, a framework which jointly processes local and non-local data is currently missing.

2.2 Security Approaches Based on Multi-Dimensional Representation of Traffic Data

In the literature, few solutions for network anomaly detection have been proposed using a 2D or 3D representation of data. A first example of image-based network traffic visualization is provided in (Kim et al., 2004). The authors exploit source and destination IP addresses and destination port number to represent traffic in 3D. They represent each flow as a point in the 3D space defined by the three attributes. In presence of attacks, regular patterns arise whereas legitimate traffic is widely and irregularly dispersed. To detect the attacks, an attack signature table is defined so that by comparing the packet signatures with the pre-defined ones it is possible to identify the presence and the type of attack. In (Kim and Reddy, 2005a) classical image processing techniques have been exploited to analyze traffic patterns. To define the images packet counts in the address domain are used. The same representation is employed in (Kim and Reddy, 2005b) where the authors compute the DCT of the image and select a set of coefficients for computing the standard deviation. This value is used as anomaly detection metric by defining a lower and upper threshold for the standard deviation under nominal conditions. Moreover, the authors propose using motion prediction techniques to predict following attack targets. Furthermore, Nataraj et al. proposed a 2D representation of malware binaries (Nataraj et al., 2011), and presented a malware classification technique based on image processing methods. More recently, deep learning techniques have been exploited. In (Wang et al., 2017), malware traffic classification is performed based on an image representation of network traffic and a Convolutional Neural Network (CNN). The authors grouped network data based on flows and sessions and found out that different types of traffic result in different images, whereas images within the same class are consistent. Taheri et al. used the same representation for detecting botnet in the IoT environment (Taheri et al., 2018). In addition, an ensemble method employing pre-trained networks and fine-tuning for malware classification has been presented in (Vasan et al., 2020). Finally, an hybrid model based on both unsupervised and supervised methods for malware detection and classification has been proposed in (Venkatraman et al., 2019). Another approach for representing general time series data as images has been proposed in (Zhang et al., 2019), where a multi-scale signature matrix is employed to characterize the system using different time steps. These matrices are then given as input to a convolutional encoder and an attention-based convolutional Long Short Term Memory (LSTM) to capture the temporal patterns. The same data representation has been used in (Luo et al., 2021), and applied to CPSs. In this case, a single scale has been used and the different time series represent measurements provided by a set of sensors. Concerning the processing, a CNN-based autoencoder has been used. Finally, a deep learning model has been employed for network intrusion detection in (Mohammadpour et al., 2018). More specifically, a one-



dimensional feature vector has been re-arranged in a 2D structure and then provided as input to a CNN.

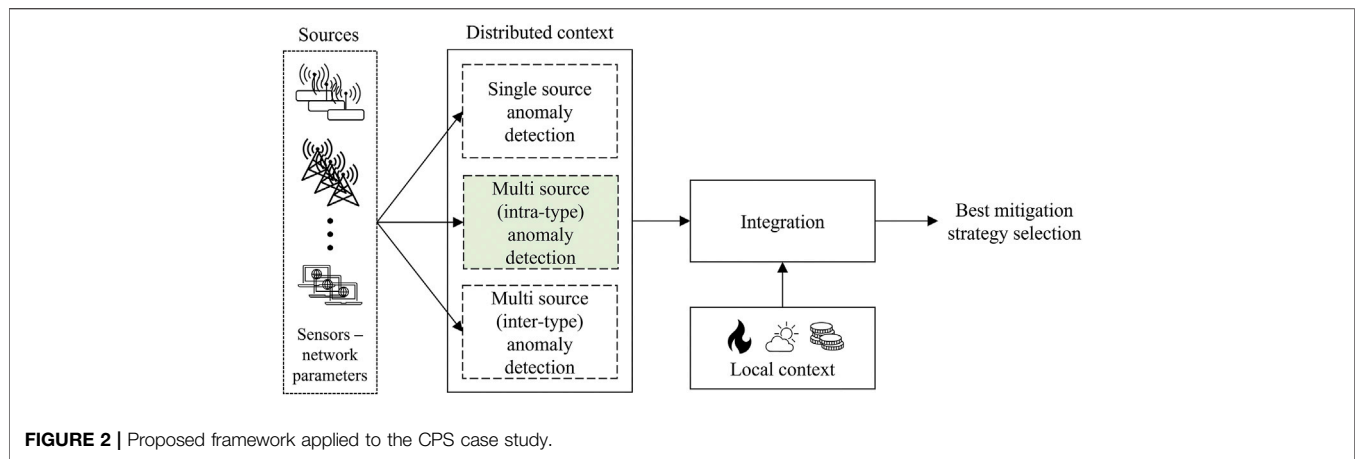
3 PROPOSED FRAMEWORK

The proposed framework aims at providing an effective tool for detecting anomalies and for suggesting the best mitigation actions. The idea underlying this framework is the exploitation of both local and distributed contexts as shown in **Figure 1**.

More specifically, the set of sources provides the input for the distributed context building block. The input type may vary depending on the specific application (e.g., sensor measurements, signals, network parameters). In this stage, three processing options are available: the detection of anomalies through the information provided by single information sources and the late fusion of the outcomes, the joint processing of the information coming from sources of the same type, or the joint analysis of data gathered from different types of sources. As previously mentioned, anomaly detection will be performed through a profile-based deep learning approach. This choice guarantees a detection system able to work independently from the specific type of attack and that consequently can be applied in a wide number of cases. The selected learning approach is based on autoencoders, which are unsupervised architectures that have already been applied for anomaly detection in different application scenarios.

The motivation behind this choice is that the autoencoder is enforced to learn important regularities of the data to minimize reconstruction errors. If the algorithm is trained on non-attacked data, a non-desired modification in the system behavior will result in large reconstruction errors.

The information coming from the local context is then integrated in order to differentiate between natural and attack-related anomalies. A natural event, in fact, may be associated to an anomalous local context, and should concern more than one source deployed in the same area. Once the root cause has been identified, and the mitigation costs have been taken into account, the best mitigation strategy can be selected.



The proposed framework is extremely flexible since it can be adapted to different application scenarios with a minimal modification of its building blocks. Depending on the specific application, the inputs of the proposed framework may change and, according to this variation, also the practical definition of the context evolves. Moreover, the number of parameters that can be detected for performing the anomaly detection can be increased to achieve a desired level of detection accuracy, (Wang et al., 2010).

3.1 CPS Security Case Study

CPSs can be described through a three-layer architecture: perception, transmission, and application. The first layer collects data in real-time, the second allows data exchange, and the last layer realizes data processing and control functionalities. CPSs directly interact with the surrounding environment and are usually deployed in groups on a pre-defined area. Therefore, the distributed nature of CPSs, together with their inter-dependency with the deployment environment makes them the perfect application scenario for the proposed security approach. The flowchart of the proposed framework applied to the CPS scenario is presented in **Figure 2**.

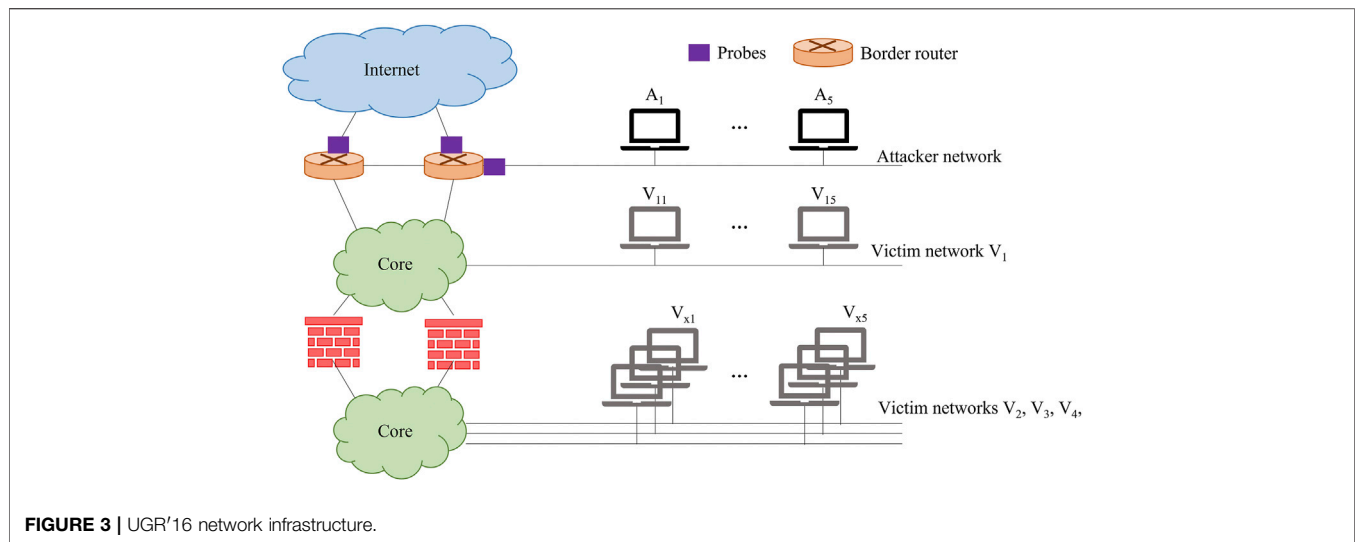
According to the CPS definition, the distributed context could work both analyzing the measurements, and processing the network traffic data. Since, as pointed out in (Luo et al., 2021), the majority of the research literature has focused on detecting anomalies from sensor and actuator data, in this work we aim at analyzing the anomaly detection issue from the network traffic data point of view. Concerning the local data, in this case, they may concern the operating conditions of the CPS (e.g., weather forecast or presence of natural emergencies like fires and earthquakes), and the cost of the different mitigation actions.

We aim at defining the multi-source intra-type anomaly detection subsystem of the distributed context building block. To this end, we investigate the use of a 2D representation of traffic data. Thanks to this representation, the information gathered by a set of distributed nodes (i.e., the distributed context), is analyzed by providing to the anomaly detection system a single input. In addition, the 2D data structure is suitable to be processed by deep-learning based algorithms. The complete realization of this

subsystem aims at defining an anomaly detection technique that will advance the state-of-the-art by working towards three specific objectives:

1. selection of the most effective two-dimensional representation of network traffic information;
2. definition of a deep learning model for the detection of anomalies based on the two-dimensional representation of traffic data;
3. design and implementation of a context-sensitive network anomaly detection system by exploiting the information gathered from a set of distributed nodes.

Among these three goals, in this contribution we focus on the first. To do so, the first issues to be solved is the research of available datasets to be used in the training phase. In fact, even if network security is a well-studied topic, the availability of verification datasets does not follow the rapid evolution trend of attack strategies and communication system development. Furthermore, the use of deep learning-based analytic methods requires a large amount of data to effectively train networks. Therefore, the choice of the dataset containing the traffic data to be analyzed, is an important step towards the realization of the proposed anomaly detection system. To this aim, two key aspects have to be considered: the sampling interval and the total duration of the data recording. The sampling interval must be short and fixed for all recorded data. If the sampling interval was long, in fact, assuming that the recorded data is analyzed exploiting time windows, in order to collect a sufficient amount of samples in each of them, a single time window would correspond to a long time period, thus impacting on the system promptness. On the contrary, the use of smaller time windows would result in the processing of a reduced number of samples, thus impairing the system effectiveness. As for the total data recording period, it should be long enough for both normal and anomalous traffic in order to perform an effective training and testing. Moreover, one of the first problems to deal with is the pre-processing of the network data and the definition of its structure so that the subsequent analysis module, based on deep learning, is more effective. More



specifically, the following characteristics need to be selected: 1) the traffic parameters (e.g., bytes, packets) to be used in the two-dimensional representation of the network status; 2) the protocol level at which data will be analyzed (e.g., IP layer, transport layer); 3) the data normalization model to guarantee that the resulting images have the same dynamic range; 4) the best domain representation of the 2D data (e.g., transform domain, time domain); 5) the type of information that needs to be represented for each traffic parameter (e.g., the traffic volume, the correlation between traffic patterns of different nodes).

4 NETWORK TRAFFIC ANALYSIS

In this section, the definition of the most effective 2D data representation is carried out. To do so, the dataset selection is performed and four 2D representation variables are identified. The suitability of these variables for anomaly detection will be assessed in **Section 5**.

4.1 Dataset

Due to the aforementioned requirements, the UGR'16 dataset has been selected (Maciá-Fernández et al., 2018). This dataset is composed of two parts: a calibration subset and a test subset. The former includes real background traffic and can be used for training, whereas the second is a combination of real background and controlled attack traffic and can be used for testing. The recording period for the calibration subset lasted 100 days, with two gaps of few hours documented in (Maciá-Fernández et al., 2018). As for the test subset, data have been recorded for approximately a month. The considered network infrastructure is shown in **Figure 3**. Data capture is performed through the netflow probes configured on the outgoing network interfaces of the border routers. The probes performed the collection of incoming and outgoing traffic.

The captured data are organized in flows and for each flow the following features are provided: timestamp of the end of a flow,

duration of the flow, source and destination IP addresses, source and destination port, protocol, flags, forwarding status, type of service, packets and bytes exchanged in the flow.

Concerning the attacks, the following classes have been simulated:

- Denial of Service (DoS):
 - DoS11: one-to-one DoS where attacker A_1 attacks the victim V_{21} ;
 - DoS53: the five attackers ($A_1 - A_5$) attack three victims. More specifically, attackers A_1 and A_2 attack the victim V_{21} , attackers A_3 and A_4 attack the victim V_{31} , and attacker A_5 attacks the victim V_{41} ;
 - DoS53a: follows the same structure as DoS53 but the attacks are sequentially executed.
- Port Scanning:
 - Scan11: one-to-one scan attack where attacker A_1 scans the victim V_{41} ;
 - Scan44: four-to-four scan attack where the attackers A_1 , A_2 , A_3 and A_4 scan the victims V_{21} , V_{11} , V_{31} and V_{41} , respectively.
- Botnet: an attack involving all the twenty victim machines is simulated.

The attacks have been simulated according to two different scheduling: a planned scheduling, in which there is no overlap between the attacks, and a random scheduling in which overlap is possible. Moreover, since the calibration set is composed of real traffic, although attack data have not been injected, real anomalies may be present. For this reason, in (Maciá-Fernández et al., 2018), a further classification has been performed to differentiate between normal and anomalous background traffic in the calibration set.

In this work, only DoS and scan attacks have been considered, and only the planned scheduling has been taken into account. The former choice is related to how the botnet traffic has been generated. More specifically, as highlighted in (Maciá-

Fernández et al., 2018), the produced traffic may not be realistic since it does not consider the effect of botnet traffic over the normal one. The authors underline that the produced traffic can be considered as sufficiently realistic for scenarios in which the influence of the botnet attack on background traffic is negligible but, in order to define the optimal 2D data representation, this attack has been currently excluded. In this work we chose to consider only the planned scheduling since the random scheduling may result in multiple simultaneous attacks. By considering the planned scheduling only we are sure that an individual attack is present for each time window, and we are able not only to analyze its effects on the 2D representation, but also the difference between the attacks. The study of multiple contemporary attacks will be the subject of future contributions.

4.2 2D Traffic Representation

The goal of this step is to obtain a 2D representation in which the attack presence is highlighted. Therefore, a traffic parameter has to be selected and a pre-processing has to be performed to produce a 2D structure in which the value of the element corresponding to the attack is significantly different from all the others. In addition, the variables associated to rows and columns have to be selected. In this work, exchanged bytes have been considered as traffic parameters and source and destination IPs have been used for indexing rows and columns. Moreover, when dealing with a 2D representation, an important aspect to define is the data structure dimension. More specifically, how data are organized in the 2D structure depends on the asset to be protected, on the threat source, and on the following processing flow. In order to define a distributed-context subsystem which can be applied in several application scenarios, the asset to be protected has been defined as a sub-network consisting of a fixed number of IP addresses. This consideration leads to the definition of the number of elements for one of the dimensions (i.e., the rows) of the 2D structure. The definition of the variable indexing the other dimension has been chosen based on how the data structure is processed. More specifically, deep learning techniques usually employ data with fixed dimensions. Therefore, in order to use the source IPs as variable which indexes the columns, an IP selection procedure was needed. However, in a realistic scenario, no assumptions can be made about the attack source, and using a number of columns equal to the number of all the possible source IPs is not feasible in practice. As a consequence, we decided to split the available source IPs in subsets of fixed dimensions. Moreover, the number of source IPs in each set has been selected to be equal to the number of monitored destination IPs to obtain a squared data structure. It is useful to highlight that this design choice allows not only to detect an anomaly, but also to identify the network IP subset from which it comes from.

In order to represent the bytes exchanged between two IP addresses, additional design choices are needed: the duration of the observation time window and how to process the bytes to define the value of a single element of the 2D structure. The time window specifies the time interval Δt (in seconds) in which data are aggregated for computing the representation variables. Starting from the timestamp associated to the first captured

sample, it is possible to define consecutive non-overlapping time windows of length Δt for obtaining a 2D representation associated to each time interval. The length of the observation time window impacts on the effectiveness of the anomaly detection method. More specifically, if it is too short it may impair the attack visibility, whereas if it is too long it reduces the detection promptness. A possible way to overcome this issue is to employ multiple windows in parallel. As for the processing, several options are available. The easiest solution is the computation of the number of bytes exchanged between a node pair. This variable will be referred to as Σ in the following. However, the choice of focusing on the traffic volume may lead to the detection of high-rate and high-volume attacks only, while losing useful information. Therefore, in order to define the representation variable for highlighting the presence of an attack, the targeted attacks have been analyzed in detail, and the features of both scan and DoS attacks are discussed in the following.

4.2.1 Scan Analysis

A port scan attack usually results in a high number of flows of similar length exchanged between the attacker and the victim. These attack features lead to the following considerations:

1. If a victim is scanned, more than one flow will be exchanged with the attacker. As a consequence, a single flow in the time window should not be related to a scanning attack. To delete the contribution of single flows we computed the sum of bytes exchanged in the time window, and subtracted the mean. More specifically, the mean has been computed as

$$\mu = \frac{1}{N} \sum_{i=1}^n b_i \quad (1)$$

where N indicates the number of samples in the time window, i.e., the number of flows for the considered dataset, and b_i is the number of bytes exchanged within each flow. In this way, single flows will result in a 0 value, whereas long runs of flows will be only slightly affected by the subtraction. The resulting variable will be referred to as Σ_μ in the following.

2. If a victim is scanned, the flows are often equal in length thus resulting in a low standard deviation. In order to highlight this feature it is possible to divide Σ_μ by the standard deviation of the flow bytes. More specifically, the standard deviation has been computed as

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^n |b_i - \mu|} \quad (2)$$

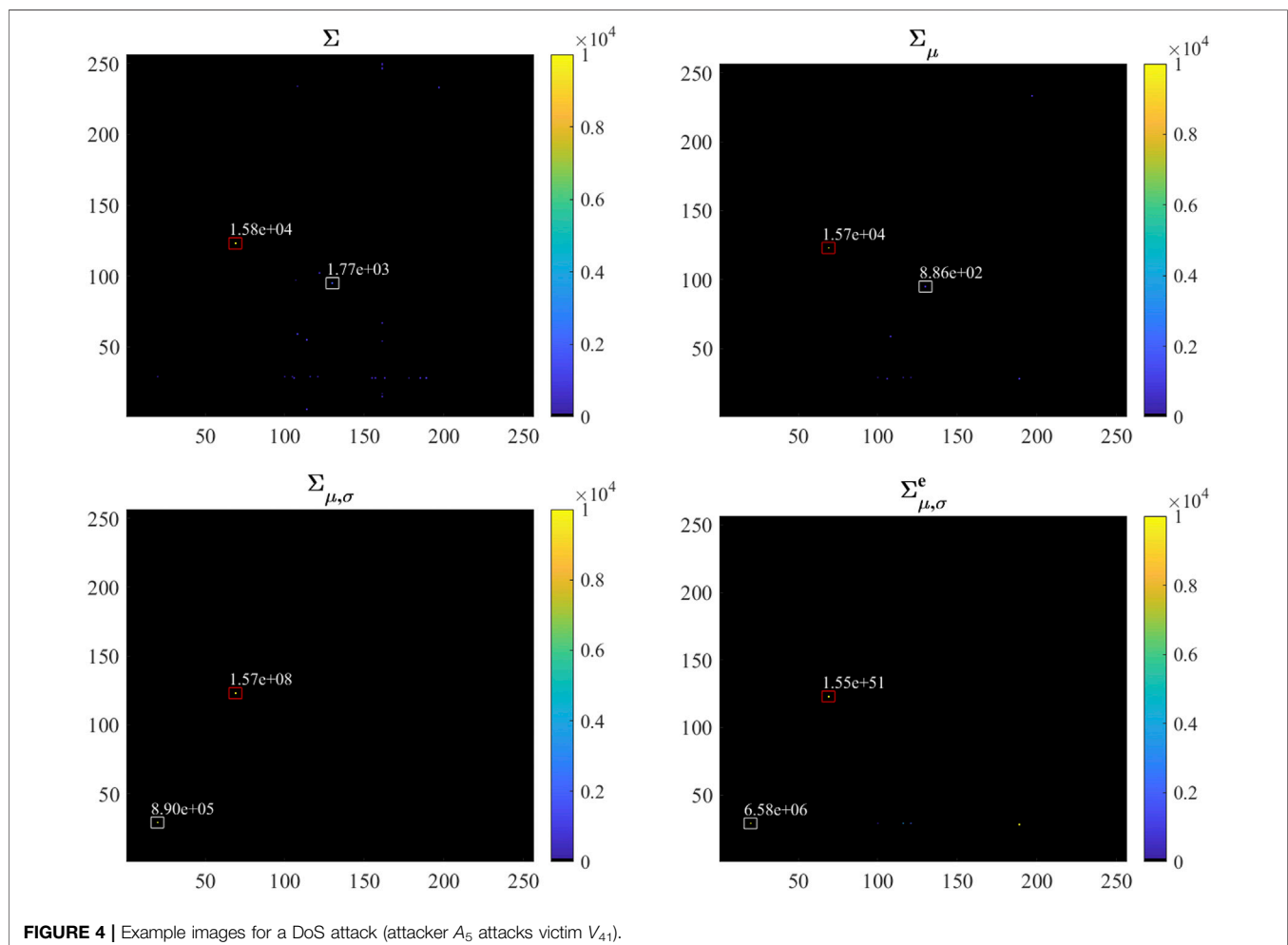
where N indicates the number of samples in the time window and b_i is the number of bytes exchanged within each flow. To avoid a denominator equal to 0, we divided Σ_μ by the standard deviation summed to a small quantity (i.e., 10^{-4}). This variable will be referred to as $\Sigma_{\mu,\sigma}$ in the following.

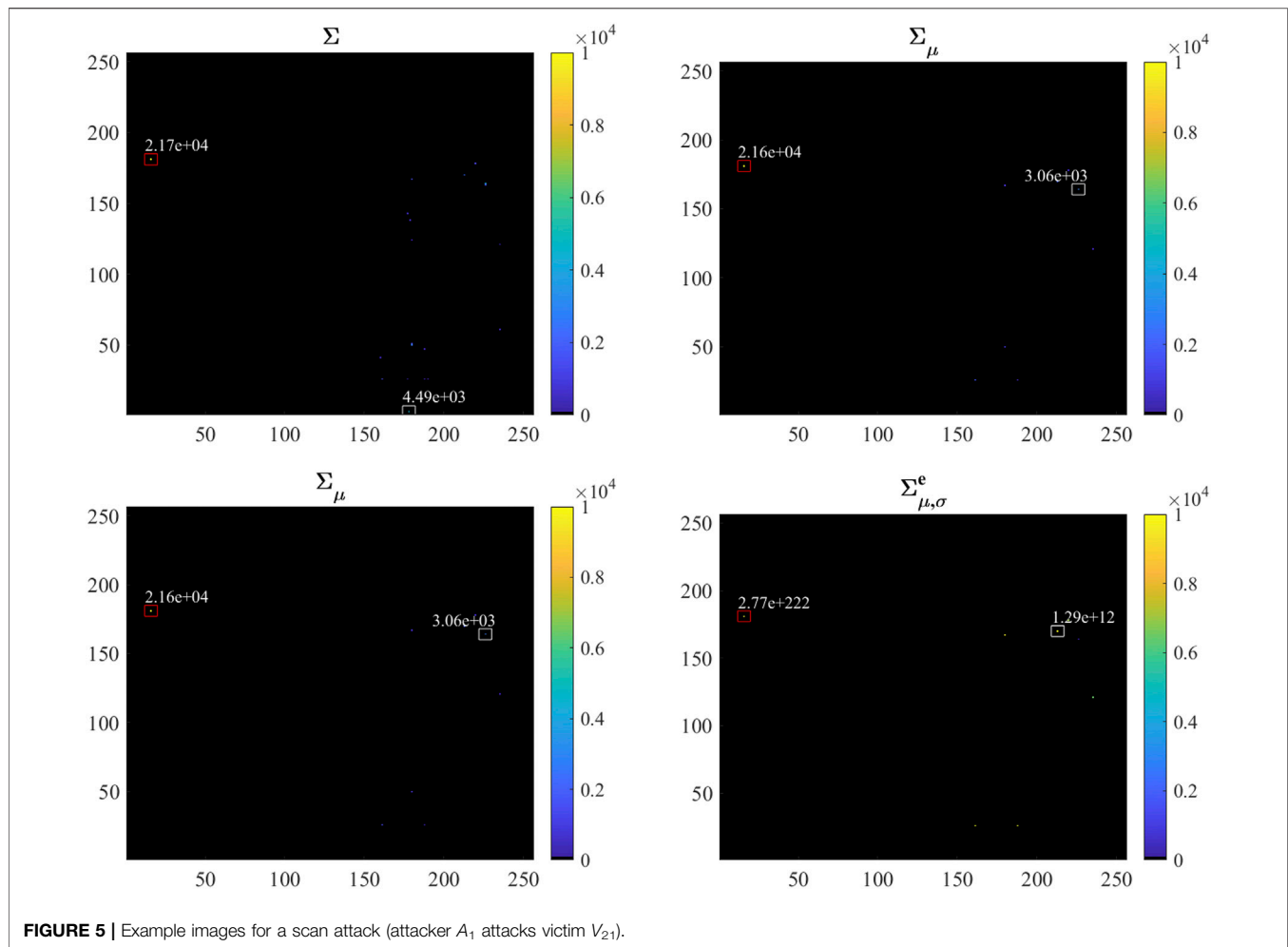
3. The variable $\Sigma_{\mu,\sigma}$ has been defined based on the attack features. However, also normal traffic consisting of a small number of flows may have a small standard deviation. To account for this issue, another attack feature has been employed: the high number of flows. More specifically, the variable $\Sigma_{\mu,\sigma}$ has been multiplied by a function of the number of flows and, to compensate for standard deviations tending to zero, the exponential function has been considered. The resulting variable will be indicated as $\Sigma_{\mu,\sigma}^e$.

4.2.2 DoS Analysis

In order to assess if the proposed variables Σ , Σ_μ , $\Sigma_{\mu,\sigma}$ and $\Sigma_{\mu,\sigma}^e$ are suitable for highlighting anomalous network behaviors associated with the DoS attack, we performed an analysis to identify the relevant features for detecting this type of attack. To achieve this goal, we rely on the studies performed in (Tang et al., 2020) and (Aamir et al., 2021). More in detail, in (Tang et al., 2020), the authors propose a low-rate DoS detection method which exploits a set of features selected based on the correlation score between features and data labels. Among the analyzed features, the ones providing the highest correlation score are: the total number of packets, the UDP ratio (proportion of the UDP traffic relative to

the total network traffic), the average of the traffic sequence, the variance of the traffic sequence, the covariance of the traffic sequence, the UDP maximum, the UDP and TCP range, the variation coefficients, and the mean absolute differences. Additional details can be found in (Tang et al., 2020). A similar study has been performed in (Aamir et al., 2021), where several machine learning approaches for detecting DoS and port scanning are analyzed. More in detail, the authors performed feature selection by analyzing the correlation coefficient scores with respect to the dependent (target variable) and, according to the study performed in (Taylor, 1990), a correlation coefficient smaller or equal to 0.35 indicates that the associated feature does not provide useful information. Based on this assumption, the most significant features among the ones analyzed in (Aamir et al., 2021) are: the maximum packet length, the minimum packet length, the mean packet length, the packet length standard deviation and variance, and the average segment size. From these results, it is possible to infer that the features considered for defining the representation variables in **Section 4.2.1**, namely the volume of traffic, the traffic mean, and its standard deviation, are relevant also for highlighting the DoS attack presence. Moreover, since we





are analyzing traffic at IP level, no information concerning the transport level layer has been included in this work although it could be exploited for future contributions. It is useful to underline that the main difference with respect to the methods presented in (Tang et al., 2020; Aamir et al., 2021) is that in our work the features are combined to provide a single value for the elements of the 2D data structure.

5 REPRESENTATION VARIABLE ASSESSMENT

In order to study if the considered variables, namely Σ , Σ_μ , $\Sigma_{\mu,\sigma}$ and $\Sigma_{\mu,\sigma}^e$, allow to define a 2D representation which highlights the presence of the attacks, a time window of 1 s has been selected, and a monitored network of 256 IPs has been considered. More specifically, for each attack, tests have been performed considering as monitored networks the one in which the victims are placed. Moreover, the available source IPs have

been split in subsets of 256 IPs to obtain squared matrices. Examples of the images both for DoS and scan attacks are provided in **Figures 4, 5**. More specifically, the attack point is surrounded by a red box, whereas the largest among the background pixels is surrounded by a white box. Moreover, the corresponding pixel values are written near the boxes.

Figures 4, 5 show that the use of the variable Σ fails to effectively highlight the attack for two main reasons. The former is that several non-zero values are present in the generated image, and the latter is that background and attack-related pixels show a similar order of magnitude thus making it difficult to detect the presence of the attack. The former issue can be mitigated with the use of the variable Σ_μ which allows the reduction of the number of non-zero values by eliminating the single-flow samples as described in **Section 4.2.1**. However in this case, due to the subtraction of the mean, the value of the pixel corresponding to the attack becomes smaller so that the separation gap between pixels corresponding to the attack and the ones associated to background traffic is further reduced. This

TABLE 1 | Percentage of samples for which the attack value is larger than the maximum of the non attacked samples (DoS).

Attack	Variable			
	Σ	Σ_{μ}	$\Sigma_{\mu,\sigma}$	$\Sigma_{\mu,\sigma}^e$
Dos 11: $A_1 - V_{21}$	71%	98.5%	98.5%	98.5%
Dos 53: $A_1 - V_{21}$	30.8%	72.6%	96.6%	100%
Dos 53: $A_2 - V_{21}$	30.8%	72.6%	96.6%	100%
Dos 53: $A_3 - V_{31}$	100%	100%	100%	100%
Dos 53: $A_4 - V_{31}$	100%	100%	100%	100%
Dos 53: $A_5 - V_{41}$	89.7%	96.6%	97.9%	100%

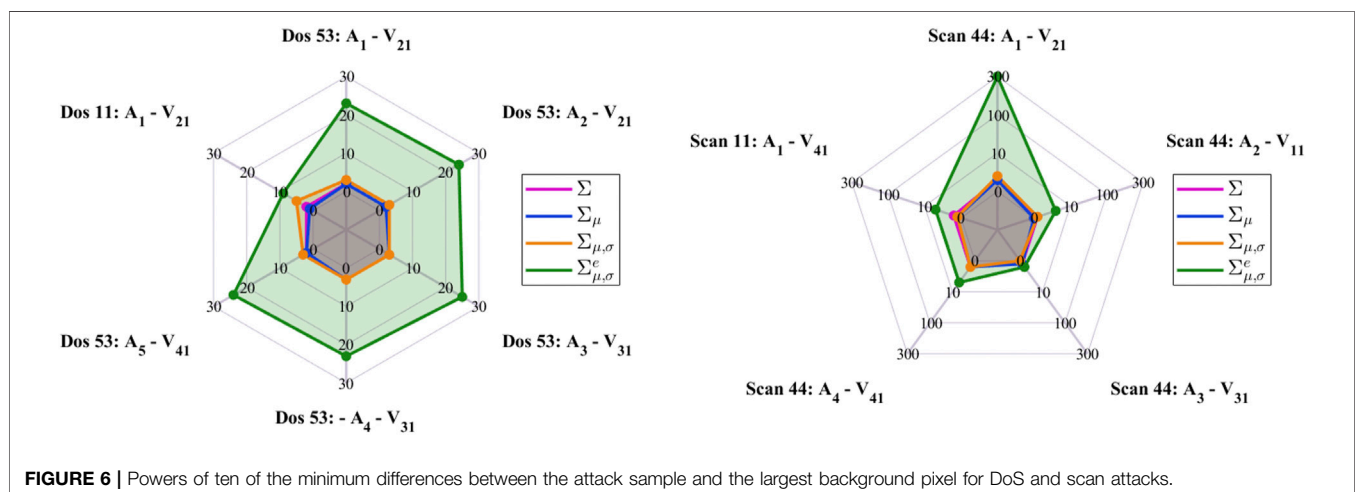
TABLE 2 | Percentage of samples for which the attack value is larger than the maximum of the non attacked samples (scan).

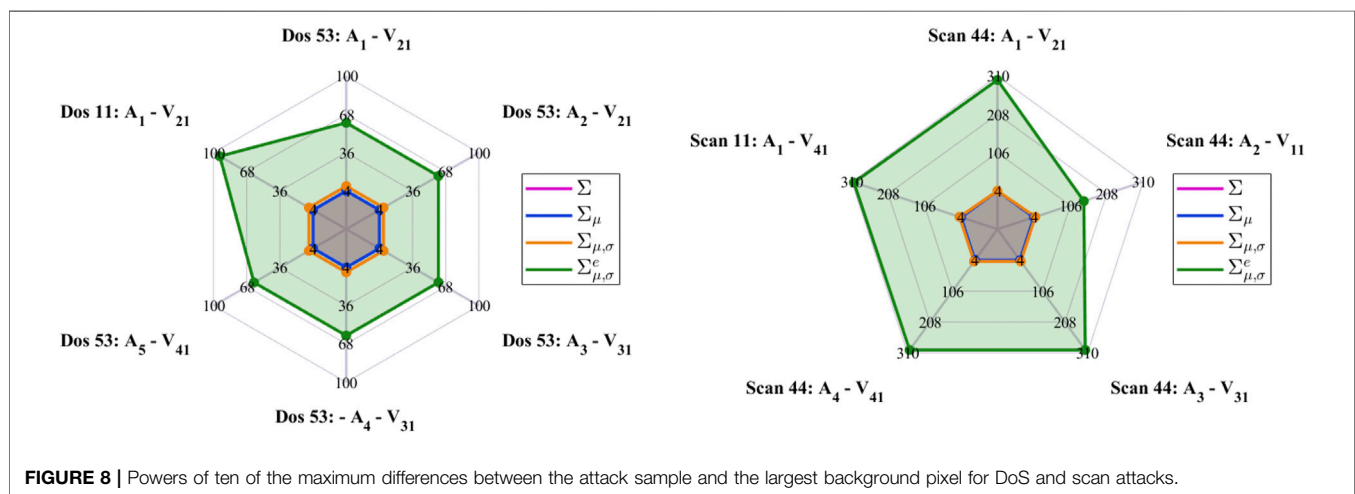
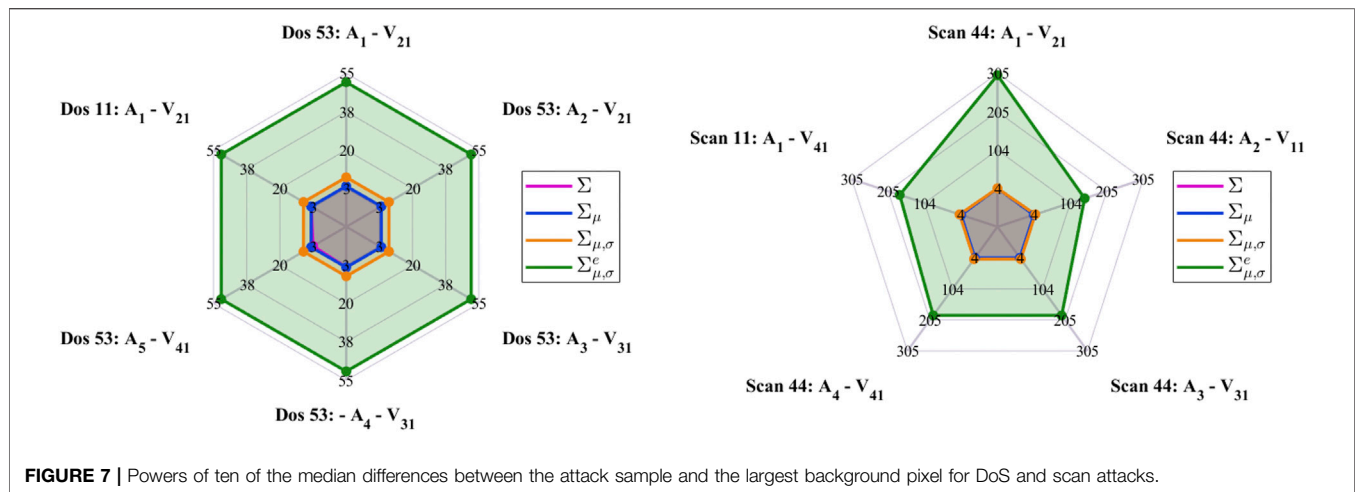
Attack	Variable			
	Σ	Σ_{μ}	$\Sigma_{\mu,\sigma}$	$\Sigma_{\mu,\sigma}^e$
Scan 11: $A_1 - V_{41}$	65.2%	78.8%	65.2%	93.9%
Scan 44: $A_1 - V_{21}$	57.1%	76.2%	88.1%	95.2%
Scan 44: $A_2 - V_{11}$	93.8%	93.8%	25%	93.8%
Scan 44: $A_3 - V_{31}$	84.4%	90.6%	90.6%	100%
Scan 44: $A_4 - V_{41}$	67.7%	76.4%	73.5%	97.1%

phenomenon is avoided using the variable $\Sigma_{\mu,\sigma}$ since in this case, dividing by the standard deviation, the value of the pixels corresponding to the attack becomes larger. However, as shown in the figures, also background pixels undergo a huge increase so that the separation between attack and background pixel values may be insufficient for detecting the attack. This behavior is due to the fact that, given the limited extension of the time window, a reduced number of flows may be present for background traffic. The corresponding standard deviation, as a consequence, will be small, thus causing the growth of $\Sigma_{\mu,\sigma}$. This effect can be mitigated by using the variable $\Sigma_{\mu,\sigma}^e$ which, due to the multiplication for a function of the number of flows, produces

large values when the number of flows increases, as occurs in presence of attacks. **Figures 4, 5** show that this variable allows to significantly increase the attack pixel values, while enlarging the gap between attack and background pixels. Although with respect to $\Sigma_{\mu,\sigma}^e$ some background pixel values increase, their entity is several orders of magnitude smaller than the attacked pixel values. In addition, **Figures 4, 5** show that the variable $\Sigma_{\mu,\sigma}^e$ results in a clear separation between DoS and scan attacks since the pixel values associated to the two types of attacks are significantly different. This could allow not only to perform attack detection, but also attack classification. The reason for this phenomenon can be found in the variable definition procedure.

In order to provide quantitative results concerning the suitability of the different variables for highlighting the attack presence, a single day of recording has been taken as example and the number of times for which the attack value is larger than the corresponding largest background sample has been computed. The results, expressed as percentages, are reported in **Tables 1, 2** for the DoS and for the scan attack, respectively. From these results, it is clear that the variable Σ is not reliable for highlighting the presence of an attack with respect to background traffic. As for the others, usually $\Sigma_{\mu,\sigma}$ performs better than Σ_{μ} . In some cases, however, $\Sigma_{\mu,\sigma}$ performs worse than Σ and Σ_{μ} . This is due to the fact that, as already mentioned, while $\Sigma_{\mu,\sigma}$ increases the attack pixel value, it also enlarges background traffic pixels without causing an enlargement of the gap between them. At last, the tables show that $\Sigma_{\mu,\sigma}^e$ usually achieves the best performances. The results shown in **Tables 1, 2** allow to evaluate only the percentage of times for which the attack value is larger than the maximum of the non-attacked samples. In order to be highlighted, however, the attack value should be significantly different (in this case larger) than the background samples. To evaluate the capability of the variables to achieve this goal, for the selected day, the difference between the attack value and the maximum





background sample has been computed. **Figures 6–8** show the power of ten of the minimum, the median and the maximum of the computed differences for DoS and scan attacks. These figures clearly show that $\Sigma_{\mu,\sigma}^e$ allows to create a significant difference between attacked and background pixels, while the other variables fail to do so. Therefore, $\Sigma_{\mu,\sigma}^e$ is eligible as representation variable for the 2D data structure definition.

6 CONCLUSION

In this contribution, a context-based security framework has been presented. It exploits information gathered both from local and distributed contexts for detecting the presence of an anomaly and estimating its cause. Through the joint processing of the two types of contexts it is possible to evaluate the impact of the detected anomaly on the system,

thus allowing the selection of the most effective mitigation strategies, also considering the associated costs.

In this work we applied the proposed framework to the CPS scenario focusing on the distributed context analysis. To this aim, a 2D representation of network traffic for anomaly detection has been investigated and a representation variable has been selected. Preliminary results demonstrate its suitability to highlight the presence of attacks. The full implementation of the distributed-context building block and its integration in the overall framework will be the subject of future contributions.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <https://nesg.ugr.es/nesg-ugr16/>.

AUTHOR CONTRIBUTIONS

SB, MC, and FB contributed to conception and design of the study. SB performed the software implementation of the system and wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

REFERENCES

- Aamir, M., Rizvi, S. S. H., Hashmani, M. A., Zubair, M., and Usman, J. A. (2021). Machine Learning Classification of Port Scanning and DDoS Attacks: A Comparative Analysis. *Mehran Univ. Res. J. Eng. Technol.* 40, 215–229. doi:10.22581/muet1982.2101.19
- Alagar, V., Alsaig, A., Ormandjiva, O., and Wan, K. (2018). “Context-Based Security and Privacy for Healthcare IoT,” in 2018 IEEE International Conference on Smart Internet of Things (SmartIoT), Xi’an, China, 122. doi:10.1109/SmartIoT.2018.00-14
- Dsouza, J., Elezabeth, L., Mishra, V. P., and Jain, R. (2019). “Security in Cyber-Physical Systems,” in 2019 Amity International Conference on Artificial Intelligence (AICAI), Dubai, United Arab Emirates, 840–844. doi:10.1109/AICAI.2019.8701411
- Ehsani-Besheli, F., and Zarandi, H. R. (2018). “Context-Aware Anomaly Detection in Embedded Systems,” in *Advances in Dependability Engineering of Complex Systems*. Editors W. Zamojski, J. Mazurkiewicz, J. Sugier, T. Walkowiak, and J. Kacprzyk (Cham: Springer International Publishing, 151–165. doi:10.1007/978-3-319-59415-6_15
- Fernandes, G., Rodrigues, J. J. P. C., and Proença, M. L. (2015). Autonomous Profile-Based Anomaly Detection System Using Principal Component Analysis and Flow Analysis. *Appl. Soft Comput.* 34, 513–525. doi:10.1016/j.asoc.2015.05.019
- Ghosh, T., Roy, A., Misra, S., and Raghuvanshi, N. S. (2021). CASE: A Context-Aware Security Scheme for Preserving Data Privacy in IoT-Enabled Society 5.0. *IEEE Internet Things J.* 1, 1. doi:10.1109/JIOT.2021.3101115
- Ivanov, R., Weimer, J., and Lee, I. (2018). “Towards Context-Aware Cyber-Physical Systems,” in 2018 IEEE Workshop on Monitoring and Testing of Cyber-Physical Systems (MT-CPS), Porto, Portugal, 10. doi:10.1109/MT-CPS.2018.00012
- Kim, H., Kang, I., and Bahk, S. (2004). Real-time Visualization of Network Attacks on High-Speed Links. *IEEE Netw.* 18, 30–39. doi:10.1109/MNET.2004.1337733
- Kim, S. S., and Reddy, A. L. N. (2005a). “Modeling Network Traffic as Images,” in IEEE International Conference on Communications, 2005, Seoul, South Korea (ICC 2005), 168–172. doi:10.1109/ICC.2005.1494341
- Kim, S. S., and Reddy, A. L. N. (2005b). “A Study of Analyzing Network Traffic as Images in Real-Time,” in Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies, Miami, FL, 2056–2067. doi:10.1109/INFCOM.2005.1498482
- Luo, Y., Xiao, Y., Cheng, L., Peng, G., and Yao, D. (2021). Deep Learning-Based Anomaly Detection in Cyber-Physical Systems. *ACM Comput. Surv.* 54, 1–36. doi:10.1145/3453155
- Maciá-Fernández, G., Camacho, J., Magán-Carrión, R., García-Teodoro, P., and Thérón, R. (2018). UGR’16: A New Dataset for the Evaluation of Cyclostationarity-Based Network IDSs. *Comput. Security* 73, 411–424. doi:10.1016/j.cose.2017.11.004
- Mohammadpour, L., Ling, T. C., Liew, C. S., and Chong, C. Y. (2018). “A Convolutional Neural Network for Network Intrusion Detection System,” in Proceedings of the Asia-Pacific Advanced Network, 50–55.
- Nataraj, L., Karthikeyan, S., Jacob, G., and Manjunath, B. S. (2011). “Malware Images,” in Proceedings of the 8th International Symposium on Visualization for Cyber Security, Pittsburgh, PA (New York, NY, USA: Association for Computing Machinery). doi:10.1145/2016904.2016908
- Pang, G., Shen, C., Cao, L., and Hengel, A. V. D. (2021). Deep Learning for Anomaly Detection. *ACM Comput. Surv.* 54, 1–38. doi:10.1145/3439950
- Park, S.-T., Li, G., and Hong, J.-C. (2020). A Study on Smart Factory-Based Ambient Intelligence Context-Aware Intrusion Detection System Using Machine Learning. *J. Ambient Intell. Hum. Comput.* 11, 1405–1412. doi:10.1007/s12652-018-0998-6
- Roukounaki, A., Efremidis, S., Soldatos, J., Neises, J., Walloschke, T., and Kefalakis, N. (2019). “Scalable and Configurable End-To-End Collection and Analysis of IoT Security Data : Towards End-To-End Security in IoT Systems,” in *Global IoT Summit (GloTS)*. Aarhus, Denmark: IEEE ComSoc. doi:10.1109/GIOTS.2019.8766407
- Sharaf Dabbagh, Y., and Saad, W. (2019). Authentication of Wireless Devices in the Internet of Things: Learning and Environmental Effects. *IEEE Internet Things J.* 6, 6692–6705. doi:10.1109/JIOT.2019.2910233
- Sylla, T., Chalouf, M. A., Krief, F., and Samaké, K. (2020). “Towards a Context-Aware Security and Privacy as a Service in the Internet of Things,” in 13th IFIP International Conference on Information Security Theory and Practice (WISTP), Paris, France. Editors M. Laurent and T. Giannetsos (Paris, France: Springer International Publishing/LNCS-12024 of Information Security Theory and Practice), 240–252. doi:10.1007/978-3-030-41702-4_15
- Taheri, S., Salem, M., and Yuan, J.-S. (2018). Leveraging Image Representation of Network Traffic Data and Transfer Learning in Botnet Detection. *Bdcc* 2, 37. doi:10.3390/bdcc2040037
- Tang, D., Tang, L., Dai, R., Chen, J., Li, X., and Rodrigues, J. J. P. C. (2020). MF-adaboost: LDoS Attack Detection Based on Multi-Features and Improved Adaboost. *Future Generation Comput. Syst.* 106, 347–359. doi:10.1016/j.future.2019.12.034
- Taylor, R. (1990). Interpretation of the Correlation Coefficient: A Basic Review. *J. Diagn. Med. Sonography* 6, 35–39. doi:10.1177/875647939000600106
- Vasan, D., Alazab, M., Wassan, S., Safaei, B., and Zheng, Q. (2020). Image-Based Malware Classification Using Ensemble of CNN Architectures (IMCEC). *Comput. Security* 92, 101748. doi:10.1016/j.cose.2020.101748
- Venkatraman, S., Alazab, M., and Vinayakumar, R. (2019). A Hybrid Deep Learning Image-Based Analysis for Effective Malware Detection. *J. Inf. Security Appl.* 47, 377–389. doi:10.1016/j.jisa.2019.06.006
- Wang, E. K., Ye, Y., Xu, X., Yiu, S. M., Hui, L. C. K., and Chow, K. P. (2010). “Security Issues and Challenges for Cyber Physical System,” in 2010 IEEE/ACM Int’l Conference on Green Computing and Communications Int’l Conference on Cyber, Physical and Social Computing, Hangzhou, China, 733–738. doi:10.1109/GreenCom-CPSCCom.2010.36
- Wang, W., Zhu, M., Zeng, X., Ye, X., and Sheng, Y. (2017). “Malware Traffic Classification Using Convolutional Neural Network for Representation Learning,” in 2017 International Conference on Information Networking (ICOIN), Da Nang, Vietnam (IEEE), 712–717. doi:10.1109/icoi.2017.7899588
- Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., et al. (2019). A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data. *Aaai* 33, 1409–1416. doi:10.1609/aaai.v33i01.33011409

ACKNOWLEDGMENTS

The research presented in this paper was partially funded by the project “ISEEYOO: AI-based Network Anomaly Detection for CPS exploiting 2D data representation” within the University of Padova funding framework “SID research grants.”

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Baldoni, Carli and Battisti. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Beamspace ESPRIT for mmWave Channel Sensing: Performance Analysis and Beamformer Design

Sina Shahsavari, Pulak Sarangi and Piya Pal*

University of California, San Diego, La Jolla, CA, United States

In this paper, we consider the beamspace ESPRIT algorithm for Millimeter-Wave (mmWave) channel sensing. We provide a non-asymptotic analysis of the beamspace ESPRIT algorithm. We derive a deterministic upper bound for the matching distance error between the true angle of arrival (AoA) of the channel paths and the estimated AoA considering a bounded noise model. Additionally, we leverage the insight obtained from our theoretical analysis to propose a novel max-min criterion for beamformer design which can enhance the performance of mmWave channel estimation algorithms, including beamspace ESPRIT. We consider a family of multi-resolution beamformers which can be implemented using phase shifters and introduce a design scheme for the optimal beamformers from this family with respect to the proposed max-min criteria. We can guarantee a minimum beamforming gain uniformly over a region of possible multipath directions, which can lead to more robust channel estimation. We provide several numerical experiments to verify our theoretical claims and demonstrate the superior performance of the proposed beamformers compared to existing beamformer design criteria.

Keywords: millimeter wave, beamspace ESPRIT, massive MIMO, channel estimation, beamformer design

OPEN ACCESS

Edited by:

Maria Sabrina Greco,
University of Pisa, Italy

Reviewed by:

Jiang Zhu,
Zhejiang University, China
Jue Wang,
Nantong University, China

*Correspondence:

Piya Pal
pipal@eng.ucsd.edu

Specialty section:

This article was submitted to
Signal Processing for
Communications,
a section of the journal
Frontiers in Signal Processing

Received: 23 November 2021

Accepted: 31 December 2021

Published: 11 February 2022

Citation:

Shahsavari S, Sarangi P and Pal P
(2022) Beamspace ESPRIT for
mmWave Channel Sensing:
Performance Analysis and
Beamformer Design.
Front. Sig. Proc. 1:820617.
doi: 10.3389/frsip.2021.820617

1 INTRODUCTION

Millimeter wave (mmWave) communication has emerged as a key technology for the next generation of wireless communication systems due to an abundance of spectrum availability in the mmWave bands, and the higher data rates enabled by larger bandwidths (Bai and Heath, 2014).¹ However, at mmWave frequencies, the wireless channel is spatially sparse and suffers from severe path loss. To ensure reliable communication, it becomes essential to perform beamforming in order to combat this path loss. Due to the large number of antennas in a mmWave system, it is impractical to implement a fully digital beamforming scheme with a dedicated radio frequency (RF) chain for every antenna, which would incur high power consumption and cost. In order to overcome this challenge, mmWave systems typically utilize either analog (Junyi Wang et al., 2009; Hur et al., 2013) or hybrid beamforming approaches with a reduced number of RF chains (Alkhateeb et al., 2014a; Han et al., 2015). Therefore, the problem of mmWave channel estimation becomes challenging, since a high-dimensional channel matrix (whose size is given by the large number of antennas) needs to be

¹This work was supported in part by the Office of Naval Research under Grants ONR N00014-19-1-2256 and ONR N00014-19-1-2227, and in part by the National Science Foundation under Grant NSF CAREER ECCS 1700 506.

estimated from only low-dimensional measurements acquired at the output of a reduced number of RF chains, especially with limited pilot overhead.

MmWave channel sensing has emerged as an active area of research, with many algorithms having been developed for both flat-fading (Alkhateeb et al., 2014b; Bogale et al., 2015; Lee et al., 2016; Méndez-Rial et al., 2016) and frequency-selective channels (Alkhateeb and Heath, 2016; Gao et al., 2016; González-Coma et al., 2018; Rodríguez-Fernández et al., 2018). Under the flat-fading model, compressed sensing based techniques that leverage the sparse nature of the channel, have been proposed (Park and Heath, 2018). Recently, adaptive schemes have also been developed for estimating the channel paths by employing hierarchical multi-resolution beamforming codebooks (Alkhateeb et al., 2014b). However, such techniques assume the multipath angles to be on a grid, which can potentially introduce bias (grid-offset). The mmWave channel model shares many similarities with the measurement models arising in array signal processing, which enables the application of *super-resolution* AOA estimation techniques such as multiple signal classification (MUSIC) and estimating signal parameters via rotational invariance techniques (ESPRIT) for mmWave channel estimation (Schmidt, 1986; Roy and Kailath, 1989). Suitable variants of these algorithms have been developed in the beamspace which leverage the structure of beamformers to enable super-resolution estimation of the AOAs (Guanghan Xu et al., 1994; Zoltowski et al., 1996; Li et al., 2020; Sarangi et al., 2020). Finally, both sparsity-based techniques (Gao et al., 2016; Park and Heath, 2018; Rodríguez-Fernández et al., 2018) and subspace based angular estimation algorithms (Guo et al., 2017; Liao et al., 2017; Zhang and Haardt, 2017; Park et al., 2019; Zhang et al., 2021) have been extended for frequency-selective channels.

In this work, we focus on the ESPRIT algorithm for channel estimation (Liao et al., 2017; Zhang and Haardt, 2017; Wen et al., 2018; Rakhimov et al., 2019; Ma et al., 2020; Zhang et al., 2021). In recent times, several works (Zhang and Haardt, 2017; Rakhimov et al., 2019; Zhang et al., 2021) have considered DFT-based beamspace ESPRIT, inspired by earlier works in array processing (Guanghan Xu et al., 1994; Haardt and Nosske, 1995; Mathews et al., 1996; Zoltowski et al., 1996). However, the large number of antennas in mmWave systems lead to very narrow DFT beams (Ma et al., 2020). To get a wide spatial coverage, a large number of RF chains are required, which may not be practical. A different beamspace ESPRIT is proposed in (Liao et al., 2017) where beamformers are designed to ensure that the low-dimensional beamspace measurements share the same shift-invariance structure as the high dimensional channel. However, in order to realize this, approximately half of the antennas need to be turned off. This strategy may suffer from a reduction in total transmitted power, and inability to perform high-resolution channel estimation (Ma et al., 2020). Recently in (Ma et al., 2020), Li et al. proposed a beamspace ESPRIT scheme which is applicable for any choice of beamformer, that satisfies some mild rank constraints. Unlike the aforementioned variants of ESPRIT, only one antenna needs to be turned off at a time, which results in a negligible drop in transmitted power and signal coverage.

Despite their wide use in mmWave channel sensing, a rigorous non-asymptotic analysis for beamspace ESPRIT is currently not available. Existing performance analysis are either asymptotic in the number of snapshots (Guanghan Xu et al., 1994; Mathews et al., 1996), or based on perturbation analysis where certain higher-order terms are ignored (Roemer et al., 2014; Steinwandt et al., 2017). Recently, in (Li et al., 2020) the authors provided a rigorous theoretical analysis of the single-snapshot antenna space ESPRIT algorithm. In this work, we will extend their analysis to multi-snapshot beamspace ESPRIT. Beyond mathematical interest, a key motivation for such analysis is to develop insights on how the choice of beamformer controls the error bound. The choice of the analog/hybrid beamformer indeed determines the quality of channel estimation. Therefore, an important consideration for beamspace algorithms involves developing suitable analog/hybrid beamforming schemes that ensure reliable channel estimation. It should be noted that typically beamformer design is performed after the channel state information is available. However, while performing channel estimation using beamspace algorithms, the channel information is not available apriori and the beamformer must be designed to ensure robust performance uniformly across a variety of channel configurations.

DFT beamformers are a common choice for analog beamforming since they automatically satisfy the constant modulus constraint, and are easy to implement using purely RF (Analog) components (Méndez-Rial et al., 2016). However, the spatial coverage obtained using DFT beamformers is limited, especially with few RF chains (Li et al., 2020). Several alternate beamformer designs have been proposed that aim to approximately ensure constant gain across a sector of interest. Approximating ideal filters using only phase shifters or hybrid architectures results in optimization problems with non-convex constraints. A variety of heuristics/iterative techniques have been proposed to solve these problems, using Orthogonal matching pursuit (OMP) (Venugopal et al., 2017), alternating minimization (Yu et al., 2016), fast search-based techniques (Chen and Qi, 2018). An outstanding limitation of these techniques is that they cannot provide guarantees on the worst-case beamforming gain over the sector of interest that is finally achieved by the design. In particular, the gain at several points in the region of interest can significantly drop below the desired level. This can degrade the performance of channel estimation algorithms for several channel realizations. In this paper, we will develop beamformer designs based on alternative criteria to overcome this drawback.

2 OUR CONTRIBUTIONS

Our contributions are twofold (i) non-asymptotic analysis of the beamspace ESPRIT algorithm, and (ii) design of beamformers that can enhance the performance of mmWave channel estimation algorithms (including beamspace ESPRIT). We first provide a non-asymptotic analysis of beamspace ESPRIT algorithm in (Ma et al., 2020), tailored to the flat-fading channel model. Inspired by the analysis of Single Snapshot

element-space ESPRIT in (Li et al., 2020), we obtain error bounds on the matching distance error between the true angle of arrival (AoA) of the channel paths and the estimated AoA. Our error analysis is non-asymptotic in the number of snapshots, does not require any statistical assumption on the noise distributions, and the error bounds are applicable for any beamformer satisfying suitable rank constraints. Furthermore, the analysis reveals that the error bounds are controlled by the smallest singular value of a suitable matrix which is shaped by the beamformer and the AoAs. We leverage this insight from our theoretical analysis to propose a novel max-min criterion for beamforming, with the goal of boosting the performance of beamspace ESPRIT. We consider a family of multi-resolution beamformers, that exploits the geometric coupling between the antenna array manifold and the beamformer. Our design can guarantee a minimum beamforming gain uniformly over a region of possible multipath directions and can be implemented with phase shifters (analog-only implementation).

3 MEASUREMENT MODEL

We consider a single user mmWave uplink system consisting of a single-antenna Mobile Station (MS), and a Base Station (BS) equipped with $M > 1$ antennas. We assume that the BS antennas are arranged in the form of a large Uniform Linear Array (ULA) with an inter-antenna spacing of $\lambda/2$, where λ denotes the carrier wavelength. It is well-known that mmWave channels exhibit sparse scattering, where each scatterer is often assumed to contribute to a single channel path (Raghavan and Sayeed, 2010; Ayach et al., 2014; Alkhateeb et al., 2014b). Based on this geometric model (Alkhateeb et al., 2014a; Alkhateeb et al., 2014b; Park and Heath, 2018), we consider a channel with S scattering paths, with $\theta_s \in [0, \pi]$ denoting the angle of arrival (AoA) of the s th path between the BS and the MS. Assuming that the AoAs remain unchanged during the training period, the uplink channel at the t th snapshot is given by (Park and Heath, 2018)

$$\mathbf{h}_t = \sum_{s=1}^S x_{s,t} \mathbf{a}(f_s), \quad t = 1, 2, \dots, T \quad (1)$$

Here T denotes the number of time snapshots in the training period, and $x_{s,t}$ represents the (time-varying) gain of the s th path. The array response vector (or steering vector) associated with the s th channel path is given by

$$[\mathbf{a}(f_s)]_m = e^{-j\pi m f_s}, \quad m = 0, \dots, M-1$$

where $f_s := \sin(\theta_s)$ denotes the spatial frequency determined by the AoA θ_s . Notice that, Eq. 1 corresponds to a flat-fading channel model, which is of interest in this paper.² We further consider a

low-mobility scenario where the AoA's do not change over the training period T (although the path gains can change).

Let $\mathcal{F} := \{f_i\}_{i=1}^S$ be the set of all spatial frequencies. The received signal at the physical array is given by

$$\mathbf{r}_t = \mathbf{A}(\mathcal{F})\mathbf{x}_t s_t + \mathbf{n}_t \quad t = 1, \dots, T \quad (2)$$

Here s_t represents the (known) transmitted pilot signal,³ $\mathbf{x}_t = [x_{1,t}, x_{2,t}, \dots, x_{S,t}]^T$, $\mathbf{A}(\mathcal{F}) = [\mathbf{a}(f_1), \dots, \mathbf{a}(f_S)] \in \mathbb{C}^{M \times S}$ is the antenna array manifold, and $\mathbf{n}_t \in \mathbb{C}^M$ represents the channel noise at time t . Since s_t is known, without loss of generality, we assume that $s_t = 1$ for the entire training duration (Haghighatshoar and Caire, 2016; Park and Heath, 2018).

In mmWave systems, the number of deployed antennas is very large, and a dedicated RF chain for every antenna significantly increases the hardware complexity and power consumption. Therefore, in order to reduce the number of RF chains, the signals received at the antennas are linearly combined in the analog domain using a network of analog beamformers, where the number of beamformers is given by the number of RF chains. Due to a limited number of RF chains, the measurement at the output of the RF chains is a low-dimensional projection of the signal received at the antennas. In this work, we assume that the BS is equipped with $N < M$ RF chains. Let $\mathbf{W} \in \mathbb{R}^{M \times N}$ be an analog beamforming matrix, that performs a linear combination of the received signal \mathbf{r}_t to obtain a compressed signal \mathbf{y}_t . It is typically realized using analog circuitry, such as switches or phase shifters. The measurements at the output of the RF chains is given by

$$\mathbf{y}_t = \mathbf{W}^H \mathbf{r}_t = \mathbf{W}^H \mathbf{A}(\mathcal{F})\mathbf{x}_t + \mathbf{W}^H \mathbf{n}_t, \quad t = 1, \dots, T \quad (3)$$

Denoting $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T] \in \mathbb{C}^{N \times T}$, we have

$$\mathbf{Y} = \mathbf{W}^H \mathbf{A}(\mathcal{F})\mathbf{X} + \mathbf{W}^H \mathbf{N} \quad (4)$$

where $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_T]$, $\mathbf{N} = [\mathbf{n}_1, \dots, \mathbf{n}_T]$. Given $\mathbf{Y} \in \mathbb{C}^{N \times T}$, our objective is to estimate the mmWave channel Eq. 1, which is equivalent to estimating f_i and \mathbf{x}_i , $i = 1, \dots, S$.

4 REVIEW OF BEAMSPACE ESPRIT FOR MMWAVE CHANNEL ESTIMATION

In recent times, there has been a renewed interest in utilizing classical subspace-based techniques from array signal processing for mmWave channel estimation, due to similarities between the two measurement models at mmWave frequencies (Guo et al., 2017; Liao et al., 2017; Zhang and Haardt, 2017; Ma et al., 2020). An obvious advantage of these subspace-based algorithms is that they enable “gridless super-resolution” estimation of the AoAs that grid-based sparse techniques fail to achieve. Specifically, variants of subspace algorithms in the beamspace have been proposed that can produce high-resolution estimates of channel parameters even with a limited number of RF chains. In (Ma et al., 2020), Li et al. proposed a beamspace ESPRIT scheme which is applicable for any

²The results can also be extended to channels that exhibit frequency selectivity by considering orthogonal frequency-division multiplexing (OFDM), where the channel vector at each subcarrier can be described by Eq. 1.

³For multi-user system, we can reduce the problem to single-user using orthogonal pilots.

choice of beamformer, that satisfies some mild rank constraints. This approach requires only one antenna to be turned off at a time, and has several advantages such as a negligible drop in transmitted power and spatial coverage.

In this paper, we will analyze this variant of the beamspace ESPRIT algorithm. For ease of exposition, we will consider a flat-fading single carrier system, although extensions are possible. Our analysis will not require any specific assumptions on the distribution of noise, other than assuming it to be bounded. Unlike the prior asymptotic analysis in (Guanghan Xu et al., 1994; Mathews et al., 1996), we will not assume a large number of snapshots, or ignore higher-order perturbation terms (Roemer et al., 2014; Steinwandt et al., 2017). We first review the algorithm from (Ma et al., 2020) in the noiseless setting adapted to the flat-fading scenario.

The key idea behind the ESPRIT algorithm is exploiting the so-called shift invariance property, which refers to arrays with two identical subarrays that are separated by a common displacement. Let $\mathbf{A}_1(\mathcal{F})$, and $\mathbf{A}_2(\mathcal{F})$ denote two subarrays of $\mathbf{A}(\mathcal{F})$ comprising of the first and last $M - 1$ antenna elements. The array $\mathbf{A}(\mathcal{F})$ exhibits shift invariance property since

$$\mathbf{A}_2 = \mathbf{A}_1 \Phi(\mathcal{F}) \quad (5)$$

where $\Phi(\mathcal{F}) = \text{diag}(e^{-j\pi f_1}, \dots, e^{-j\pi f_S})$. One way to realize such subarrays and the corresponding shift invariance is by successively turning off the first and the last antennas. In (Ma et al., 2020), a two-stage approach was utilized to obtain this invariance structure. In the first stage, the M th antenna is turned off, which corresponds to a beamforming matrix $\tilde{\mathbf{W}}_1 := [\mathbf{W}^H, \mathbf{0}_N]^H \in \mathbb{C}^{M \times N}$. In the second stage, the first antenna is turned off, yielding a beamforming matrix $\tilde{\mathbf{W}}_2 := [\mathbf{0}_N, \mathbf{W}^H]^H \in \mathbb{C}^{M \times N}$. Here $\mathbf{W} \in \mathbb{C}^{(M-1) \times N}$ is an analog beamforming matrix which satisfies the following rank condition:

$$\text{rank}(\tilde{\mathbf{W}}_1^H \mathbf{A}(\mathcal{F})) = \text{rank}(\tilde{\mathbf{W}}_2^H \mathbf{A}(\mathcal{F})) = S \text{ for all } \mathcal{F} \quad (6)$$

A necessary condition for Eq. 6 is $N \geq S$. Let $\tilde{\mathbf{Y}}_1 = \tilde{\mathbf{W}}_1^H \mathbf{A}_M(\mathcal{F})\mathbf{X}$, and $\tilde{\mathbf{Y}}_2 = \tilde{\mathbf{W}}_2^H \mathbf{A}_M(\mathcal{F})\mathbf{X}$ be the beamspace measurements acquired using this scheme. We define an augmented observation $\tilde{\mathbf{Y}}$ as

$$\tilde{\mathbf{Y}} := \begin{bmatrix} \tilde{\mathbf{Y}}_1 \\ \tilde{\mathbf{Y}}_2 \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{W}}_1^H \\ \tilde{\mathbf{W}}_2^H \end{bmatrix} \mathbf{A}_M(\mathcal{F})\mathbf{X} \quad (7)$$

Define

$$\mathbf{B} := \begin{bmatrix} \mathbf{W}^H \mathbf{A}_1 \\ \mathbf{W}^H \mathbf{A}_1 \Phi \end{bmatrix} \quad (8)$$

where $\Phi(\mathcal{F}) = \text{diag}(e^{-j\pi f_1}, \dots, e^{-j\pi f_S})$, and $\mathbf{A}_1(\mathcal{F}) \in \mathbb{C}^{(M-1) \times S}$, $\mathbf{A}_2(\mathcal{F}) \in \mathbb{C}^{(M-1) \times S}$ comprise of the first and last $M - 1$ rows of $\mathbf{A}_M(\mathcal{F})$, respectively. For the rest of paper, we suppress the dependence on \mathcal{F} and simply use \mathbf{A}_1 , \mathbf{A}_2 , Φ . Thus, Eq. 7 can be represented as

$$\tilde{\mathbf{Y}} = \mathbf{B}\mathbf{X} = \begin{bmatrix} \mathbf{W}^H \mathbf{A}_1 \\ \mathbf{W}^H \mathbf{A}_1 \Phi \end{bmatrix} \mathbf{X} \quad (9)$$

Note that under the assumption Eq. 6, we have $\text{rank}(\mathbf{B}) = S$. We further assume that \mathbf{X} has full row rank which together with $\text{rank}(\mathbf{B}) = S$ implies that $\text{rank}(\tilde{\mathbf{Y}}) = S$. Let $\mathbf{U}_y \Sigma_y \mathbf{V}_y^H = \tilde{\mathbf{Y}}$ be a reduced singular value decomposition (SVD) of $\tilde{\mathbf{Y}} \in \mathbb{C}^{2N \times S}$, where $\mathbf{U}_y \in \mathbb{C}^{2N \times S}$, $\Sigma_y \in \mathbb{C}^{S \times S}$, $\mathbf{V}_y \in \mathbb{C}^{T \times S}$. Since $\text{rank}(\mathbf{B}) = S$, its columns form a basis for $\mathcal{R}(\tilde{\mathbf{Y}})^4$ (which coincides with $\mathcal{R}(\mathbf{U}_y)$). Thus, there exists an invertible matrix $\mathbf{P} \in \mathbb{C}^{S \times S}$ which provides a mapping between these two bases for $\mathcal{R}(\tilde{\mathbf{Y}})$,

$$\mathbf{U}_y = \mathbf{B}\mathbf{P} \quad (10)$$

Let \mathbf{U}_1 , and \mathbf{U}_2 be two submatrices of \mathbf{U}_y , comprising of its first and last N rows, respectively. Then,

$$\mathbf{U}_y = \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{W}^H \mathbf{A}_1 \mathbf{P} \\ \mathbf{W}^H \mathbf{A}_1 \Phi \mathbf{P} \end{bmatrix} \quad (11)$$

$$\mathbf{U}_2 = \mathbf{U}_1 \mathbf{P}^{-1} \Phi \mathbf{P} \quad (12)$$

Notice that $\Psi := \mathbf{U}_1^\dagger \mathbf{U}_2 = \mathbf{P}^{-1} \Phi \mathbf{P}$.⁵ Since Ψ is diagonalized by \mathbf{P} , we can determine the AoAs (contained in Φ) from the S eigenvalues $\{\lambda_i\}_{i=1}^S$ of Ψ as follows:

$$f_i = -\frac{\arg(\lambda_i)}{\pi}, \quad i = 1, \dots, S \quad (13)$$

where $\arg(\lambda) \in [-\pi, \pi)$ denotes the phase of the complex number λ .

The noiseless beamspace ESPRIT described above can be directly extended to a noisy setting. Consider the following noisy version of the measurement model introduced in Eq. 7:

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{\mathbf{Y}}_1 \\ \hat{\mathbf{Y}}_2 \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{W}}_1 \\ \tilde{\mathbf{W}}_2 \end{bmatrix} (\mathbf{A}_M(\mathcal{F})\mathbf{X} + \mathbf{N}) \quad (14)$$

where $\mathbf{N} \in \mathbb{C}^{M \times T}$ denotes the additive noise. The noisy version of the beamspace ESPRIT algorithm follows on similar lines as its noiseless version, and is summarized in **Algorithm 1**.

Algorithm 1: Channel estimation using beamspace ESPRIT

- 1 Input: $\hat{\mathbf{Y}}$, S
 - 2 Compute an SVD of $\hat{\mathbf{Y}}$: $\hat{\mathbf{Y}} = \hat{\mathbf{U}} \hat{\Sigma} \hat{\mathbf{V}}^H$, $\hat{\mathbf{U}} \in \mathbb{C}^{2N \times 2N}$
 - 3 Obtain $\hat{\mathbf{U}}_1 = \hat{\mathbf{U}}_{1:N, 1:S}$, $\hat{\mathbf{U}}_2 = \hat{\mathbf{U}}_{N+1:2N, 1:S}$
 - 4 Compute $\hat{\Psi} := \hat{\mathbf{U}}_1^\dagger \hat{\mathbf{U}}_2$
 - 5 Compute S eigenvalues of $\hat{\Psi}$: $\{\hat{\lambda}_i\}_{i=1}^S$
 - 6 Estimate $\hat{f}_i = -\frac{\arg(\hat{\lambda}_i)}{\pi}$, $i = 1, \dots, S$
 - 7 **return** $\hat{\mathcal{F}} = \{\hat{f}_i\}_{i=1}^S$
-

⁴ $\mathcal{R}(\tilde{\mathbf{Y}})$ denotes the range space of $\tilde{\mathbf{Y}}$.

⁵ \mathbf{U}_1^\dagger denotes the Moore-Penrose Pseudo-inverse of the rectangular matrix \mathbf{U}_1 .

5 PERFORMANCE ANALYSIS OF BEAMSPACE ESPRIT

We will analyze the performance of the noisy beamspace ESPRIT algorithm. Our analysis extends the recent results from (Li et al., 2020) (for single-snapshot ESPRIT), to beamspace and multi-snapshot scenarios. Our error bounds will explicitly capture the role of the beamforming matrix \mathbf{W} , and provide insights into how the error is shaped by the interaction between the AOA (\mathcal{F}) and beamforming matrix (\mathbf{W}). We will use this characterization to develop new criteria for robust beamformer design in **Section 2**.

We first define some key quantities and metrics. The wrap around distance between two spatial frequencies $f_i, f_j \in [0, 1]$ over the unit interval is defined as:

$$|f_i - f_j|_{\mathbb{T}_u} := \min_{n \in \{0,1\}} |f_i - f_j - n|$$

Our error metric will be the “matching distance” between the estimated $\hat{\mathcal{F}}$ and ground truth \mathcal{F} , defined as:

$$md(\mathcal{F}, \hat{\mathcal{F}}) := \min_{\psi} \max_i |\hat{f}_{\psi(i)} - f_i|_{\mathbb{T}_u} \quad (15)$$

where ψ is taken over all permutations of $\{1, 2, \dots, S\}$. Matching distance between the eigenvalues of $\hat{\Psi}$ and Ψ is similarly defined as

$$md(\Psi, \hat{\Psi}) := \min_{\psi} \max_i |\hat{\lambda}_{\psi(i)} - e^{-j\pi f_i}|_{\mathbb{T}_u} \quad (16)$$

We will use the notation $\sigma_k(\mathbf{Q})$ to denote the k th largest singular value of a matrix \mathbf{Q} .

The following theorem provides an upper bound on the matching distance error between the true AoAs \mathcal{F} and its estimate $\hat{\mathcal{F}}$ obtained from beamspace ESPRIT:

Theorem 1. Consider the noisy measurement model **Eq. 14**. Let $\hat{\mathcal{F}}$ be the estimated frequencies obtained from the beamspace ESPRIT algorithm (**Algorithm 1**). Assume that $\text{rank}(\mathbf{B}\mathbf{X}) = S$. If the noise level is moderately small such that

$$\|\mathbf{N}\|_2 \leq \frac{\sigma_S(\mathbf{B})\sigma_S(\mathbf{X})\sigma_S(\mathbf{U}_1)}{16\sqrt{S}\|\mathbf{W}\|_2} \quad (17)$$

then the matching distance error between $\hat{\mathcal{F}}$ and \mathcal{F} is bounded as

$$md(\mathcal{F}, \hat{\mathcal{F}}) \leq \frac{CS^{1.5}\|\mathbf{B}\|_2\|\mathbf{W}\|_2\|\mathbf{N}\|_2}{\sigma_S(\mathbf{B})^2\sigma_S(\mathbf{X})\sigma_S(\mathbf{U}_1)^2} \quad (18)$$

Here C is a universal constant, and $\mathbf{U}_1 \in \mathbb{C}^{N \times S}$ is defined in **Eq. 11**.

Proof. The proof follows by combining Lemma 6 and 8 of Appendix. See **Supplementary Appendix** for the details.

Remark 1. When $S = 1$, the bound **Eq. 18** can be simplified to

$$md(\mathcal{F}, \hat{\mathcal{F}}) \leq \frac{C'\|\mathbf{W}\|_2\|\mathbf{N}\|_2}{\|\mathbf{W}^H \mathbf{a}(f_1)\|_2\|\mathbf{X}\|_2} \quad (19)$$

where C' is a constant. The quantity $\|\mathbf{W}^H \mathbf{a}(f_1)\|_2$ controls the error bound and represents the beamformer response to the

spatial frequency f_1 (direction θ_1). Note that a simple scaling of the beamforming matrix \mathbf{W} cannot improve performance since it boosts both the noise and signal components. For $N = 1$, $\|\mathbf{W}^H \mathbf{a}(f_1)\|_2 = |\mathbf{w}^H \mathbf{a}(f_1)|$ represents the beamforming gain in direction θ_1 . Of course, if we knew f_1 , we would choose $\mathbf{w} = \mathbf{a}(f_1)$ to maximize $|\mathbf{w}^H \mathbf{a}(f_1)|$. In that case, $\|\mathbf{W}^H \mathbf{a}(f_1)\|_2/\|\mathbf{W}\|_2 = \sqrt{M}$, and the error will scale as $1/\sqrt{M}$. However, during channel sensing, f_1 is unknown and \mathbf{w} needs to be designed to ensure that a certain beamforming gain is achieved over a target sector of interest. Such a design will also decrease the error bound uniformly over that region. In the next section, this will be the basis for beamformer design.

6 ANALOG BEAMFORMER DESIGN FOR MAXIMIZING THE MINIMUM GAIN

6.1 Review of Existing Beamformer Design Approaches

As explained earlier, the choice of the beamformer is implicitly tied to \mathcal{F} . However, prior to channel estimation, the AoAs (\mathcal{F}) of the multipaths are unknown, and it becomes impossible to beamform along these directions. As an alternative, in order to ensure beamforming gain over all possible multipath angles, it is common to assume that the AoAs belong to a sector of interest (Alkhateeb et al., 2014b; Ma et al., 2020). Let $\mathbb{T}: [f_{\min}, f_{\max}]$ be a spatial sector of interest, and suppose we have prior knowledge that the AoAs $f_i \in \mathbb{T}$. We will now review beamformer designs that utilize this prior information to enhance the performance of channel estimation algorithms when AoAs are within this sector of interest (Chen et al., 2019; Ma et al., 2020). The most widely used criterion for designing a hybrid beamformer for mmWave channel sensing is to ensure (i) a constant beamforming gain over the sector of interest, and (ii) zero gain outside the region \mathbb{T} , i.e.,

$$|\mathbf{w}^H \mathbf{a}(f)| = \begin{cases} g, & f \in \mathbb{T} \\ 0, & f \notin \mathbb{T} \end{cases} \quad (20)$$

where g is the desired gain. The criterion **Eq. 20** represents an *ideal brick-wall filter*, and it cannot be realized in practice. A common approach is to ensure the desired gain g only on a finite grid of discretized frequencies (Chen et al., 2019; Alkhateeb et al., 2014b; Ma et al., 2020). Specifically, Let $\mathbf{A}_g = [\mathbf{a}_{\tilde{f}_1}, \mathbf{a}_{\tilde{f}_2}, \dots, \mathbf{a}_{\tilde{f}_{N_g}}] \in \mathbb{C}^{M \times N_g}$ be a dictionary of steering vectors $\mathbf{a}(\tilde{f}_k)$ corresponding to N_g grid points with

$$\tilde{f}_k = \frac{(k-1)}{N_g}, \quad 1 \leq k \leq N_g.$$

Suppose k_1 and k_2 are respectively the smallest and largest integers such that $\tilde{f}_{k_1}, \tilde{f}_{k_2} \in \mathbb{T}$. We introduce a vector $\mathbf{g} \in \mathbb{C}^{N_g}$:

$$[\mathbf{g}]_k = \begin{cases} g e^{j\phi_i}, & k_1 \leq k \leq k_2 \\ 0, & \text{otherwise} \end{cases},$$

which enforces the gain constraints **Eq. 20** at the discretized directions \tilde{f}_k . Note that a phase term ϕ_i is introduced to the response to provide additional flexibility of design. The first step

towards designing the desired beamformer \mathbf{w} involves estimating an *ideal beamformer* (\mathbf{v}^*) design by solving the following least square problem (Alkhateeb et al., 2014b; Chen and Qi, 2018; Chen et al., 2019; Ma et al., 2020):

$$\mathbf{v}^* = \arg \min_{\mathbf{v} \in \mathbb{C}^M} \|\mathbf{A}_g^H \mathbf{v} - \mathbf{g}\|_2 \quad (21)$$

Typically, the grid is chosen to satisfy $N_g > M$, which implies that $\text{rank}(\mathbf{A}_g) = M$, and a closed form solution for the design is given by $\mathbf{v}^* = (\mathbf{A}_g \mathbf{A}_g^H)^{-1} \mathbf{A}_g \mathbf{g} = 1/M \mathbf{A}_g \mathbf{g}$. The beamformers are normalized to obtain $\mathbf{v}_0 = \mathbf{v}^*/\|\mathbf{v}^*\|_2$. In a typical hybrid mmWave system, the beamformer \mathbf{v}_0 is realized by a hybrid structure where $\mathbf{W}_{\text{RF}} \in \mathbb{C}^{M \times N}$ is an RF (analog) beamformer implemented using phase shifters, i.e.

$$[\mathbf{W}_{\text{RF}}]_{m,n} = e^{j\phi_{m,n}}, \quad 1 \leq m \leq M, 1 \leq n \leq N,$$

and a digital beamformer $\mathbf{w}_{\text{BB}} \in \mathbb{C}^{N \times 1}$. Therefore, the second stage of the beamformer design involves approximating the ideal design \mathbf{v}_0 under these additional constraints imposed by the hardware, resulting in the following optimization problem (Alkhateeb et al., 2014b; Chen et al., 2019)

$$\begin{aligned} \min_{\mathbf{W}_{\text{RF}} \in \mathbb{C}^{M \times N}, \mathbf{w}_{\text{BB}} \in \mathbb{C}^N} \quad & \|\mathbf{v}_0 - \mathbf{W}_{\text{RF}} \mathbf{w}_{\text{BB}}\|_2 \\ \text{s.t.} \quad & [\mathbf{W}_{\text{RF}}]_{m,n} = e^{j\phi_{m,n}}, \|\mathbf{W}_{\text{RF}} \mathbf{w}_{\text{BB}}\|_2 = 1 \end{aligned} \quad (22)$$

Several algorithms have been proposed to approximately solve Eq. 22 based on Orthogonal Matching Pursuit (Ayach et al., 2014; Alkhateeb et al., 2014b), alternating minimization (Yu et al., 2016; Chen et al., 2019), and fast search based techniques (Chen and Qi, 2018; Chen et al., 2019) using suitable initialization schemes.

Recently, instead of enforcing the constraint Eq. 20, the authors in (Ma et al., 2020) consider parametric beamformers of the following form, parameterized by $a \in \mathbb{R}$

$$[\mathbf{w}(a)]_{m,n} = \frac{g\sqrt{M}}{2} \left(\frac{e^{j((m-1)\pi-a)f(n-1)} - e^{j((m-1)\pi-a)f(n)}}{j(M-1)\pi-a} \right) \quad (23)$$

where each beamformer is responsible for a partition of \mathbb{T} determined by $\{f(n)\}_{n=0}^N \in \mathbb{T}$. They propose to maximize the following ratio as a function of the parameter a

$$S(a) := \frac{\int_{\mathbb{T}} |\mathbf{w}(a)^H \mathbf{a}(f)|^2 df}{\int_0^1 |\mathbf{w}(a)^H \mathbf{a}(f)|^2 df} \quad (24)$$

The numerator of $S(a)$ represents the power concentrated in the sector of interest \mathbb{T} , and the denominator represents the total power. This criterion is maximized by performing a grid-based search over a after simplifying the ratio Eq. 24. One drawback of both of the aforementioned beamforming strategies is that the design is not guaranteed to ensure a constant gain of g even on the grid points. More importantly, the beamforming gain can drop below the desired level (g) at several regions in \mathbb{T} . There is no analytical characterization of how small the gain can become in these regions. This can lead to significant performance degradation of beamspace channel sensing techniques, especially if the multipath directions are aligned with the above regions where gain is small. In order to overcome these

drawbacks, in the next section, we will propose a new “max-min” criterion for beamformer design to boost the minimum beamforming gain over \mathbb{T} . Such a criterion will allow more robust channel estimation uniformly over \mathbb{T} .

6.2 Beamformer Design Strategy

We motivate our approach for beamformer design by focusing on the quantity $\sigma_S(\mathbf{W}^H \mathbf{A})$, and relate it to the beamforming gain. It can be seen from Eq. 18 that larger the value of $\sigma_S(\mathbf{W}^H \mathbf{A})$, smaller the error of beamspace ESPRIT. Hence, one can aim to design \mathbf{W} that maximizes $\sigma_S(\mathbf{W}^H \mathbf{A})$. But such a \mathbf{W} will depend on $\mathbf{A}(\mathcal{F})$, and we do not know the AoA's \mathcal{F} to begin with. In many cases however, we can assume that the AoA belong to a region/sector of interest given by $\mathbb{T} := [f_{\min}, f_{\max}]$. In other words

$$f_i \in \mathbb{T}, \quad i = 1, 2, \dots, S$$

In this case, we wish to ensure that $\sigma_S(\mathbf{W}^H \mathbf{A}(\mathcal{F}))$ stays uniformly large over the entire set \mathbb{T} . Let α_W be the smallest value that $\sigma_S(\mathbf{W}^H \mathbf{A}(\mathcal{F}))$ can assume over \mathbb{T} , i.e.,

$$\alpha_W := \min_{\mathcal{F} \in \mathbb{T}^S} \sigma_S(\mathbf{W}^H \mathbf{A}(\mathcal{F}))$$

We wish to design \mathbf{W} in order to maximize α_W (under constant modulus constraints on \mathbf{W}), which leads to the following problem:

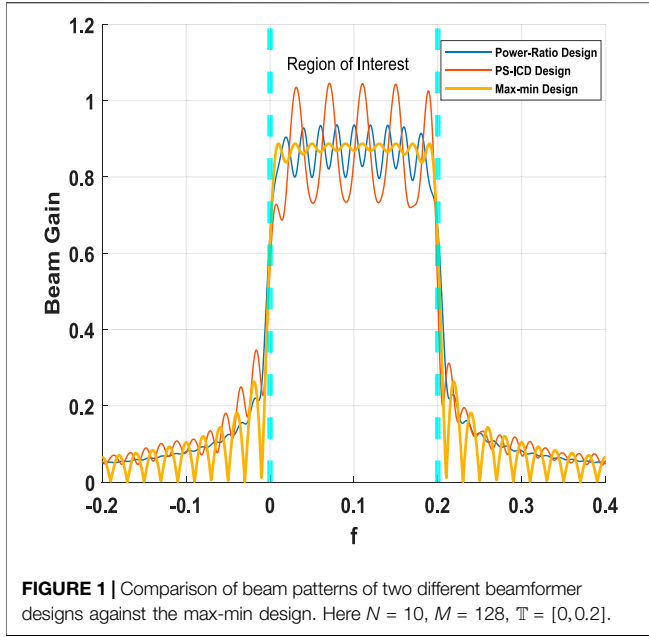
$$\alpha^* := \max_{\mathbf{W} \in \mathbb{C}^{M \times N}} \alpha_W, \quad \text{s.t. } |\mathbf{W}_{m,n}| \in \{0, 1\} \quad (25)$$

This problem belongs to the family of non-convex max-min optimization problems, and it is challenging to solve it for the most general setting. In the next section, we focus on providing the solution of such an optimization problem for the scenario when there is a single source $S = 1$, single RF chain $N = 1$, and contrast the distinctions between the proposed criteria to the existing beamformer designs reviewed in Section 1.

6.3 Optimal Solution for Single Source and Single RF Chain

We consider the single path scenario ($S = 1$) where the channel is given by $\mathbf{h}_t = \alpha_t \mathbf{a}(f)$. This model has been widely used in the mmWave communication literature where the path losses are high and the channel is assumed to have only a single Line of Sight (LOS) path that is dominant (Alkhateeb et al., 2014b; Chiu et al., 2019). Our goal will be to optimize the design of a single RF chain ($N = 1$), which is again motivated by typical hybrid mmWave hardware systems that are equipped with large antenna arrays but often just 1 RF chain (Roh et al., 2014). For $S = 1$, $N = 1$, it can be verified that $\sigma_S(\mathbf{W}^H \mathbf{A}(\mathcal{F})) = \sigma_1(\mathbf{w}^H \mathbf{a}(f)) = |\mathbf{w}^H \mathbf{a}(f)|$. We first develop a framework for designing \mathbf{w} that is optimized to *maximize the minimum gain over the entire sector of interest*. Specifically, this yields the following max-min problem:

$$\begin{aligned} \eta_{\mathbb{T}}^* := \max_{\mathbf{w} \in \mathbb{C}^M} \min_{f \in \mathbb{T}} & |\mathbf{w}^H \mathbf{a}(f)|, \\ \text{s.t. } & |w_m| \in \{0, 1\}, m = 1, 2, \dots, M \end{aligned} \quad (26)$$

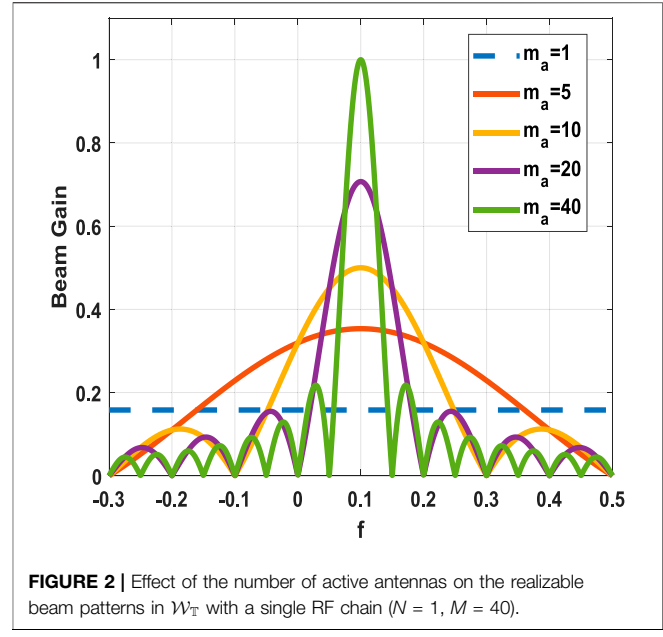


Notice that **Eq. 26** aims to maximize the minimum (or worst case) gain of the beamformer over the sector \mathbb{T} . At this point, we will like to distinguish the criterion **Eq. 26** from those discussed in **Section 1**. Although the quantity of interest is $\mathbf{w}^H \mathbf{a}(f)$ in both cases, the design criteria reviewed in **Section 1** are fundamentally different from **Eq. 26**. Firstly, the approaches in (Alkhateeb et al., 2014b; Chen and Qi, 2018; Chen et al., 2019) solve a grid-based least square loss **Eq. 21**, and therefore the obtained design is not guaranteed to ensure constant gain g even on the grid points. Indeed, there could be an adversarial multipath angle $f_0 \in \mathbb{T}$ where the observed gain is lower:

$$\mathbf{w}_0^H \mathbf{a}(f_0) < g$$

In contrast, the criterion **Eq. 26** can uniformly guarantee beamforming gain of at least $\eta_{\mathbb{T}}^*$ in the entire sector \mathbb{T} . This is illustrated in **Figure 1** where we plot the gain of two different beamformers against the max-min design over the sector of interest $\mathbb{T} = [0, 0.2]$. As can be seen, the gain for “power-ratio” (Ma et al., 2020) and “constant gain” designs (Chen et al., 2019) both significantly fall below the desired level along several directions in the sector \mathbb{T} , whereas for the max-min design the smallest (worst case) gain is larger than the other designs.

Solving **Eq. 26** over the set of all unimodular \mathbf{w} can be a challenging problem, and it can be difficult to quantify and analyze the optimal solution. To make **Eq. 26** tractable so that we can obtain a closed form solution with “quantifiable” minimum beamforming gain over the entire sector \mathbb{T} , we propose to choose \mathbf{w} from a parametric class $\mathcal{W}_{\mathbb{T}}$ of beamformers, which already obey the unimodular constraints. We define $\mathcal{W}_{\mathbb{T}}$ as follows. Given a “center frequency” $f_c \in \mathbb{T}$, and an integer m_a , $1 \leq m_a \leq M$, define $\mathbf{b}(f_c, m_a) \in \mathbb{C}^M$ as



$$[\mathbf{b}(f_c, m_a)]_m = \begin{cases} \exp(-j f_c \pi (m-1)), & m \leq m_a \\ 0, & m > m_a \end{cases}$$

Hence, $\mathbf{b}(f_c, m_a)$ represents a DFT beamformer with m_a active antennas and whose pass band is centered at f_c . Let $m_r := \min\{\frac{4}{f_{\max} - f_{\min}}, M\}$. Given an integer m_a satisfying $1 \leq m_a \leq m_r$, we define $\mathcal{C}_{m_a, \mathbb{T}} \subset \mathbb{C}^M$ as the set of all DFT beamformers with m_a active antennas, generated by varying the center frequency f_c over the interval $[f_{\max} - \frac{2}{m_a}, f_{\min} + \frac{2}{m_a}]$:

$$\mathcal{C}_{m_a, \mathbb{T}} = \left\{ \mathbf{b}(f_c, m_a) \in \mathbb{C}^M, f_c \in \left[f_{\max} - \frac{2}{m_a}, f_{\min} + \frac{2}{m_a} \right] \right\}$$

All beamformers in $\mathcal{C}_{m_a, \mathbb{T}}$ therefore have the same beamwidth determined by m_a with the flexibility of shifting the beam centers to any location f_c such that the desired coverage region \mathbb{T} remains in the main lobe of the beam, i.e., $\mathbb{T} \subset [f_c - \frac{2}{m_a}, f_c + \frac{2}{m_a}]$. Finally, we define the set $\mathcal{W}_{\mathbb{T}}$ that comprises of DFT beamformers of all possible mainlobe widths, i.e.,

$$\mathcal{W}_{\mathbb{T}} := \bigcup_{m_a=1}^{m_r} \mathcal{C}_{m_a, \mathbb{T}} \quad (27)$$

Using this class $\mathcal{W}_{\mathbb{T}}$, we propose to solve

$$\bar{\eta}_{\mathbb{T}}^* := \max_{\mathbf{w} \in \mathcal{W}_{\mathbb{T}}} \min_{f \in \mathbb{T}} |\mathbf{w}^H \mathbf{a}(f)| \quad (28)$$

Notice that the class of beamformers $\mathcal{W}_{\mathbb{T}}$ is quite broad, consisting of multi-resolution beams of varying beam widths (determined by m_a), and for each resolution/beamwidth the permissible beams are shifted copies of each other. Such beamformers have two fold-advantages (i) they inherently satisfy the desired constant modulus constraint for hardware implementation using phase shifters and switches, and (ii) they are amenable to theoretical analysis due to the parametric

structure. A key distinction compared to the designs described in **Section 1** is that we are only considering purely RF or analog beamformers, without any baseband processing. As will be shown in the simulations, with the same budget of RF chains, this max-min design strategy yields superior performance compared to the hybrid designs, especially for adversarial multipath configurations.

The max-min design criterion involves a natural trade-off between gain and coverage. As shown in **Figure 2**, when m_a is large the resulting beams are sharper and can offer higher gains. Despite their higher gain, their coverage is limited owing to the narrow main lobes. Since our goal is to ensure a certain minimum gain uniformly across the entire sector \mathbb{T} , we must design the beam centers and select the widths appropriately to satisfy this objective. The parametric structure of $\mathcal{W}_{\mathbb{T}}$ allows us to obtain the following expression for the beamforming gain for $\mathbf{w} = \mathbf{b}(f_c, m_a)$:

$$\|\mathbf{w}^H \mathbf{a}(f)\|_2^2 = \left(\frac{\sin\left(\frac{(f-f_c)\pi m_a}{2}\right)}{\sin\left(\frac{(f-f_c)\pi}{2}\right)} \right)^2$$

Owing to the structure of $\mathcal{W}_{\mathbb{T}}$, searching for the optimum \mathbf{w} reduces to finding the optimum center f_c and active antennas m_a that jointly maximize the minimum gain over the entire sector of interest \mathbb{T} . Theorem 2 provides the optimal choice of this design.

Theorem 2. Let $\mathbb{T} = [f_{\min}, f_{\max}] \in [0, 1]$, $\Delta_f = f_{\max} - f_{\min}$. Given the class of beamformers $\mathcal{W}_{\mathbb{T}}$ defined in **Eq. 27**, the optimal value of **Eq. 28** is given by

$$\bar{\eta}_{\mathbb{T}}^* = \begin{cases} \frac{\sin\left(\left\lceil \frac{2}{\Delta_f} \right\rceil \Delta_f \pi / 4\right)}{\sin(\Delta_f \pi / 4)} & \text{if } \Delta_f \geq \frac{2}{M} \\ \left| \frac{\sin(M \Delta_f \pi / 4)}{\sin(\Delta_f \pi / 4)} \right| & \text{if } \Delta_f \leq \frac{2}{M} \end{cases} \quad (29)$$

The optimal value is attained by the beamformer $\mathbf{b}(f_c^*, m_a^*) \in \mathcal{W}_{\mathbb{T}}$ where

$$\begin{aligned} f_c^* &= f_{\text{mid}} = \frac{f_{\max} + f_{\min}}{2} \\ m_a^* &= \begin{cases} \left\lceil \frac{2}{\Delta_f} \right\rceil & \text{if } \Delta_f \geq \frac{2}{M} \\ M & \text{if } \Delta_f \leq \frac{2}{M} \end{cases} \end{aligned} \quad (30)$$

where notation $\lceil x \rceil$ refers to the closest integer to x .

Proof. The problem **Eq. 28** is equivalent to the following problem

$$\max_{(f_c, m_a) \in \mathcal{D}_{M, \mathbb{T}}} \min_{f \in \mathbb{T}} |\mathbf{b}(f_c, m_a)^H \mathbf{a}(f)| \quad (31)$$

where $\mathcal{D}_{M, \mathbb{T}}$ denotes the set of (f_c, m_a) pairs such that $\mathbf{b}(f_c, m_a) \in \mathcal{W}_{\mathbb{T}}$:

$$\mathcal{D}_{M, \mathbb{T}} = \left\{ (f_c, m_a) \mid f_c \in \left[f_{\max} - \frac{2}{m_a}, f_{\min} + \frac{2}{m_a} \right], m_a \leq m_r \right\}$$

Fix m_a . Define $g_{(f_c, m_a)}: \mathbb{T} \rightarrow \mathbb{R}_+$ as:

$$\begin{aligned} g_{(f_c, m_a)}(f) &:= |\mathbf{b}(f_c, m_a)^H \mathbf{a}(f)| = \left| \sum_{i=0}^{m_a-1} e^{j\pi(f_c-f)i} \right| \\ &= \left| \frac{\sin\left((f_c-f)\frac{m_a\pi}{2}\right)}{\sin\left(\pi\frac{f_c-f}{2}\right)} \right| \end{aligned} \quad (32)$$

The function $g_{(f_c, m_a)}(f)$ is symmetric around f_c and is monotonically increasing for $f \in [f_c - 2/m_a, f_c]$ and monotonically decreasing for $f \in [f_c, f_c + 2/m_a]$. Since $(f_c, m_a) \in \mathcal{D}_{M, \mathbb{T}}$, we have $\mathbb{T} \subset [f_c - 2/m_a, f_c + 2/m_a]$ for all $(f_c, m_a) \in \mathcal{D}_{M, \mathbb{T}}$. Therefore, it holds that

$$\min_{f \in \mathbb{T}} g_{(f_{\text{mid}}, m_a)}(f) = g_{(f_{\text{mid}}, m_a)}(f_{\min}) = g_{(f_{\text{mid}}, m_a)}(f_{\max}) \quad (33)$$

We first consider

$$f_c > f_{\text{mid}} \quad (34)$$

Based on the definition of $g_{(f_c, m_a)}(f)$, for all $(f_c, m_a) \in \mathcal{D}_{M, \mathbb{T}}$ we have

$$g_{(f_{\text{mid}}, m_a)}(f_{\min}) = g_{(f_c, m_a)}(f_{\min} + (f_c - f_{\text{mid}})), \quad (35)$$

Further the fact $(f_c, m_a) \in \mathcal{D}_{M, \mathbb{T}}$ along with **Eq. 34** implies that

$$f_c - 2/m_a < f_{\min} < f_{\min} + (f_c - f_{\text{mid}}) < f_c \quad (36)$$

Thus

$$g_{(f_c, m_a)}(f_{\min}) \leq g_{(f_c, m_a)}(f_{\min} + (f_c - f_{\text{mid}})) \quad (37)$$

Using **Eqs 33–35, 37** we have

$$\begin{aligned} \min_{f \in \mathbb{T}} g_{(f_{\text{mid}}, m_a)}(f) &= g_{(f_{\text{mid}}, m_a)}(f_{\min}) \\ &= g_{(f_c, m_a)}(f_{\min} + (f_c - f_{\text{mid}})) \\ &\stackrel{(a)}{\geq} g_{(f_c, m_a)}(f_{\min}) \\ &\stackrel{(b)}{\geq} \min_{f \in \mathbb{T}} g_{(f_c, m_a)}(f) \end{aligned}$$

where the inequality (a) follows from the monotonically increasing behavior of g over the interval $[f_c - 2/m_a, f_c]$, and **Eq. 36** which implies $f_{\min} \in [f_c - 2/m_a, f_c]$. The inequality (b) follows from the fact that $f_{\min} \in \mathbb{T}$. Hence, we get

$$\min_{f \in \mathbb{T}} g_{(f_{\text{mid}}, m_a)}(f) \geq \min_{f \in \mathbb{T}} g_{(f_c, m_a)}(f) \quad (38)$$

Using a similar argument we can show that **Eq. 38** also holds when $f_c \leq f_{\text{mid}}$.

Notice **Eq. 32** implies that

$$\operatorname{argmax}_{m_a \leq m_r} |g(f_{\text{mid}}, m_a)(f_{\text{min}})| = \operatorname{argmax}_{m_a \leq m_r} \left| \sin\left(\Delta_f \frac{m_a \pi}{4}\right) \right| \quad (39)$$

Therefore, it can be easily verified that the maximum value of $g(f_{\text{mid}}, m_a)(f_{\text{min}})$ over all $(f_c = f_{\text{mid}}, m_a) \in \mathcal{D}_{M, \mathbb{T}}$ pairs is given as

$$\max_{(f_c = f_{\text{mid}}, m_a) \in \mathcal{D}_{M, \mathbb{T}}} g(f_{\text{mid}}, m_a)(f_{\text{min}}) \quad (40)$$

$$= \max_{(f_c = f_{\text{mid}}, m_a) \in \mathcal{D}_{M, \mathbb{T}}} g(f_{\text{mid}}, m_a)(f_{\text{max}}) \\ = \begin{cases} \frac{\sin\left(\left[\frac{2}{\Delta_f}\right] \Delta_f \pi / 4\right)}{\sin(\Delta_f \pi / 4)} & \text{if } \Delta_f \geq \frac{2}{M} \\ \frac{\sin(M \Delta_f \pi / 4)}{\sin(\Delta_f \pi / 4)} & \text{if } \Delta_f \leq \frac{2}{M} \end{cases} \quad (41)$$

which will be attained at

$$m_a^* = \begin{cases} \left\lceil \frac{2}{\Delta_f} \right\rceil & \text{if } \Delta_f \geq \frac{2}{M} \\ M & \text{if } \Delta_f \leq \frac{2}{M} \end{cases} \quad (42)$$

Optimality of (f_{mid}, m_a^*) implies that for all $(f_c = f_{\text{mid}}, m_a) \in \mathcal{D}_{M, \mathbb{T}}$ we have,

$$\min_{f \in \mathbb{T}} g(f_{\text{mid}}, m_a^*)(f) \geq \min_{f \in \mathbb{T}} g(f_{\text{mid}}, m_a)(f) \quad (43)$$

Hence as the result of Eq. 38, and Eq. 43, for any $(f_c, m_a) \in \mathcal{D}_{M, \mathbb{T}}$ we have

$$\min_{f \in \mathbb{T}} g(f_c, m_a)(f) \leq \min_{f \in \mathbb{T}} g(f_{\text{mid}}, m_a)(f) \leq \min_{f \in \mathbb{T}} g(f_{\text{mid}}, m_a^*)(f)$$

which completes the proof.

6.4 A Sub-band Splitting Design for Multiple RF Chains

We now develop a heuristic for $N > 1$ RF chains that utilizes the optimal design from Theorem 2. When $N > 1$, let $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_N] \in \mathbb{C}^{M \times N}$ be the analog beamforming matrix. In this case, $\sigma_S(\mathbf{W}^H \mathbf{a}(f))$ assumes the form

$$\|\mathbf{W}^H \mathbf{a}(f)\|_2^2 = \sum_{k=1}^N \|\mathbf{w}_k^H \mathbf{a}(f)\|_2^2$$

Similar to our prior objective, we now wish to ensure the minimum value of $\|\mathbf{W}^H \mathbf{a}(f)\|_2^2$ is maximized:

$$\bar{\eta}_I^* := \max_{\mathbf{w}_1, \dots, \mathbf{w}_N \in \mathcal{W}_{\mathbb{T}}} \min_{f \in \mathbb{T}} \sum_{k=1}^N |\mathbf{w}_k^H \mathbf{a}(f)|_2^2 \quad (44)$$

Instead of exactly solving Eq. 44, we will use our design for 1 RF chain to develop a subband splitting approach for designing $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_N$. In particular, given the interval \mathbb{T} and a budget of N RF-chains, we partition \mathbb{T} into N subbands,

$$\mathbb{T} = \bigcup_{k=1}^N \mathbb{T}_k \quad (45)$$

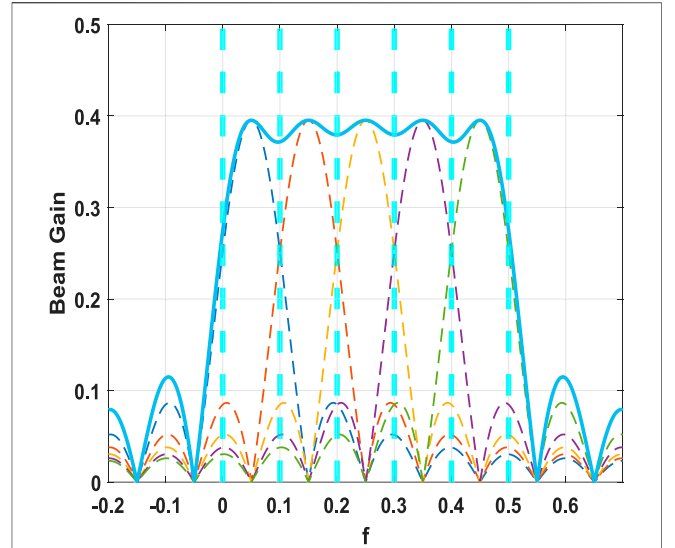


FIGURE 3 | Subband splitting approach for designing $N = 5$ beamformers by partitioning the region of interest into $N = 5$ subbands and choosing the optimal beamformer within each subband. A single RF chain is responsible for maximizing the minimum gain in each subband/partition.

where $\mathbb{T}_k := [f_{\text{min}} + (\Delta_f) \frac{k-1}{N}, f_{\text{min}} + (\Delta_f) \frac{k}{N}]$ represents the k th partition or subband of the sector. Such a subband design was also considered in (Ma et al., 2020), but the criterion was different. In particular, in (Ma et al., 2020), each beamformer is responsible for maximizing the ratio Eq. 24 in the subband. Instead of maximizing the power-ratio, we propose to maximize the worst-case gain within each subband, which will prove to be more robust, especially for adversarial settings. We solve N different optimization problems to find \mathbf{w}_i , $i = 1, 2, \dots, N$ as follows

$$\max_{\mathbf{w}_k \in \mathcal{W}_{\mathbb{T}_k}} \min_{f \in \mathbb{T}_k} |\mathbf{w}_k^H \mathbf{a}(f)|, 1 \leq k \leq N \quad (46)$$

As a result of the partition (Stewart, 1990), we can adopt the optimal design obtained from Theorem 2 for each of the N problems in (Hansen, 1987). This approach greedily designs the columns of the matrix \mathbf{W} such that the k th beamformer \mathbf{w}_k maximizes the minimum gain over \mathbb{T}_k . Let $f_c^{(k)}$ and $m_a^{(k)}$ be the center and number of active antenna for the optimum \mathbf{w}_k that solves Eq. 46. Then, Theorem 2 dictates

$$f_c^{(k)} = f_{\text{min}} + (\Delta_f) \frac{2k-1}{2N}, m_a^{(k)} = \min\left(\left\lceil \frac{2N}{\Delta_f} \right\rceil, M\right) \quad (47)$$

This design follows from Theorem 2, where each of the N intervals $\{\mathbb{T}_k\}_{k=1}^N$ are of length Δ_f/N , hence the optimal value of $m_a^{(k)}$ is given by Eq. 30. In this case, we can obtain a lower bound on $\|\mathbf{W}^H \mathbf{a}(f)\|_2^2$ by using the gain characterization from Theorem 2.

An example of this sub-band max-min design scheme is illustrated in Figure 3, where the sector of interest $\mathbb{T} = [0, 0.5]$ is partitioned into 5 subbands (partitions are shown using the

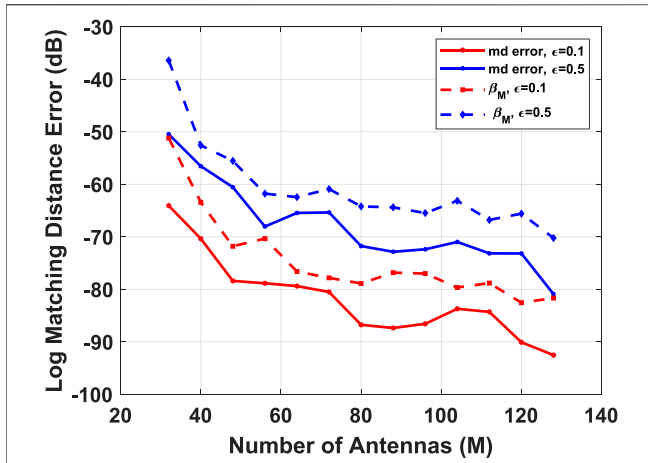


FIGURE 4 | Comparative study of matching distance error as a function of the number of antennas M , using the max-min beamforming scheme for a multipath channel with $S = 4$ paths, $N = 20$ RF chains, $T = 50$ Snapshots. The dotted lines illustrate the trend predicted by the bounds in Theorem 1.

dotted lines). Each of the available 5 RF chains maximize the beamforming gain in each subband. The solid line indicates the overall gain in the sector of interest showing reduced gain drop over \mathbb{T} .

7 NUMERICAL RESULTS

In this section, we experimentally validate our analysis of beamspace ESPRIT (in Theorem 1), and also evaluate the performance of the max-min beamformer proposed in Section 2.

In the first experiment, we study the effect of varying the number (M) of antennas on the matching distance error. We fix the AOAs of the channel paths to be $\mathcal{F} = \{0.21, 0.29, 0.36, 0.38\}$, which belong to the region of interest $\mathbb{T} = [0.2, 0.4]$. For each M , the channel gains \mathbf{X} are normalized to satisfy $\|\mathbf{X}\|_2 = 1$. The received signal in Eq. 4 is corrupted by bounded random noise, normalized to satisfy $\|\mathbf{N}\|_2 = \epsilon$. We keep \mathbf{X} fixed and only the noise is randomly generated during the Monte Carlo experiments. We compute the average matching distance error of beamspace ESPRIT for this channel configuration averaged over $L = 500$ different noise realizations. In Figure 4, we plot the average matching distance error of the beamspace ESPRIT algorithm using the max-min beamformer proposed in Section 2. Although the design was proposed for $S = 1$, it can be deployed for channels with $S > 1$ multipath components as well. We vary the number of antennas from $M = 32$ to $M = 128$ for two different noise levels $\epsilon = 0.1, 0.5$. From Theorem 1, for a fixed channel configuration (fixed S , and \mathcal{F}), the matching distance error bound is proportional to $\beta_M(\mathbf{W}, \mathbf{N})$, given by

$$\beta_M(\mathbf{W}, \mathbf{N}) = \frac{\|\mathbf{B}\|_2 \|\mathbf{W}\|_2 \|\mathbf{N}\|_2}{\sigma_S(\mathbf{B})^2 \sigma_S(\mathbf{X}) \sigma_S(\mathbf{U}_1)^2}$$

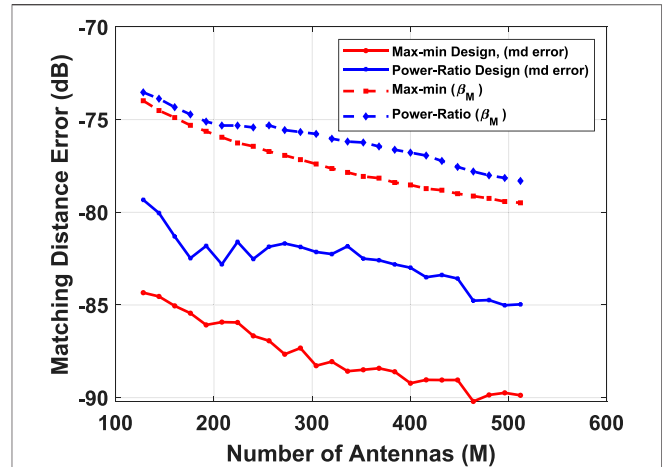


FIGURE 5 | Comparative study of matching distance error as a function of the number of antennas M , using the max-min and power ratio beamforming schemes for a LOS channel with $S = 1$ path along $\mathcal{F} = [0.18]$, $N = 10$ RF chains, $T = 5$ snapshots, and noise level $\epsilon = 0.5$. The dotted lines illustrate the trend predicted by the bounds in Theorem 1.

In order to validate the trend predicted by our theoretical result, we overlay the average $\beta_M(\mathbf{W}, \mathbf{N})$ (averaged over the noise realizations), scaled by a factor of 10^{-3} . As shown in Figure 4, the average matching distance error follows the trend predicted by the bound in Theorem 1. As expected, both the empirical error and the trend based on $\beta_M(\mathbf{W}, \mathbf{N})$ increase with ϵ . The fluctuations in error (which are also consistent with the fluctuations in the bound) can be attributed to the fact that as we vary M , the gain of the beamformers at the (fixed) AoAs also fluctuate.

We now compare the proposed max-min beamformer against three other beamformer design strategies (Li et al., 2020; Chen et al., 2019) which were reviewed in Section 1. In all of the following experiments, we choose $\mathbb{T} = [0, 0.2]$ as our region of interest. In all figures, “Power-Ratio” refers to the beamformer designed using Eq. 23 (Li et al., 2020), “PS-ICD” refers to the design in (Chen et al., 2019) which approximates the solution of Eq. 21 using phase shifters, and “DFT” refers to a sub-selection of the columns of a DFT matrix (according to the region of interest and number of available RF chains). “Max-Min design” denotes the beamformer described in Section 2, where we partition the region of interest \mathbb{T} and construct the optimal beamformers corresponding to each region.

We first generate a single LOS path with $\mathcal{F} = \{0.18\} \in \mathbb{T} = [0, 0.2]$, and a fixed channel gain matrix \mathbf{X} satisfying $\|\mathbf{X}\|_2 = M$. Similar to the previous setting, we consider bounded noise with $\|\mathbf{N}\|_2 = \epsilon M$ to ensure that the ratio $\|\mathbf{X}\|_2 / \|\mathbf{N}\|_2$ is fixed. In Figure 5, we plot the average matching distance error for three different beamformers, and overlay the average trend predicted by $\beta_M(\mathbf{W}, \mathbf{N})$. The empirical average matching distance error exhibits a similar trend as predicted by $\beta_M(\mathbf{W}, \mathbf{N})$. The performance gap observed between the bounds is also reflected in the actual matching

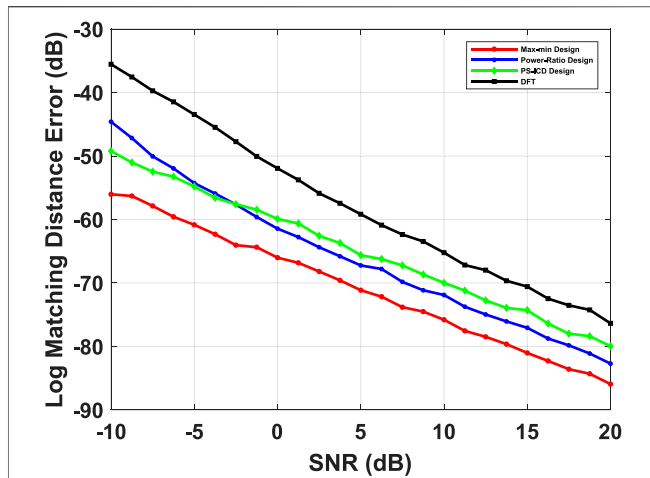


FIGURE 6 | Comparison of performance of max-min beamformer against power-ratio, PS-ICD and DFT beamformers as a function of SNR, with channel path directions given by $\mathcal{F} = [0.18]$. Each beamformer design is realized with $N = 5$ RF chains where $M = 64$.

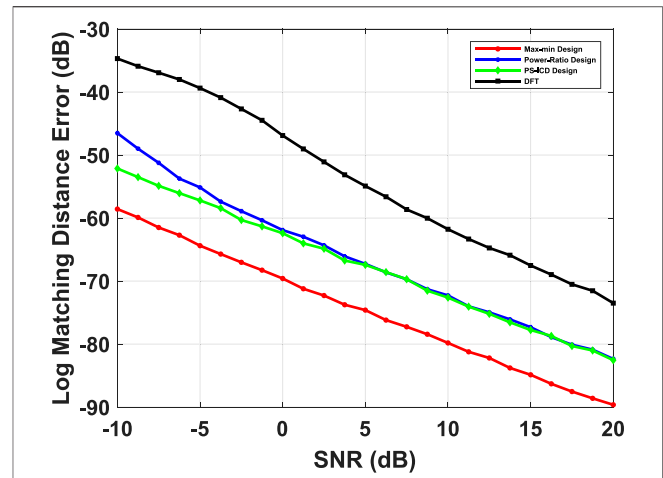


FIGURE 8 | Comparison of performance of max-min beamformer against power-ratio, PS-ICD and DFT beamformers as a function of SNR, with channel path directions given by $\mathcal{F} = [0.02, 0.1, 0.16, 0.18]$. Each beamformer design is realized with $N = 10$ RF chains where $M = 128$.

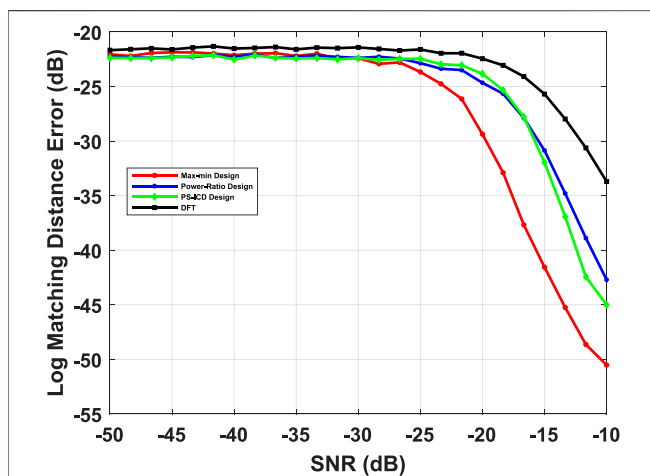


FIGURE 7 | Comparison of channel estimation performance of max-min beamformer against power-ratio, PS-ICD and DFT beamformers in the SNR regime of -50 dB to -10 dB, with channel path directions given by $\mathcal{F} = [0.18]$. Each beamformer design is realized with $N = 5$ RF chains where $M = 64$.

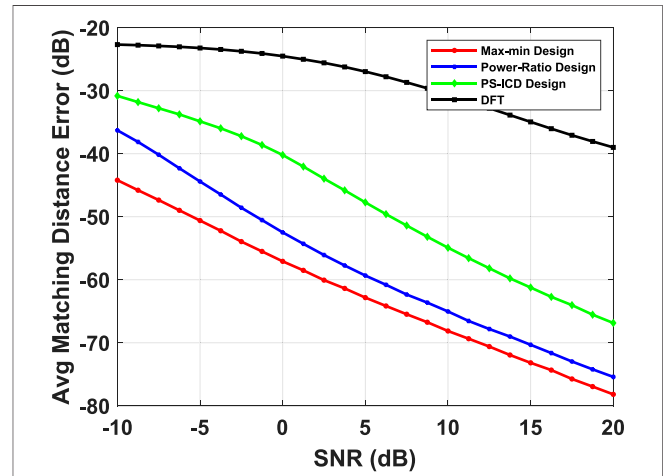


FIGURE 9 | Comparison of beamspace channel estimation performance for a LOS channel with $S = 1$ path, $M = 64$ antennas, $N = 5$ RF chains as a function of SNR. The plot shows the matching distance error averaged over $K = 100$ different channel realizations by varying one of the AOAs uniformly on a grid in the region of interest.

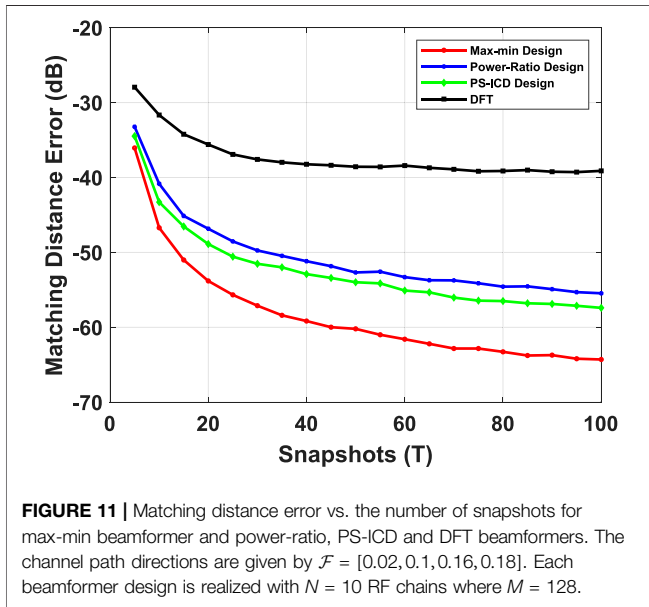
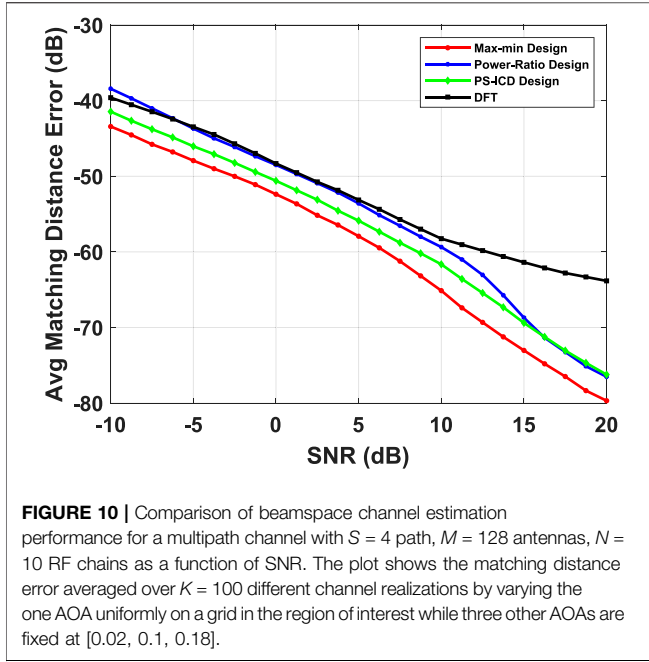
distance error, hence, illustrating the effectiveness of the proposed beamformer design by leveraging the error analysis.

In **Figures 6–12**, we assume that the path gains are i. i.d random variables distributed as $x_{s,t} \sim \mathcal{NC}(0, \sigma_x)$. Furthermore, the noise are assumed to be i. i.d random variables distributed as $n_{m,t} \sim \mathcal{NC}(0, \sigma_n)$ and independent from the channel gains. We define SNR as

$$\text{SNR} := 10 \log \frac{\sigma_x}{\sigma_n}$$

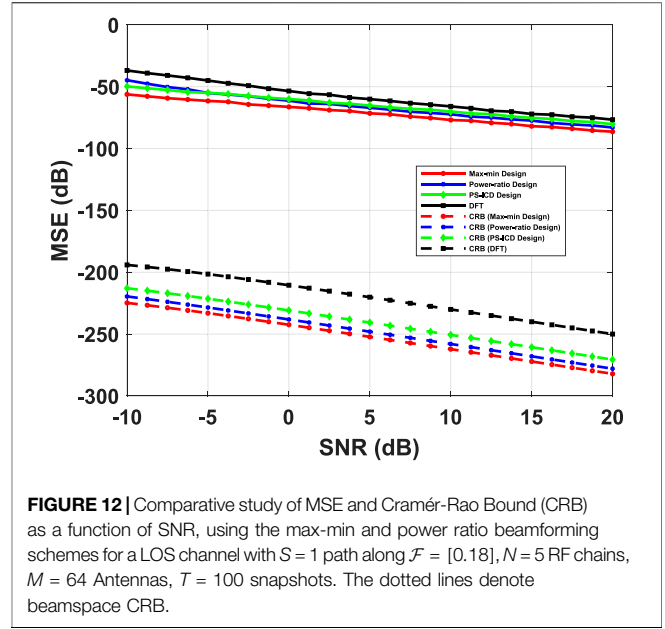
In **Figures 6–8**, we compare the average Matching Distance error of beamspace ESPRIT as a function of SNR, using different

beamformers. In **Figures 6, 7**, the channel is assumed to have a single LOS path, with $M = 64$ antennas, $N = 5$ RF chains, and $T = 50$ snapshots. In **Figure 8**, we consider a multi-path channel with $S = 4$ paths, $M = 128$ antennas, $N = 10$ RF chains, and $T = 50$ temporal snapshots. We considered a specific channel configuration where the AoAs are chosen from the set $\mathcal{F} = \{0.02, 0.1, 0.16, 0.18\}$. As can be observed, the max-min design outperforms other schemes uniformly over the entire range of SNR, maintaining a gap in error of about 10 dB. In **Figure 7**, we further show how the performance of all the beamformers degrade when the SNR becomes very small. In



such a low SNR regime, the matching distance error of one channel path converges to half of the length of the region of interest, i. e. $20 \log_{10} (0.1) = -20$ dB.

In the next experiment, we investigate the performance of max-min beamformer with respect to different path directions over the region of interest. In order to do so, we choose the path angles from a uniform grid of size $K = 100$ over the region of interest \mathbb{T} , and plot the average matching distance error (Avg-md) over all possible channel configurations. For a single path LOS channel, let $\mathcal{F}_k = \{\frac{0.2(k-1)}{K}\}$ be the AoA direction while $\hat{\mathcal{F}}_{k,l}$ refers to the estimate of \mathcal{F}_k for the l th realization. The Avg-md is given as follows



$$\text{Avg-md}(\mathbb{T}, K, L) = 20 \log \frac{1}{KL} \sum_{k=1}^K \sum_{l=1}^L \text{md}(\mathcal{F}_k, \hat{\mathcal{F}}_{l,k})$$

In **Figure 9** we plot the Avg-md as a function of SNR, under a similar setting as **Figure 6**. In **Figure 10**, we repeat this for a multi-path channel, where three path directions are fixed, and the AoA of the fourth path is varied over \mathbb{T} , i. e. $\mathcal{F}_k = \{0.02, 0.1, 0.18, \frac{0.2(k-1)}{K}\}$. The superior performance of the Max-Min beamformer can be attributed to the fact that its minimum gain over \mathbb{T} is always larger than that of the other beamformers.

Owing to a higher “worst-case gain” over the region of interest, max-min beamformers enable a more *robust channel estimation* in face of certain (adversarial) channel configurations/multipath directions, compared to the other beamformers whose pass-band gain can drop significantly below the desired (constant) level for these directions, resulting in an overall degradation of the average error.

In **Figure 11**, we demonstrate the effect of the total number of temporal snapshots on the ESPRIT matching distance error for different beamforming schemes. In this experiment, we fix the SNR at -5 dB, and the other parameters are identical to those used in **Figure 8**. The plot shows that our schemes is effective even in the limited snapshot regime. Additionally, the gap between the performance of Min-Max design and the other beamformers steadily increases with the number of snapshots.

Finally, **Figure 12** compares the MSE of beamspace ESPRIT against the beamspace Cramér-Rao Bound (CRB) for different beamformers (Van Trees, 2004), under a similar setting as **Figure 6**. As we can observe, the trend exhibited by the empirical MSE is consistent with the trend shown by the CRB, and the max-min beamformers exhibit a smaller CRB compared to other beamformers.

8 CONCLUSION

In this work, we have extended the analysis of single-snapshot ESPRIT for beamspace and multi-snapshot scenarios. Our analysis is non-asymptotic in the number of snapshots, and provides an upper bound on the matching distance error without requiring any specific distribution for the noise. The error analysis revealed the role of the beamformer design. Based on our theoretical analysis, we have proposed a novel max-min criterion for designing beamformers which ensures a minimum beamforming gain uniformly over a region of possible path directions. We have considered a family of multi-resolution beamformers which can be implemented with phase shifters, and proposed the optimal beamformers from this family with respect to the new max-min criteria. By conducting several numerical experiments, we have empirically established the superior performance of our designed beamformers compared to other beamformers. In future, an interesting question would be to extend the max-min design over a broader class of beamformers.

REFERENCES

- Alkhateeb, A., El Ayach, O., Leus, G., and Heath, R. W. (2014). Channel Estimation and Hybrid Precoding for Millimeter Wave Cellular Systems. *IEEE J. Sel. Top. Signal. Process.* 8 (5), 831–846. doi:10.1109/jstsp.2014.2334278
- Alkhateeb, A., and Heath, R. W. (2016). Frequency Selective Hybrid Precoding for Limited Feedback Millimeter Wave Systems. *IEEE Trans. Commun.* 64 (5), 1801–1818. doi:10.1109/tcomm.2016.2549517
- Alkhateeb, A., Mo, J., Gonzalez-Prelcic, N., and Heath, R. W. (2014). MIMO Precoding and Combining Solutions for Millimeter-Wave Systems. *IEEE Commun. Mag.* 52 (12), 122–131. doi:10.1109/mcom.2014.6979963
- Ayach, O. E., Rajagopal, S., Abu-Surra, S., Pi, Z., and Heath, R. W. (2014). Spatially Sparse Precoding in Millimeter Wave MIMO Systems. *IEEE Trans. Wireless Commun.* 13 (3), 1499–1513. doi:10.1109/twc.2014.011714.130846
- Bai, T., and Heath, R. W. (2014). Coverage and Rate Analysis for Millimeter-Wave Cellular Networks. *IEEE Trans. Wireless Commun.* 14 (2), 1100–1114.
- Bauer, F. L., and Fike, C. T. (1960). Norms and Exclusion Theorems. *Numer. Math.* 2 (1), 137–141. doi:10.1007/bf01386217
- Bogale, T. E., Le, L. B., and Wang, X. (2015). Hybrid Analog-Digital Channel Estimation and Beamforming: Training-Throughput Tradeoff. *IEEE Trans. Commun.* 63 (12), 5235–5249. doi:10.1109/tcomm.2015.2495191
- Chen, K., and Qi, C. (2018). “Beam Design with Quantized Phase Shifters for Millimeter Wave Massive MIMO,” in 2018 IEEE Global Communications Conference (GLOBECOM) (IEEE), 1–7. doi:10.1109/glocom.2018.8647641
- Chen, K., Qi, C., and Li, G. Y. (2019). Two-step Codeword Design for Millimeter Wave Massive MIMO Systems with Quantized Phase Shifters. *IEEE Trans. Signal Process.* 68, 170–180.
- Chiu, S.-E., Ronquillo, N., and Javidi, T. (2019). Active Learning and CSI Acquisition for Mmwave Initial Alignment. *IEEE J. Select. Areas Commun.* 37 (11), 2474–2489. doi:10.1109/jsac.2019.2933967
- Franklin, J. N. (2012). *Matrix Theory*. Courier Corporation.
- Gao, Z., Hu, C., Dai, L., and Wang, Z. (2016). Channel Estimation for Millimeter-Wave Massive MIMO with Hybrid Precoding over Frequency-Selective Fading Channels. *IEEE Commun. Lett.* 20 (6), 1259–1262. doi:10.1109/lcomm.2016.2555299
- González-Coma, J. P., Rodríguez-Fernández, J., González-Prelcic, N., Castedo, L., and Heath, R. W. (2018). Channel Estimation and Hybrid Precoding for Frequency Selective Multiuser Mmwave MIMO Systems. *IEEE J. Sel. Top. Signal. Process.* 12 (2), 353–367. doi:10.1109/jstsp.2018.2819130
- Guanghan Xu, G., Silverstein, S. D., Roy, R. H., and Kailath, T. (1994). Beamspace Esprit. *IEEE Trans. Signal. Process.* 42 (2), 349–356. doi:10.1109/78.275607

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

SS, PS, and PP developed the main conceptual idea and theoretical results presented in paper. SS, PS, and PP planned the experiments and numerical results presented in the manuscript. All authors contributed in writing the manuscript and approved the content of the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frsip.2021.820617/full#supplementary-material>

- Guo, Z., Wang, X., and Heng, W. (2017). Millimeter-wave Channel Estimation Based on 2-d Beamspace MUSIC Method. *IEEE Trans. Wireless Commun.* 16 (8), 5384–5394. doi:10.1109/twc.2017.2710049
- Haardt, M., and Nossek, J. A. (1995). Unitary Esprit: How to Obtain Increased Estimation Accuracy with a Reduced Computational Burden. *IEEE Trans. Signal. Process.* 43 (5), 1232–1242. doi:10.1109/78.382406
- Haghighatshoar, S., and Caire, G. (2016). Massive MIMO Channel Subspace Estimation from Low-Dimensional Projections. *IEEE Trans. Signal Process.* 65 (2), 303–318.
- Han, S., I, C.-L., Xu, Z., and Rowell, C. (2015). Large-scale Antenna Systems with Hybrid Analog and Digital Beamforming for Millimeter Wave 5g. *IEEE Commun. Mag.* 53 (1), 186–194. doi:10.1109/mcom.2015.7010533
- Hansen, P. C. (1987). The Truncatedsvd as a Method for Regularization. *Bit* 27 (4), 534–553. doi:10.1007/bf01937276
- Hur, S., Kim, T., Love, D. J., Krogmeier, J. V., Thomas, T. A., and Ghosh, A. (2013). Millimeter Wave Beamforming for Wireless Backhaul and Access in Small Cell Networks. *IEEE Trans. Commun.* 61 (10), 4391–4403. doi:10.1109/tcomm.2013.090513.120848
- Junyi Wang, J., Zhou Lan, Z., Chang-woo Pyo, C.-w., Baykas, T., Chin-sean Sum, C.-s., Rahman, M. A., et al. (2009). Beam Codebook Based Beamforming Protocol for Multi-Gbps Millimeter-Wave Wpan Systems. *IEEE J. Select. Areas Commun.* 27 (8), 1390–1399. doi:10.1109/jsac.2009.091009
- Lee, J., Gil, G.-T., and Lee, Y. H. (2016). Channel Estimation via Orthogonal Matching Pursuit for Hybrid MIMO Systems in Millimeter Wave Communications. *IEEE Trans. Commun.* 64 (6), 2370–2386. doi:10.1109/tcomm.2016.2557791
- Li, W., Liao, W., and Fannjiang, A. (2020). Super-resolution Limit of the Esprit Algorithm. *IEEE Trans. Inform. Theor.* 66 (7), 4593–4608. doi:10.1109/tit.2020.2974174
- Liao, A., Gao, Z., Wu, Y., Wang, H., and Alouini, M.-S. (2017). 2d Unitary Esprit Based Super-resolution Channel Estimation for Millimeter-Wave Massive MIMO with Hybrid Precoding. *IEEE Access* 5, 24747–24757. doi:10.1109/access.2017.2768579
- Ma, W., Qi, C., and Li, G. Y. (2020). High-resolution Channel Estimation for Frequency-Selective Mmwave Massive MIMO Systems. *IEEE Trans. Wireless Commun.* 19 (5), 3517–3529. doi:10.1109/twc.2020.2974728
- Mathews, C. P., Haardt, M., and Zoltowski, M. D. (1996). Performance Analysis of Closed-form, Esprit Based 2-d Angle Estimator for Rectangular Arrays. *IEEE Signal. Process. Lett.* 3 (4), 124–126. doi:10.1109/97.489068
- Méndez-Rial, R., Rusu, C., González-Prelcic, N., Alkhateeb, A., and Heath, R. W. (2016). Hybrid MIMO Architectures for Millimeter Wave Communications:

- Phase Shifters or Switches. *Ieee Access* 4, 247–267. doi:10.1109/access.2015.2514261
- Park, S., Ali, A., González-Prelcic, N., and Heath, R. W. (2019). Spatial Channel Covariance Estimation for Hybrid Architectures Based on Tensor Decompositions. *IEEE Trans. Wireless Commun.* 19 (2), 1084–1097.
- Park, S., and Heath, R. W. (2018). Spatial Channel Covariance Estimation for the Hybrid MIMO Architecture: A Compressive Sensing-Based Approach. *IEEE Trans. Wireless Commun.* 17 (12), 8047–8062. doi:10.1109/twc.2018.2873592
- Raghavan, V., and Sayeed, A. M. (2010). Sublinear Capacity Scaling Laws for Sparse MIMO Channels. *IEEE Trans. Inf. Theor.* 57 (1), 345–364.
- Rakhimov, D., Zhang, J., de Almeida, A., Nadeev, A., and Haardt, M. (2019). “Channel Estimation for Hybrid Multi-Carrier Mmwave MIMO Systems Using 3-d Unitary Tensor-Esprit in Dft Beamspace,” in 2019 53rd Asilomar Conference on Signals, Systems, and Computers (IEEE), 447–451. doi:10.1109/IEEECONF44664.2019.9048951
- Rodríguez-Fernández, J., González-Prelcic, N., Venugopal, K., and Heath, R. W. (2018). Frequency-domain Compressive Channel Estimation for Frequency-Selective Hybrid Millimeter Wave MIMO Systems. *IEEE Trans. Wireless Commun.* 17 (5), 2946–2960. doi:10.1109/twc.2018.2804943
- Roemer, F., Haardt, M., and Del Galdo, G. (2014). Analytical Performance Assessment of Multi-Dimensional Matrix-And Tensor-Based Esprit-type Algorithms. *IEEE Trans. Signal Process.* 62 (10), 2611–2625.
- Roh, W., Seol, J.-Y., Park, J., Lee, B., Lee, J., Kim, Y., et al. (2014). Millimeter-wave Beamforming as an Enabling Technology for 5g Cellular Communications: Theoretical Feasibility and Prototype Results. *IEEE Commun. Mag.* 52 (2), 106–113. doi:10.1109/mcom.2014.6736750
- Roy, R., and Kailath, T. (1989). Esprit-estimation of Signal Parameters via Rotational Invariance Techniques. *IEEE Trans. Acoust. Speech, Signal Process.* 37 (7), 984–995. doi:10.1109/29.32276
- Sarangi, P., Shahsavari, S., and Pal, P. (2020). “Robust Doa and Subspace Estimation for Hybrid Channel Sensing,” in 2020 54th Asilomar Conference on Signals, Systems, and Computers (IEEE), 236–240. doi:10.1109/IEEECONF51394.2020.9443309
- Schmidt, R. (1986). Multiple Emitter Location and Signal Parameter Estimation. *IEEE Trans. Antennas Propagat.* 34 (3), 276–280. doi:10.1109/tap.1986.1143830
- Steinwandt, J., Roemer, F., Haardt, M., and Galdo, G. D. (2017). Performance Analysis of Multi-Dimensional Esprit-type Algorithms for Arbitrary and Strictly Non-circular Sources with Spatial Smoothing. *IEEE Trans. Signal Process.* 65 (9), 2262–2276. doi:10.1109/tsp.2017.2652388
- Stewart, G. W. (1990). *Matrix Perturbation Theory*.
- Van Trees, H. L. (2004). *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons.
- Venugopal, K., Alkhateeb, A., Heath, R. W., and Prelcic, N. G. (2017). “Time-domain Channel Estimation for Wideband Millimeter Wave Systems with Hybrid Architecture,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE), 6493–6497. doi:10.1109/ICASSP.2017.7953407
- Wedin, P. Å. (1972). Perturbation Bounds in Connection with Singular Value Decomposition. *Bit* 12 (1), 99–111. doi:10.1007/bf01932678
- Wen, F., Garcia, N., Kulmer, J., Witrisal, K., and Wymeersch, H. (2018). “Tensor Decomposition Based Beamspace Esprit for Millimeter Wave MIMO Channel Estimation,” in 2018 IEEE Global Communications Conference (GLOBECOM) (IEEE), 1–7. doi:10.1109/glocom.2018.8647176
- Yu, X., Shen, J.-C., Zhang, J., and Letaief, K. B. (2016). Alternating Minimization Algorithms for Hybrid Precoding in Millimeter Wave MIMO Systems. *IEEE J. Sel. Top. Signal. Process.* 10 (3), 485–500. doi:10.1109/jstsp.2016.2523903
- Zhang, J., and Haardt, M. (2017). “Channel Estimation for Hybrid Multi-Carrier Mmwave MIMO Systems Using Three-Dimensional Unitary Esprit in Dft Beamspace,” in 2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP) (IEEE), 1–5. doi:10.1109/camsap.2017.8313174
- Zhang, J., Rakhimov, D., and Haardt, M. (2021). Gridless Channel Estimation for Hybrid Mmwave MIMO Systems via Tensor-Esprit Algorithms in Dft Beamspace. *IEEE J. Sel. Top. Signal. Process.* 15 (3), 816–831. doi:10.1109/jstsp.2021.3063908
- Zoltowski, M. D., Haardt, M., and Mathews, C. P. (1996). Closed-form 2-d Angle Estimation with Rectangular Arrays in Element Space or Beamspace via Unitary Esprit. *IEEE Trans. Signal. Process.* 44 (2), 316–328. doi:10.1109/78.485927

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Shahsavari, Sarangi and Pal. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A. PROOF OF AUXILIARY LEMMAS FOR THEOREM 1

Our proof follows similar arguments as [1] with necessary modifications for beamspace and multi-snapshot scenario. For completeness, we provide all auxiliary lemmas used.

Preliminaries

Let $\mathbf{S}_1, \mathbf{S}_2$ be any orthonormal bases for $\mathcal{R}(\mathbf{U}_y)$ and $\mathcal{R}(\hat{\mathbf{U}}_y)$, respectively. The principal (or canonical) angles between the subspaces $\mathcal{R}(\mathbf{U}_y)$ and $\mathcal{R}(\hat{\mathbf{U}}_y)$ are defined as the $\Theta(\mathbf{S}_1, \mathbf{S}_2) := [\omega_1, \omega_2, \dots, \omega_S]^T$ where $\omega_k \in [0, \pi/2]$ satisfies:

$$\cos(\omega_i) = \sigma_i(\mathbf{S}_1^H \mathbf{S}_2) \quad (48)$$

We consider the SVD of $\mathbf{S}_1^H \mathbf{S}_2 = \tilde{\mathbf{U}} \tilde{\Sigma} \tilde{\mathbf{V}}^H$. Since ESPRIT is invariant to the exact choice of the basis, for our analysis we will consider the orthonormal bases for $\mathcal{R}(\mathbf{U}_y)$ and $\mathcal{R}(\hat{\mathbf{U}}_y)$ as $\mathbf{U}_y = \mathbf{S}_1 \tilde{\mathbf{U}}$, and $\hat{\mathbf{U}}_y = \mathbf{S}_2 \tilde{\mathbf{V}}$. In this case, it can be verified that the principal angles defined in (1) can be written as:

$$\cos(\omega_i) = |\mathbf{u}_i^H \hat{\mathbf{u}}_i|$$

Here we assumed that the singular vectors are ordered such that $\omega_1 \geq \omega_2 \geq \dots \geq \omega_S$. We also denote

$$\sin(\Theta(\mathbf{U}_y, \hat{\mathbf{U}}_y)) := [\sin(\omega_1), \sin(\omega_2), \dots, \sin(\omega_S)]^T$$

The augmented noise matrix is given by:

$$\mathbf{N}_s := \begin{bmatrix} \mathbf{N}_1 \\ \mathbf{N}_2 \end{bmatrix}$$

where $\mathbf{N}_1, \mathbf{N}_2 \in \mathbb{C}^{M-1 \times T}$ represent matrices formed by selecting the first $M-1$ rows and last $M-1$ rows of \mathbf{N} , respectively. Let $\tilde{\mathbf{N}} = \mathbf{W}^H \mathbf{N}_s$, we have the following bound:

$$\begin{aligned} \|\tilde{\mathbf{N}}\|_2^2 &\leq \|\mathbf{W}\|_2^2 (\|\mathbf{N}_1\|_2^2 + \|\mathbf{N}_2\|_2^2) \\ &\leq 2\|\mathbf{W}\|_2^2 \|\mathbf{N}\|_2^2 \end{aligned} \quad (49)$$

where the first inequality follows from the fact that $\|\mathbf{N}_s\|_2^2 \leq \|\mathbf{N}_1\|_2^2 + \|\mathbf{N}_2\|_2^2$, and the second inequality holds since both $\mathbf{N}_1, \mathbf{N}_2$ are submatrices of \mathbf{N} .

For any matrix \mathbf{F} , we adopt the notation $\sigma_{\max}(\mathbf{F}) := \|\mathbf{F}\|_2$, and $\sigma_{\min}(\mathbf{F}) := 1/\|\mathbf{F}^\dagger\|_2$. We first use Wedin's theorem [2] to bound $\|\mathbf{U}_y - \hat{\mathbf{U}}_y\|_2$.

Lemma 1. (Wedin's Theorem [2]). Consider matrices $\mathbf{A}, \mathbf{B}, \mathbf{N} \in \mathbb{C}^{M \times N}$ such that

$$\mathbf{B} = \mathbf{A} + \mathbf{N}$$

Consider the Singular Value Decompositions of \mathbf{A} and \mathbf{B} :

$$\begin{aligned} \mathbf{A} &= [\mathbf{U}_1 \ \mathbf{U}_0] \begin{bmatrix} \Sigma_1 & \\ & \Sigma_0 \end{bmatrix} \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_0 \end{bmatrix}^H \\ \mathbf{B} &= [\tilde{\mathbf{U}}_1 \ \tilde{\mathbf{U}}_0] \begin{bmatrix} \tilde{\Sigma}_1 & \\ & \tilde{\Sigma}_0 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{V}}_1 \\ \tilde{\mathbf{V}}_0 \end{bmatrix}^H \end{aligned}$$

where $\mathbf{U}_1 \in \mathbb{C}^{M \times L}$, $\tilde{\mathbf{U}}_1 \in \mathbb{C}^{M \times L}$ consist of the L principal singular vectors of \mathbf{A} and \mathbf{B} , respectively. Define $\mathbf{A}_1 := \mathbf{U}_1 \Sigma_1 \mathbf{V}_1^H$,

$\mathbf{A}_0 := \mathbf{U}_0 \Sigma_0 \mathbf{V}_0^H$, $\mathbf{B}_1 := \tilde{\mathbf{U}}_1 \tilde{\Sigma}_1 \tilde{\mathbf{V}}_1^H$, $\mathbf{B}_0 := \tilde{\mathbf{U}}_0 \tilde{\Sigma}_0 \tilde{\mathbf{V}}_0^H$. If $\sigma_{\max}(\mathbf{A}_0) \leq \alpha$ and $\sigma_{\min}(\mathbf{B}_1) \geq \alpha + \delta$ for some $\alpha \geq 0$ and $\delta > 0$, the following holds

$$\|\sin \Theta(\mathcal{R}(\mathbf{A}_1), \mathcal{R}(\mathbf{B}_1))\|_\infty \leq \frac{\max\{\|\mathbf{N} \mathbf{V}_1\|_2, \|\mathbf{N}^H \mathbf{U}_1\|_2\}}{\delta}$$

Lemma 2. Consider the matrices $\mathbf{A}, \mathbf{B}_1, \mathbf{U}_1, \mathbf{V}_1$ defined in Lemma 1. If $\text{rank}(\mathbf{A}) = L$, and $\|\mathbf{N}\|_2 \leq \sigma_L(\mathbf{A})/2$, the following holds

$$\|\sin \Theta(\mathcal{R}(\mathbf{A}), \mathcal{R}(\mathbf{B}_1))\|_\infty \leq \frac{2 \max\{\|\mathbf{N} \mathbf{V}_1\|_2, \|\mathbf{N}^H \mathbf{U}_1\|_2\}}{\sigma_L(\mathbf{A})}$$

Proof. Note that since $\text{rank}(\mathbf{A}) = L$, we have $\mathbf{A}_0 = \mathbf{0}$, and $\sigma_{\min}(\mathbf{A}) = \sigma_L(\mathbf{A})$. Using Weyl's theorem [3] for matrix perturbation, we can write

$$\sigma_{\min}(\mathbf{B}_1) \geq \sigma_{\min}(\mathbf{A}) - \|\mathbf{N}\|_2 \geq \frac{\sigma_L(\mathbf{A})}{2}$$

where the last inequality follows from the assumption $\|\mathbf{N}\|_2 \leq \sigma_L(\mathbf{A})/2$. The conditions of Lemma 1 are satisfied with $\alpha = 0$ and $\delta = \sigma_L(\mathbf{A})$ completing the proof of Lemma 2. \square

We will also be using the following standard result from [4, Pg. 36].

Lemma 3. For any matrices $\mathbf{A} \in \mathbb{C}^{M \times K}$, and $\mathbf{B} \in \mathbb{C}^{K \times T}$, ($M > K$) where $\text{rank}(\mathbf{A}) = K$, we have

$$\sigma_K(\mathbf{AB}) \geq \sigma_K(\mathbf{A})\sigma_K(\mathbf{B})$$

Lemma 4. Let $\hat{\mathbf{Y}} = \mathbf{B}\mathbf{X} + \tilde{\mathbf{N}}$, where $\text{Rank}(\mathbf{B}\mathbf{X}) = S$. Consider the Singular Value Decompositions: $\mathbf{B}\mathbf{X} = \mathbf{U}_y \Sigma_y \mathbf{V}_y^H$, $\hat{\mathbf{Y}} = [\hat{\mathbf{U}}_y \ \hat{\mathbf{U}}_n] \hat{\Sigma}_y [\hat{\mathbf{V}}_y^H \ \hat{\mathbf{V}}_n^H]^H$, where $\mathbf{U}_y, \hat{\mathbf{U}}_y \in \mathbb{C}^{2N \times S}$ consists of the S principle singular vectors. Assuming that the noise is bounded as $\|\tilde{\mathbf{N}}\|_2 \leq \sigma_S(\mathbf{B})\sigma_S(\mathbf{X})/2$, the following holds

$$\|\mathbf{U}_y - \hat{\mathbf{U}}_y\|_2 \leq \frac{2\sqrt{2S}\|\tilde{\mathbf{N}}\|_2}{\sigma_S(\mathbf{B})\sigma_S(\mathbf{X})} \quad (50)$$

Proof. When the noise $\tilde{\mathbf{N}}$ is bounded by $\|\tilde{\mathbf{N}}\|_2 \leq \sigma_S(\mathbf{B})\sigma_S(\mathbf{X})/2 \leq \sigma_S(\mathbf{B}\mathbf{X})/2$, the assumptions of Lemma 2 are satisfied for $L = S$, which implies

$$\begin{aligned} \|\sin \Theta(\mathbf{U}_y, \hat{\mathbf{U}}_y)\|_\infty &\leq \frac{2 \max\{\|\tilde{\mathbf{N}} \mathbf{V}_y\|_2, \|\tilde{\mathbf{N}}^H \mathbf{U}_y\|_2\}}{\sigma_S(\mathbf{B}\mathbf{X})} \\ &\leq \frac{2 \max\{\|\tilde{\mathbf{N}} \mathbf{V}_y\|_2, \|\tilde{\mathbf{N}}^H \mathbf{U}_y\|_2\}}{\sigma_S(\mathbf{B})\sigma_S(\mathbf{X})} \end{aligned}$$

Using the fact $\|\mathbf{V}_y\|_2 = 1$, $\|\mathbf{U}_y\|_2 = 1$, we have

$$\|\sin \Theta(\mathbf{U}_y, \hat{\mathbf{U}}_y)\|_\infty \leq \frac{2\|\tilde{\mathbf{N}}\|_2}{\sigma_S(\mathbf{B})\sigma_S(\mathbf{X})} \quad (51)$$

Now, under the canonical basis assumption, we have $\|\sin \Theta(\mathbf{U}_y, \hat{\mathbf{U}}_y)\|_\infty = \sin(\omega_1)$ and for $i = 1, 2, \dots, S$

$$\|\hat{\mathbf{u}}_i - \mathbf{u}_i\|_2^2 = 2(1 - \cos \omega_k) \leq 2(1 - \cos^2 \omega_k) \leq 2 \sin^2 \omega_k$$

Therefore,

$$\begin{aligned} \|\mathbf{U}_y - \hat{\mathbf{U}}_y\|_2 &\leq \|\mathbf{U}_y - \hat{\mathbf{U}}_y\|_F = \left(\sum_{i=1}^S \|\hat{\mathbf{u}}_i - \mathbf{u}_i\|_2^2 \right)^{1/2} \\ &\leq (2S \sin^2 \omega_1)^{1/2} = \sqrt{2S} \sin \omega_1 \end{aligned} \quad (52)$$

The proof is completed by combining (52) and (51). \square

Lemma 5. Consider the measurement model in (14). If $\text{rank}(\mathbf{B}\mathbf{X}) = S$, and $\|\mathbf{U}_y - \hat{\mathbf{U}}_y\|_2 \leq \sigma_S(\mathbf{U}_1)/2$, then

$$\|\Psi - \hat{\Psi}\|_2 \leq \frac{7\|\mathbf{U}_y - \hat{\mathbf{U}}_y\|_2}{\sigma_S(\mathbf{U}_1)^2} \quad (53)$$

Proof. Notice that

$$\begin{aligned} \|\Psi - \hat{\Psi}\|_2 &= \|(\hat{\mathbf{U}}_1^\dagger - \mathbf{U}_1^\dagger)\hat{\mathbf{U}}_2 + \mathbf{U}_1^\dagger(\hat{\mathbf{U}}_2 - \mathbf{U}_2)\|_2 \\ &\leq \|(\hat{\mathbf{U}}_1^\dagger - \mathbf{U}_1^\dagger)\|_2 \|\hat{\mathbf{U}}_2\|_2 + \|\mathbf{U}_1^\dagger\|_2 \|(\hat{\mathbf{U}}_2 - \mathbf{U}_2)\|_2 \\ &\leq \|(\hat{\mathbf{U}}_1^\dagger - \mathbf{U}_1^\dagger)\|_2 + \|\mathbf{U}_1^\dagger\|_2 \|(\hat{\mathbf{U}}_y - \mathbf{U}_y)\|_2 \end{aligned}$$

where the last inequality follows from the fact that $\hat{\mathbf{U}}_2, \hat{\mathbf{U}}_2 - \mathbf{U}_2$ are submatrices of $\hat{\mathbf{U}}_y$ and $\hat{\mathbf{U}}_y - \mathbf{U}_y$, respectively. Therefore, we have $\|\hat{\mathbf{U}}_2\|_2 \leq \|\hat{\mathbf{U}}_y\|_2 = 1$, and $\|\hat{\mathbf{U}}_2 - \mathbf{U}_2\|_2 \leq \|\hat{\mathbf{U}}_y - \mathbf{U}_y\|_2$. By the assumption in this lemma, we have,

$$\|\hat{\mathbf{U}}_1 - \mathbf{U}_1\|_2 \leq \|\hat{\mathbf{U}}_y - \mathbf{U}_y\|_2 \leq \frac{\sigma_S(\mathbf{U}_1)}{2} \quad (54)$$

We use a result from [5, Theorem 3.2] which states that a matrix \mathbf{F} with rank S , and its perturbed matrix $\tilde{\mathbf{F}} = \mathbf{F} + \mathbf{E}$ satisfy the following inequality:

$$\|\mathbf{F}^\dagger - \tilde{\mathbf{F}}^\dagger\|_2 \leq \frac{3\|\mathbf{E}\|_2}{\sigma_S(\mathbf{F})(\sigma_S(\mathbf{F}) - \|\mathbf{E}\|_2)}$$

provided the perturbation satisfies $\|\mathbf{E}\|_2 < \sigma_S(\mathbf{F})$. We use this result by substituting \mathbf{F} with \mathbf{U} , and $\tilde{\mathbf{F}}$ with $\hat{\mathbf{U}}_1$.

From (54), the perturbation condition is satisfied and this result leads to:

$$\begin{aligned} \|\hat{\mathbf{U}}_1^\dagger - \mathbf{U}_1^\dagger\|_2 &\leq \frac{3\|\hat{\mathbf{U}}_1 - \mathbf{U}_1\|_2}{\sigma_S(\mathbf{U}_1)(\sigma_S(\mathbf{U}_1) - \|\hat{\mathbf{U}}_1 - \mathbf{U}_1\|_2)} \\ &\leq \frac{6\|\hat{\mathbf{U}}_y - \mathbf{U}_y\|_2}{\sigma_S(\mathbf{U}_1)^2} \end{aligned} \quad (55)$$

Therefore, we have that

$$\begin{aligned} \|\Psi - \hat{\Psi}\|_2 &\leq \left(\frac{6}{\sigma_S(\mathbf{U}_1)^2} + \frac{1}{\sigma_S(\mathbf{U}_1)} \right) \|\hat{\mathbf{U}}_y - \mathbf{U}_y\|_2 \\ &\leq \frac{7\|\hat{\mathbf{U}}_y - \mathbf{U}_y\|_2}{\sigma_S(\mathbf{U}_1)^2} \end{aligned} \quad (56)$$

Lemma 6. Consider the measurement model in (14) such that (17) holds. Then the following bound is satisfied:

$$\|\Psi - \hat{\Psi}\|_2 \leq \frac{14\sqrt{2S}\|\tilde{\mathbf{N}}\|_2}{\sigma_S(\mathbf{B})\sigma_S(\mathbf{X})\sigma_S(\mathbf{U}_1)^2} \quad (57)$$

Proof. From (17) and (49), we have

$$\|\tilde{\mathbf{N}}\|_2 \leq \frac{\sigma_S(\mathbf{B})\sigma_S(\mathbf{X})\sigma_S(\mathbf{U}_1)}{8\sqrt{2S}} \leq \frac{\sigma_S(\mathbf{B})\sigma_S(\mathbf{X})}{2} \quad (58)$$

where the second inequality follows from the fact that $\sigma_S(\mathbf{U}_1) \leq 1$ and $S \geq 1$. By applying Lemma 4, (50) holds. Now, (50) and (58) together imply that $\|\mathbf{U}_y - \hat{\mathbf{U}}_y\|_2 \leq \sigma_S(\mathbf{U}_1)/2$. This ensures that the conditions of Lemma 5 are satisfied. Combining (53) and (50) leads to the desired result. \square

Lemma 7.

$$md(\mathcal{F}, \hat{\mathcal{F}}) \leq \frac{1}{2} md(\Psi, \hat{\Psi}) \quad (59)$$

Proof. The proof follows directly from eq. (III.1) in [1] \square

Lemma 8. Consider the measurement model in (14). If $\text{rank}(\mathbf{B}\mathbf{X}) = S$, then

$$md(\mathcal{F}, \hat{\mathcal{F}}) \leq \frac{S\|\mathbf{B}\|_2}{\sigma_S(\mathbf{B})} \|\Psi - \hat{\Psi}\|_2 \quad (60)$$

Proof. Based on (9), Ψ is diagonalizable by the invertible matrix \mathbf{P} . Using Bauer-Fike theorem, [6], [4, Theorem 3.3] and Lemma 7, we have

$$md(\mathcal{F}, \hat{\mathcal{F}}) \leq \frac{1}{2} (2S - 1) \kappa(\mathbf{P}^{-1}) \|\Psi - \hat{\Psi}\|_2 \quad (61)$$

where $\kappa(\mathbf{P}^{-1}) = \|\mathbf{P}\|_2 \|\mathbf{P}^{-1}\|_2$. To bound $\kappa(\mathbf{P}^{-1})$, we use the fact that $\mathbf{U}_y = \mathbf{B}\mathbf{P}$ and $\|\mathbf{U}_y\|_2 = 1$. This implies that

$$\kappa(\mathbf{P}^{-1}) \leq \kappa(\mathbf{B}) = \frac{\|\mathbf{B}\|_2}{\sigma_S(\mathbf{B})} \quad (62)$$

References

- [1] W. Li, W. Liao, and A. Fannjiang, "Super-resolution limit of the esprit algorithm," *IEEE Transactions on Information Theory*, vol. 66, no. 7, pp. 4593–4608, 2020.
- [2] P.-A. Wedin, "Perturbation bounds in connection with singular value decomposition," *BIT Numerical Mathematics*, vol. 12, no. 1, pp. 99–111, 1972.
- [3] J. N. Franklin, *Matrix theory*. Courier Corporation, 2012.
- [4] G. W. Stewart, "Matrix perturbation theory," 1990.
- [5] P. C. Hansen, "The truncatedsvd as a method for regularization," *BIT Numerical Mathematics*, vol. 27, no. 4, pp. 534–553, 1987.
- [6] F. L. Bauer and C. T. Fike, "Norms and exclusion theorems," *Numerische Mathematik*, vol. 2, no. 1, pp. 137–141, 1960.



VIPDA: A Visually Driven Point Cloud Denoising Algorithm Based on Anisotropic Point Cloud Filtering

Tiziana Cattai, Alessandro Delfino, Gaetano Scarano and Stefania Colonnese*

Department of Information Engineering, Electronics and Telecommunications, University La Sapienza of Rome, Rome, Italy

OPEN ACCESS

Edited by:

Hagit Messer,
Tel Aviv University, Israel

Reviewed by:

Milos Brajovic,
University of Montenegro,
Montenegro
Miaohui Wang,
Shenzhen University, China

*Correspondence:

Stefania Colonnese
stefania.colonnese@uniroma1.it

Specialty section:

This article was submitted to
Statistical Signal Processing,
a section of the journal
Frontiers in Signal Processing

Received: 23 December 2021

Accepted: 15 February 2022

Published: 16 March 2022

Citation:

Cattai T, Delfino A, Scarano G and
Colonnese S (2022) VIPDA: A Visually
Driven Point Cloud Denoising
Algorithm Based on Anisotropic Point
Cloud Filtering.
Front. Sig. Proc. 2:842570.
doi: 10.3389/frsip.2022.842570

Point clouds (PCs) provide fundamental tools for digital representation of 3D surfaces, which have a growing interest in recent applications, such as e-health or autonomous means of transport. However, the estimation of 3D coordinates on the surface as well as the signal defined on the surface points (vertices) is affected by noise. The presence of perturbations can jeopardize the application of PCs in real scenarios. Here, we propose a novel visually driven point cloud denoising algorithm (VIPDA) inspired by visually driven filtering approaches. VIPDA leverages recent results on local harmonic angular filters extending image processing tools to the PC domain. In more detail, the VIPDA method applies a harmonic angular analysis of the PC shape so as to associate each vertex of the PC to suit a set of neighbors and to drive the denoising in accordance with the local PC variability. The performance of VIPDA is assessed by numerical simulations on synthetic and real data corrupted by Gaussian noise. We also compare our results with state-of-the-art methods, and we verify that VIPDA outperforms the others in terms of the signal-to-noise ratio (SNR). We demonstrate that our method has strong potential in denoising the point clouds by leveraging a visually driven approach to the analysis of 3D surfaces.

Keywords: point cloud, denoising, non-Euclidean domain, angular harmonic filtering, graph signal processing

1 INTRODUCTION

Digital representation of real 3D surfaces has a crucial importance in a variety of cutting-edge applications, such as autonomous navigation (Huang J. et al., 2021), UAV fleets (Ji et al., 2021), extended reality streaming, or telesurgery (Huang T. et al., 2021). Point clouds represent 3D surfaces by means of a set of 3D locations of points on the surface. In general, those points can be acquired by active or passive techniques (Chen et al., 2021; Rist et al., 2021), in presence of random errors, and they may be associated with color and texture information as well. Point cloud denoising can in general be applied as an enhancement stage at the decoder side of an end-to-end communication system, involving volumetric data, for e.g., for extended reality or mixed reality services. Although lossless compression of point clouds is feasible (Ramalho et al., 2021), color point cloud lossy coding based on 2D point cloud projection (Xiong et al., 2021) is increasingly relevant both in sensor networks (de Hoog et al., 2021) and autonomous systems (Sun et al., 2020). Nonlocal estimation solutions (Zhu et al., 2022) or color-based (Irfan and Magli 2021b) solutions as well as solutions for point cloud sequences (Hu et al., 2021a) have been proposed.

The extraction of visually relevant features on point cloud is needed for tasks as pattern recognition, registration, compression and quality evaluation (Yang et al., 2020; Diniz et al., 2021), and semantic segmentation. Point cloud (PC) processing has been widely investigated, and many of the proposed processing methods are based on the geometric properties (Hu et al., 2021b;

Erçelik et al., 2021). Still, feature extraction on PC is mostly focused on information related to the point cloud shape. In addition, new PC acquisition systems for surveillance (Dai et al., 2021) or extended reality (XR) (Yu et al., 2021) require processing tools operating both on geometry and texture. Shape and texture processing needs development of new tools because of the non-Euclidean nature of the real surfaces modeled by a point cloud. In this direction, few studies in the literature simultaneously leverage both geometry and texture information. Furthermore, in the context of classical image processing, several effective tools have been inspired by the human visual systems (HVSs), which process both texture and shape, and it is sensitive to angular patterns such as edges, forks, and corners (Beghdadi et al., 2013). In particular, two-dimensional circular harmonic functions (CHFs) have been investigated for visually driven image processing and specifically for angular pattern detection. CHFs have been successfully applied to interpolation (Colonnese et al., 2013), deconvolution (Colonnese et al., 2004), and texture synthesis (Campisi and Scarano 2002). On the contrary, point cloud processing lacks HVS-inspired processing tools, which can, in principle, provide alternative perspectives.

In this article, we leverage a class of point cloud multiscale anisotropic harmonic filters (MAHFs) inspired by HVS. MAHFs were recently introduced in our conference article (Conti et al. (2021)). First, we recall the MAHF definition and describe their local anisotropic behavior that highlights directional components of the point cloud texture or geometry. In addition, we show their applicability to both geometric and textured PC data. Second, we illustrate how MAHF can be applied to visually driven PC denoising problems. Denoising is a crucial preprocessing step for many further point cloud processing techniques. In real acquisition scenarios, the perturbations on the PC vertices or on the associated signal severely affect the PC usability. MAHF is used to drive an iterative denoising algorithm so as to adapt the restoration to the local information. The proposed method differs from other competitors (Zhu et al., 2022 and the references) in linking the denoising with visually relevant features, as estimated by suitable anisotropic filtering in the vertex domain. We test the performance of the visually driven point cloud denoising algorithm (VIPDA) on synthetic and real data from the public database (Turk and Levoy 1994; d'Eon et al., 2017). Specifically, considering different signal-to-noise-ratios by adding Gaussian noise to the original data, we verify that our method outperforms state-of-the-art alternatives in denoising data.

The structure of the article is as follows. In **Section 2**, we review a particular class of HVS-inspired image filters, namely the circular harmonic filters, which are needed to introduce our point cloud filtering approach. In **Section 3**, we present a class of multiscale anisotropic filters, formerly introduced in Conti et al. (2021), and we illustrate their relation with visually driven image filters. In **Section 5**, we present the visually driven point cloud denoising algorithm (VIPDA) based on the proposed manifold filters. In **Section 6**, we show by numerical simulations that the VIPDA outperforms state-of-the-art competitors. **Section 7** concludes the article.

2 CIRCULAR HARMONIC FUNCTIONS FOR HVS-BASED IMAGE FILTERING: A REVIEW

Before the introduction of the MAHFs, a step back is necessary in order to contextualize the research problems by investigating other filter methods in the Euclidean domain.

In several important applications in the field of image processing, circular harmonic functions (CHFs) have been used (Panci et al., 2003; Colonnese et al., 2010). As mentioned previously, CHFs have been widely applied in image processing applications because they are able to detect relevant image features, such as edges, lines, and crosses, i.e., they perform the analysis in an analogous way to the behavior of the HVS during the pre-attentive step. It is important to remark here that the results of the order-1 CHFs are complex images, in which the module corresponds to the edge magnitude while the phase describes the orientation. Taken together, this filtering procedure returns precious information about the structures of the output image; in fact, it underlines the edges by simultaneously measuring their intensity and direction. The interest in CHFs also stems from the fact that they can be integrated within an invertible filter bank, thereby being exploited for suitable processing, for e.g., image enhancement, in the CHF-transformed domain (Panci et al., 2003).

CHFs' properties relate to the specific way in which they characterize the information belonging to two points. In fact, they encode the distance as well as the geometric direction that joins them.

These aspects are evident in the mathematical formulation of CHFs. Let us consider the 2D domain of the continuous CHF described by the polar coordinates (r, ϑ) that, respectively, represent the distance from the origin and the angle with the reference x axis. The CHF of order k is the complex filter defined as:

$$h^{(k)}(r, \vartheta) = g_k(r)e^{ik\vartheta}, \quad (1)$$

where the influence of the radial (r) and the angular (ϑ) contributions are separated by the two factors. With the aim of preserving the isomorphism with the frequency space, the functions $g_k(r)$ in **Eq. 1** are usually isotropic Gaussian kernels. The variable k defines the angular structure of the model. For $k = 0$, the zero-order CHF returns output as a real image, represented by the low-pass version of the original one. As a general consideration, when the order k increases, CHFs are able to identify more and more complex structures on the images, such as edges (for $k = 1$), lines (for $k = 2$), forks (for $k = 3$), and crosses (for $k = 4$). We can see the effect of increasing the k order of the heat kernel on a sphere in Conti et al. (2021).

Based on the definition of CHF and the introduction of a scale parameter α , circular harmonic wavelet (CHW) (Jacovitti and Neri 2000) of order k can be introduced and can be typically applied in the context of multi-resolution problems.

Finally, it is worth observing that the CHF output has been shown to be related to the Fisher information of the input w.r.t rotation and translation parameters. The Fisher information of an image w.r.t. shift/rotation estimation is associated with the power of the image first derivative w.r.t. the parameter under concern

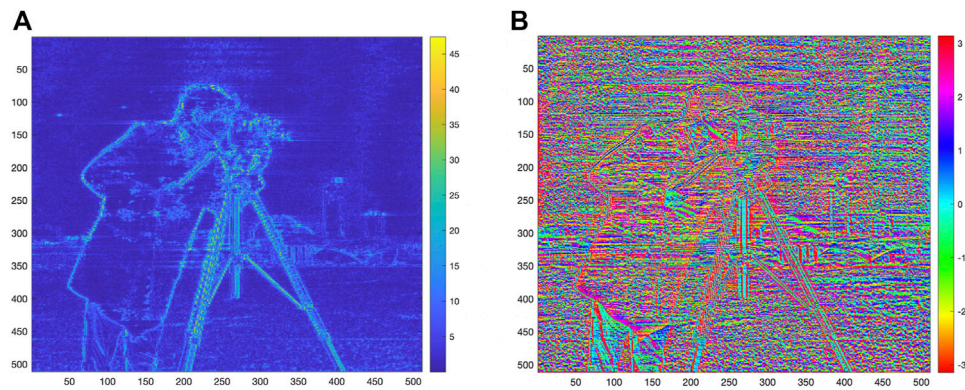


FIGURE 1 | Results of CHF application on the “cameraman image” (with $k = 1$). In panel **(A)**, we have the module of the CHF $|h^{(1)}|$, and in panel **(B)**, we have the associated phase $\angle h^{(1)}$.

TABLE 1 | Table of main notation.

Notation	Description
\mathcal{M}	Manifold
\mathcal{G}	Graph associated to the point cloud
\mathbf{A}	Adjacency matrix
\mathbf{D}	Degree matrix
\mathbf{L}	Laplacian matrix
λ, \mathbf{u}	Eigenvalues and eigenvectors
\mathbf{p}_0, \mathbf{p}	Coordinates of point on the continuous manifold
$\mathbf{p}_i, \mathbf{p}_j$	Coordinates of point on the discrete manifold
\mathbf{q}_i	Noisy coordinates of points on the discrete manifold
$\tilde{\mathbf{p}}_i$	Reconstructed coordinates after denoising
$\mathbf{n}_p, \mathbf{n}_{p_i}$	Direction orthogonal to the continuous and discrete manifold
$\phi^{(k)}$	MAHF of k order on the continuous Manifold
$\varphi^{(k)}$	MAHF of k order on the discrete Manifold

(Friedlander 1984). The CHF in the 2D account for a local derivative of the signal has been shown to be related to the Fisher information w.r.t. localization parameters (Neri and Jacovitti 2004).

In order to show a visual example of the application of CHF on real data, we consider the “cameraman image,” and we apply first-order CHF as an example to represent the effect of CHF on a real image. We report the results in **Figure 1**, in which we can see the module in panel A and the phase in panel B.

Stemming on these studies on directional harmonic analysis of 2D signals, we introduce in the following sections the multiscale anisotropic harmonic filters to be adopted in the non-Euclidean manifold domain.

3 MULTISCALE ANISOTROPIC HARMONIC FILTERS

Although the HVS is very complex in nature, its low-level behavior, as determined by the primary visual cortex, is well characterized by being bandpass and orientation selective (Wu et al., 2017). Therefore, the CHF mimics these features on 2D images, and in this section, we show how to extend this behavior

to manifold filters. Specifically, in this section, we describe a new class of visually driven filters operating on a manifold in the 3D space; the preliminary results on such filters appear in Conti et al.(2021). We extend the presentation in Conti et al. (2021) by an in-depth analysis of their relation with the CHF and by providing new results about their applications to point cloud filtering.

Our general idea consists in the extension of CHFs to 2D manifolds embedded in 3D domains. In this direction, the two key points to adapt to this new scenario are as follows: 1) we need to define a smoothing kernel that corresponds to the isotropic Gaussian smoothing in the 2D case; and 2) we have to identify an angular measurement on the surface of the manifold in the 3D space.

In the following sections, we elaborate on the filters description first in the case of a 2D manifold defined in a continuous 3D domain and then in the case of its discretized version, as represented by a point cloud. The main notation is reported in **Table 1**.

3.1 MAHF on Manifolds

We first introduce the multiscale anisotropic harmonic filter (MAHF) on a continuous manifold \mathcal{M} in \mathbb{R}^3 . The first step consists in the definition of a smoothing kernel, which is necessary to adapt to **Eq. 1** in this scenario. The smoothing kernel should account on the intrinsic (non-Euclidean) distance between a point \mathbf{p}_0 and a different point \mathbf{p} on the manifold surface.

To this aim, we resort to the heat kernel that describes the diffusion of the heat from a point-wise source located at a point \mathbf{p}_0 on the manifold to a generic other manifold point \mathbf{p} , after the time t . In formulas, the heat kernel $K_t: \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R}$ is found as the fundamental solution of the heat propagation equation¹ under the initial condition $f_0(\mathbf{p}) = \delta(\mathbf{p} - \mathbf{p}_0)$.

¹Let Δ denote the Laplace–Bertrami operator, i.e., a linear operator computing the sum of directional derivatives of a function defined on the manifold. The heat propagation equation is written as:

$$\Delta f(\mathbf{p}, t) = -\partial_t f(\mathbf{p}, t), f(\mathbf{p}, t = 0) = f_0(\mathbf{p}), \quad (2)$$

where $f(\mathbf{p}, t)$ is the solution under initial condition $f_0(\mathbf{p})$.

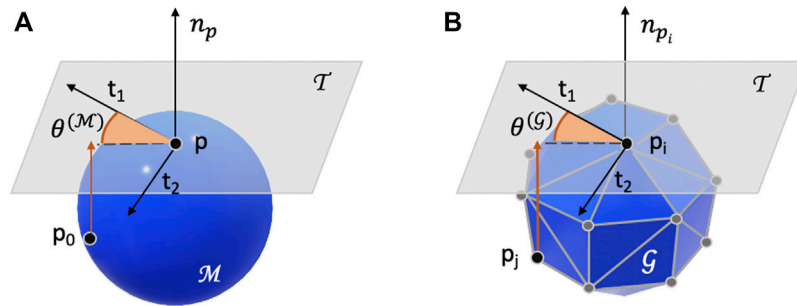


FIGURE 2 | Graphical representation of angles in the 3D space. In panel (A), we represent the continuous manifold \mathcal{M} , in which we highlight the angle $\vartheta^{(\mathcal{M})}$ in orange. In panel (B), we have the discrete manifold on which a graph \mathcal{G} is defined. The angle $\vartheta^{(\mathcal{G})}$ is plotted in orange.

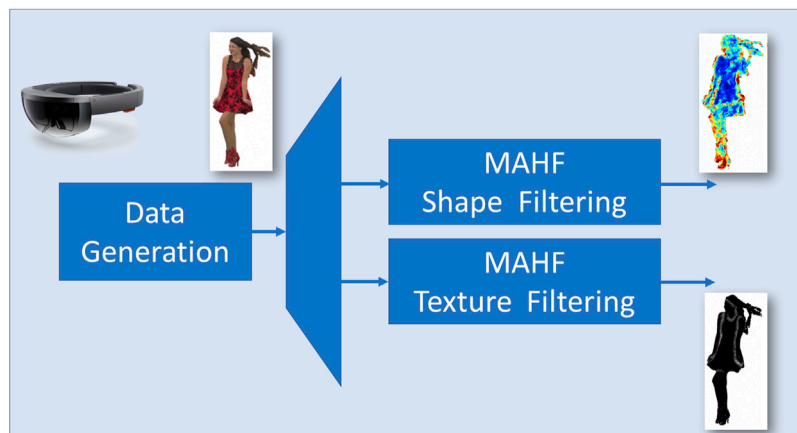


FIGURE 3 | Application of MAHF to texture (luminance) and shape (point cloud normals).

With these positions, given two points \mathbf{p}_0 and \mathbf{p} on the manifold surface, the heat kernel is expressed as an infinite of suitable functions on the manifold. Specifically, let $\chi_s(\mathbf{p}, s = 0, \dots)$ be the eigenfunctions of the Laplace–Bertrami (sum of directional derivatives) manifold operator. Therefore, $\chi_s: \mathcal{M} \rightarrow \mathbb{R}$. For any point pair $(\mathbf{p}, \mathbf{p}_0)$ on the manifold, the heat kernel is written as:

$$K_t^{(\mathcal{M})}(\mathbf{p}, \mathbf{p}_0) = \sum_{s=0}^{\infty} e^{-t\alpha_s} \chi_s(\mathbf{p}) \chi_s(\mathbf{p}_0). \quad (3)$$

The heat kernel $K_t^{(\mathcal{M})}(\mathbf{p}, \mathbf{p}_0)$ has the interesting property to represent a smooth function on the \mathcal{M} manifold (Hou and Qin 2012), smoothly decreasing as a bell-shaped function at a rate depending on the parameter t (Conti et al., 2021).

For what concerns angles, several definitions have been proposed in the context of convolutional neural networks on the manifold. Specifically, the polar coordinates on the geodesic can be defined using angular bins. Alternatively, a representation of points on the plane \mathcal{T} , which is the plane tangent to the manifold in the point \mathbf{p} . In this work, we define the angle $\vartheta^{(\mathcal{M})}$ in a similar way to the second approach. As represented in the panel A in **Figure 2**,

the angle $\vartheta^{(\mathcal{M})}$ corresponds to the azimuth in the spherical coordinates of the point \mathbf{p} , when the reference is the system (t_1, t_2, n) with the origin centered in \mathbf{p} and the n axis is normal to the tangent plane \mathcal{T} .

With these positions, the multiscale anisotropic harmonic filters (MAHFs) ϕ of k order and centered in \mathbf{p}_0 are defined as follows (Conti et al., 2021):

$$\phi^{(k)}(\mathbf{p}, \mathbf{p}_0) = \underbrace{K_t^{(\mathcal{M})}(\mathbf{p}, \mathbf{p}_0) \cos(k \vartheta^{(\mathcal{M})}(\mathbf{p}, \mathbf{p}_0))}_{\phi_R^{(k)}} + j \underbrace{K_t^{(\mathcal{M})}(\mathbf{p}, \mathbf{p}_0) \sin(k \vartheta^{(\mathcal{M})}(\mathbf{p}, \mathbf{p}_0))}_{\phi_I^{(k)}}, \quad (4)$$

with $K_t^{(\mathcal{M})}(\mathbf{p}, \mathbf{p}_0)$ defined as in 3 and where we recognize the real $\phi_R^{(k)}$ and imaginary $\phi_I^{(k)}$ part of the complex function $\phi^{(k)}(\mathbf{p}, \mathbf{p}_0)$.

3.2 MAHF on Point Clouds

In this subsection, we focus on the definition of MAHF on a 3D point cloud. Let us consider the graph \mathcal{G} associated with the point cloud and defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where \mathcal{V} is the set of N point

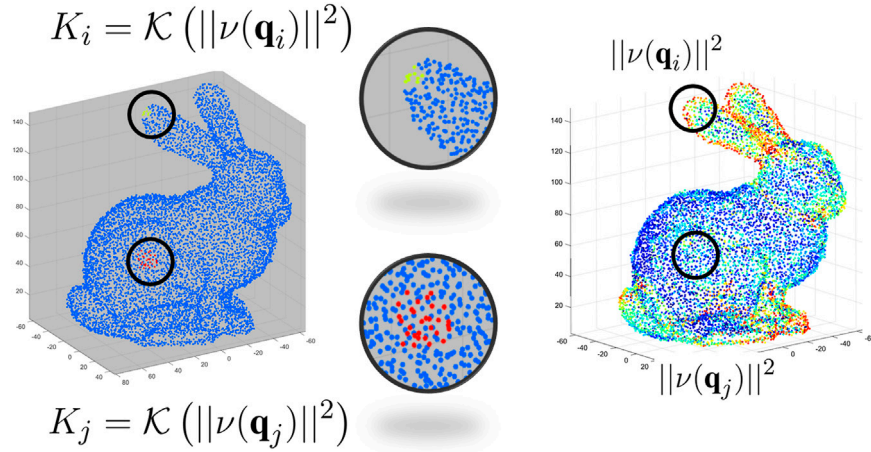


FIGURE 4 | Examples of two neighborhoods of different sizes K_i and K_j (left) around points characterized by different values of the mean square value of the MAHF output $\|\nu(\mathbf{q})\|^2$ (right). The considered point cloud is a low-resolution version of the Stanford bunny in Turk and Levoy 1994).

cloud vertices \mathbf{p}_i with $i = 1..N$, and \mathcal{E} is the set of edges (or links). The edge weights are represented by the $N \times N$ weighted adjacency matrix \mathbf{A} or by the Laplacian graph $\mathbf{L} = \mathbf{D} - \mathbf{A}$, being \mathbf{D} , the degree matrix, which is a diagonal matrix with elements on the principal diagonal computed as $d_{ii} = \sum_{j=1}^N a_{ij}$. For 3D point clouds, the edge weights are selected such that the Laplacian \mathbf{L} approximates the continuous domain Laplace–Beltrami operator (Belkin et al., 2009).

Let λ_n and \mathbf{u}_n , $n = 0, \dots, N-1$ denote the eigenvalues and eigenvectors of \mathbf{L} , respectively. In this case, the heat kernel at the i -th and j -th point cloud points $(\mathbf{p}_i, \mathbf{p}_j)$ is obtained as follows:

$$K_t^{(\mathcal{G})}(\mathbf{p}_i, \mathbf{p}_j) = \sum_{n=0}^{N-1} e^{-t\lambda_n} \mathbf{u}_n[i] \mathbf{u}_n[j], \quad (5)$$

i.e., it depends on the weighted sum of the products of the i -th and j -th coefficients of each and every Laplacian eigenvector. The eigenvectors corresponding to small eigenvalues, i.e., the low-frequency vectors of the graph Fourier transform defined on the graph, dominate the sum for large values of the parameter t .

As the continuous case, in the discrete scenario, we define the angle $\vartheta^{(\mathcal{G})}$ as the azimuth of the \mathcal{T} tangent plane to \mathbf{p}_i , as graphically represented in **Figure 2B**.

Similar to the continuous case, the multiscale anisotropic harmonic filters (MAHFs) φ of k order and centered in \mathbf{p}_i are defined as:

$$\varphi^{(k)}(\mathbf{p}_i, \mathbf{p}_j) = \underbrace{K_t^{(\mathcal{G})}(\mathbf{p}_i, \mathbf{p}_j) \cos(k \vartheta^{(\mathcal{G})}(\mathbf{p}_i, \mathbf{p}_j))}_{\varphi_R^{(k)}} + j \underbrace{K_t^{(\mathcal{G})}(\mathbf{p}_i, \mathbf{p}_j) \sin(k \vartheta^{(\mathcal{G})}(\mathbf{p}_i, \mathbf{p}_j))}_{\varphi_I^{(k)}}, \quad (6)$$

with $K_t^{(\mathcal{G})}(\mathbf{p}_i, \mathbf{p}_j)$ defined as in **Eq. 5** and where we recognize the real $\varphi_R^{(k)}$ and imaginary $\varphi_I^{(k)}$ parts of the complex function $\varphi^{(k)}(\mathbf{p}_i, \mathbf{p}_j)$.

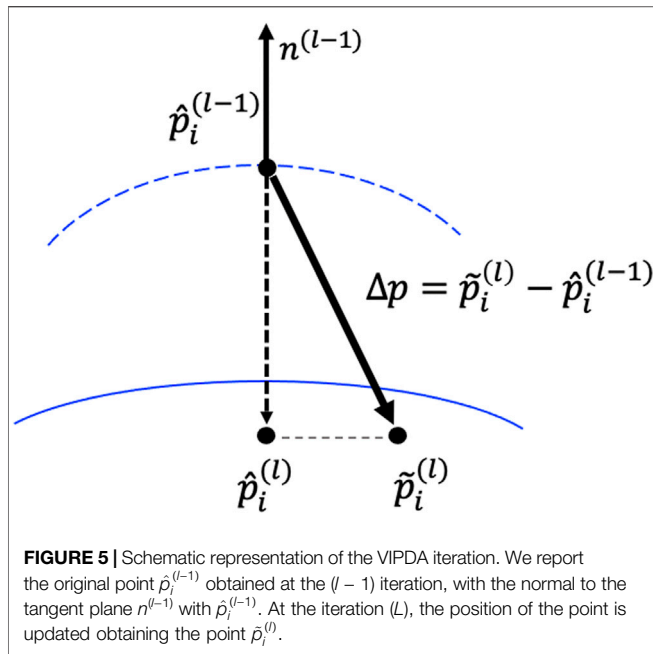
4 VISUALLY DRIVEN POINT CLOUD FILTERING: MAHF AS ANISOTROPIC ANALYSIS OF TEXTURE AND SHAPE IN POINT CLOUDS

Let us consider a real-valued D -dimensional signal on the point cloud vertices $\mathbf{s}(\mathbf{p}_i) \in \mathbb{R}^D$, $i = 0, \dots, N-1$. Applying k -th-order MAHF for the vertex domain signal on graph filtering obtains the output point cloud signal $\mathbf{r}(\mathbf{p}_i)$ as:

$$\mathbf{r}(\mathbf{p}_i) = \sum_{j=0}^{N-1} \varphi^{(k)}(\mathbf{p}_i, \mathbf{p}_j) \mathbf{s}(\mathbf{p}_j). \quad (7)$$

The filtering realized by the MAHFs performs an anisotropic harmonic angular filtering of the signal defined on the point cloud. The period of the harmonic analysis decreases as the filter order increases, and MAHF of different orders k are matched to different angular patterns of the input signal $\mathbf{s}(\mathbf{p}_i)$ $i = 0, \dots, N-1$.

MAHF applies to both texture and shape signals, depending on the choice of the input signal $\mathbf{s}(\mathbf{p}_i)$ $i = 0, \dots, N-1$. For video point clouds, the signal $\mathbf{s}(\mathbf{p}_i)$ can represent the luminance and the chrominances observed at the point \mathbf{p}_i . If this is the case, the MAHF output highlights texture patterns on the surface. On the other hand, the signal $\mathbf{s}(\mathbf{p}_i)$ $i = 0, \dots, N-1$ can be selected so as to represent geometric information. A relevant case is when the signal represents the normal to the point cloud surface at each vertex \mathbf{p}_i as follows:



$$\mathbf{s}(\mathbf{p}_i) = \begin{bmatrix} n_x(\mathbf{p}_i) \\ n_y(\mathbf{p}_i) \\ n_z(\mathbf{p}_i) \end{bmatrix} = \mathbf{n}(\mathbf{p}_i), \quad i = 0, \dots, N-1. \quad (8)$$

The application of MAHF to the signal defined as in Eq. 8 will be exploited in the following derivation of VIPDA.

To sum up, the MAHFs can be applied to different kinds of data defined on point clouds, and they provide a way to extract several point-wise shape and texture point cloud features for different values of the order k . A schematic representation of MAHF application to texture (luminance) and shape (point cloud normals) information is illustrated in Figure 3.

5 VISUALLY DRIVEN POINT CLOUD DENOISING ALGORITHM

In this section, we illustrate the VIPDA approach, based on application of the aforementioned MAHF to the problem of PC denoising.

Let us consider the case in which the point cloud vertices are observed in presence of an additive noise. Thereby, the observed coordinates are written as

$$\mathbf{q}_i = \mathbf{p}_i + \mathbf{w}_i, \quad i = 0, \dots, N-1, \quad (9)$$

where \mathbf{w}_i is an i.i.d. random noise. The denoising provides an estimate $\hat{\mathbf{p}}_i$ of the original locations \mathbf{p}_i . Let us remark that this problem is different from recovery of a signal defined at the vertices, which will be addressed in the future work.

Point cloud denoising algorithms typically leverage 1) data fidelity (Irfan and Magli 2021a), 2) manifold smoothness (low-rankedness) (Dinesh et al., 2020), and 3) local or cooperative averaging (Chen et al., 2019) objectives.

Here, the local manifold smoothness is accounted by adapting the estimator $\hat{\mathbf{p}}_i$ to the local shape variability as estimated at the first algorithm stage. Specifically, the MAHF is applied to the point cloud estimated normals $\mathbf{n}^{(0)}(\mathbf{q}_i)$ as:

$$\mathbf{v}^{(0)}(\mathbf{q}_i) = \sum_{j=0}^{N-1} \varphi^{(k)}(\mathbf{q}_i, \mathbf{q}_j) \mathbf{n}^{(0)}(\mathbf{q}_j). \quad (10)$$

Thereby, each point \mathbf{q}_i is assigned a weight related to the normal variations in its neighborhood. The key idea is that when fast variations of the normal are observed around \mathbf{q}_i , the surface smoothness is reduced, and the set of neighboring points to be exploited to compute the estimate $\hat{\mathbf{p}}_i$ should be reduced accordingly. Therefore, the size K_i of the neighborhood of the point \mathbf{q}_i to be used in the estimation stage is selected based on the MAHF-filtered signal $\mathbf{v}(\mathbf{q}_i)$. Specifically, K_i is selected as a function of the mean square value of the MAHF output, namely

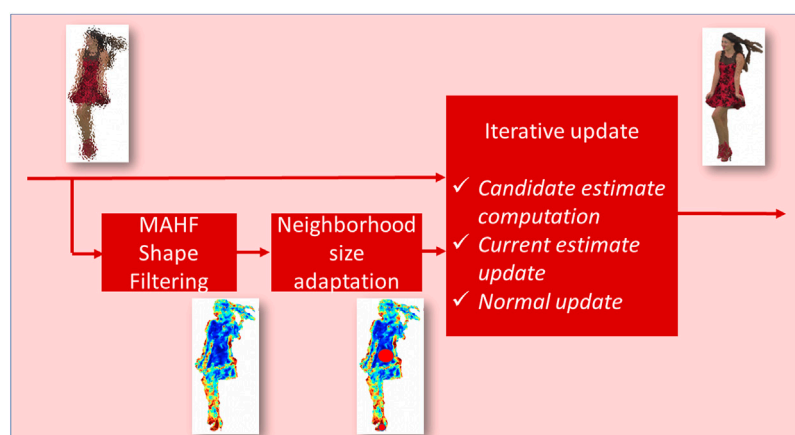


FIGURE 6 | VIPDA overview.

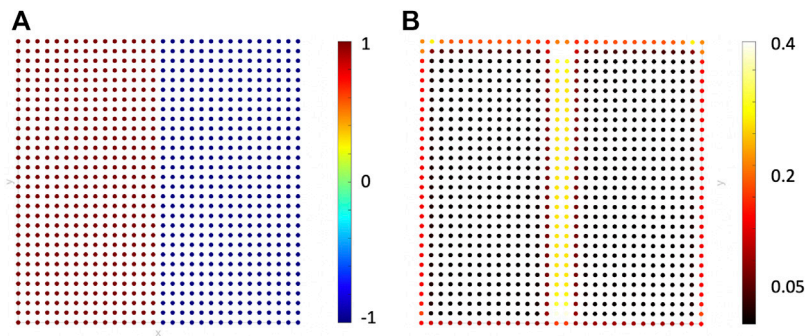


FIGURE 7 | Results of the application of MAHF ($k = 1$) on a two-valued signal (in panel **(A)**). We compute the MAHF, and we report the square of the module of the output in panel **(B)**.

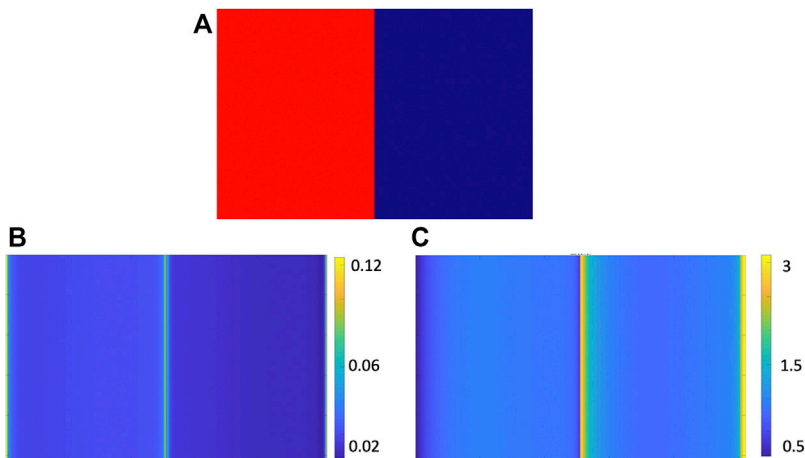


FIGURE 8 | Results of the application of CHF ($k = 1$) on a discrete bidimensional step function in panel **(A)**. We compute the output of the MAHF, and we report the square of the module in panel **(B)** and the phase in panel **(C)**.

$$K_i = \mathcal{K}(\|\mathbf{v}(\mathbf{q}_i)\|^2),$$

where \mathcal{K} is an integer function defined on \mathbb{R} . This is exemplified in **Figure 4** (left), which illustrates a point cloud (namely a low-resolution version of the Stanford bunny in Turk and Levoy (1994)) and two neighborhoods of different sizes K_i and K_j around points characterized by different values of the mean square value of the MAHF output, namely $\|\mathbf{v}(\mathbf{q}_i)\|^2$, plotted in **Figure 4** (right).

After the size of the estimation window at each point is given, the denoising algorithm iteratively alternates 1) the computation of a candidate estimate of the point location based on spatially adaptive averaging over the K_i -size neighborhood of the i -th vertex and 2) the update of the current estimate, along the direction of the normal to the surface.

In formulas, at the l -th iteration, the candidate estimate of the i -th point cloud vertex is computed as

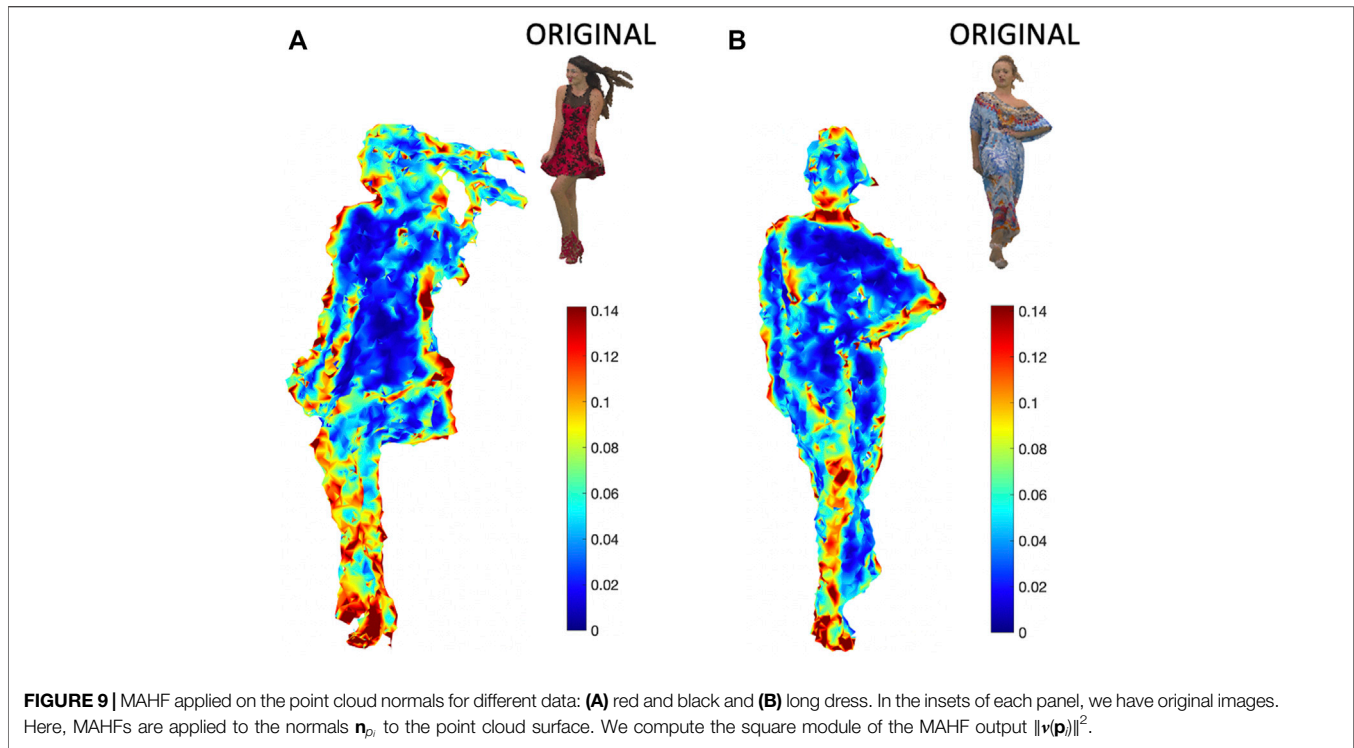
$$\tilde{\mathbf{p}}_i^{(l)} = \alpha_0 \hat{\mathbf{p}}_i^{(l-1)} + \sum_{j \in \eta(i; K_i)} \alpha_j \hat{\mathbf{p}}_j^{(l-1)},$$

where $\eta(i; K_i)$ denotes the set of K_i nearest neighbors of the i -th point cloud vertex. Then, the estimate is updated as

$$\hat{\mathbf{p}}_i^{(l)} = \hat{\mathbf{p}}_i^{(l-1)} + \rho_0^l \mathbf{n}^{(l-1)}(\hat{\mathbf{p}}_i^{(l-1)}) \mathbf{n}^{(l-1)}(\hat{\mathbf{p}}_i^{(l-1)})^T (\tilde{\mathbf{p}}_i^{(l)} - \hat{\mathbf{p}}_i^{(l-1)}),$$

where $\rho_0 \in (0, 1]$ is a parameter controlling the update rate throughout the iterations. Finally, the normals $\mathbf{n}^{(l)}(\hat{\mathbf{p}}_i^{(l)})$ are recomputed on the estimated point cloud $\hat{\mathbf{p}}_i^{(l)}$, $i = 0, \dots, N - 1$.

To sum up, the MAHF is applied once for all at the beginning of the iterations. For each vertex, the set of neighboring points is identified. Then, a candidate new point is computed as a weighted average of the neighbor and of the point itself. Then, the point estimated at the previous iteration is updated only by projection of the correction on the direction of the normal to the surface, as illustrated in **Figure 5**.



The normals to the surface are recomputed. The algorithm is terminated after few iterations (1-3 in the presented simulation results).

The algorithm is summarized as follows:

Input: Noisy point clouds coordinates \mathbf{q}_i , $i = 0, \dots, N - 1$; heat kernel spread t ; update rate control parameter ρ_0 .

Output: Denoised point clouds coordinates $\hat{\mathbf{p}}_i$, $i = 0, \dots, N - 1$.

Initialization:

Computation of the heat kernel. $K_t^{(g)}(\mathbf{p}_i, \mathbf{p}_j)$, $j = 0, \dots, N - 1$, $i = 0, \dots, N - 1$

Computation of the normals $\mathbf{n}^{(0)}$ and of their filtered version. $\mathbf{v}(\mathbf{p}_i) = \sum_{j=0}^{N-1} \varphi^{(1)}(\mathbf{q}_i, \mathbf{q}_j) \mathbf{n}^{(0)}(\mathbf{q}_j)$

Computation of $K_i = \mathcal{K}(\|\mathbf{v}(\mathbf{q}_i)\|^2)$ and of the K_i nearest neighborhood $\eta(i; K_i)$

Iteration: for $l = 1, \dots, L - 1$

Computation of the candidate estimate $\tilde{\mathbf{p}}_i^{(l)} = \alpha_0 \hat{\mathbf{p}}_i^{(l-1)} + \sum_{j \in \eta(i; K_i)} \alpha_j \hat{\mathbf{p}}_j^{(l-1)}$, with $\alpha_j = (\alpha_0)/K_i$

Update of the current estimate. $\hat{\mathbf{p}}_i^{(l)} = \hat{\mathbf{p}}_i^{(l-1)} + \rho^l [\mathbf{n}^{(l-1)}(\hat{\mathbf{p}}_i^{(l-1)})][\mathbf{n}^{(l-1)}(\hat{\mathbf{p}}_i^{(l-1)})]^T (\tilde{\mathbf{p}}_i^{(l)} - \hat{\mathbf{p}}_i^{(l-1)})$

Update of the normals $\mathbf{n}^{(l)}(\hat{\mathbf{p}}_i^{(l)})$.

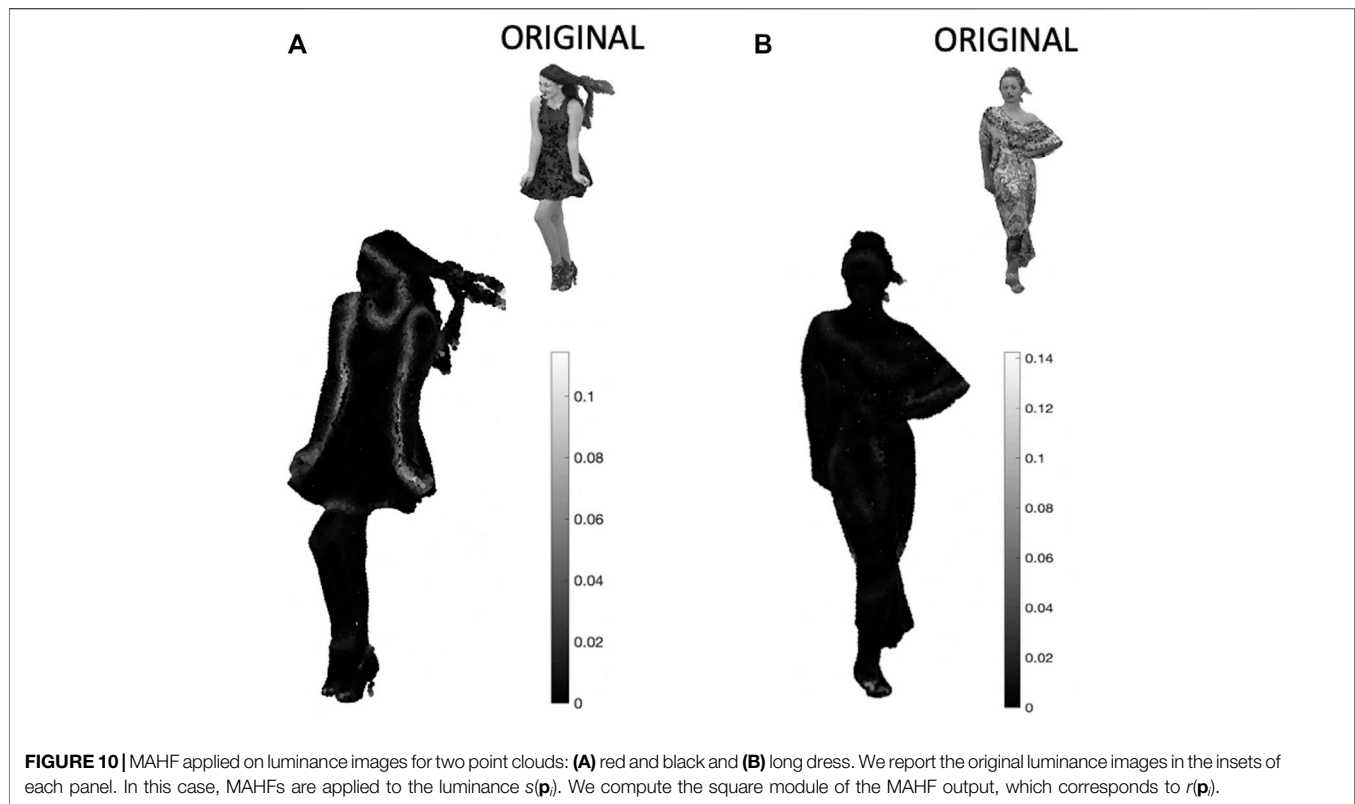
An overview of the VIPDA algorithm stages appears in **Figure 6**.

5.1 Remarks

As far as the computational complexity of VIPDA is concerned, a few remarks are in order. First, VIPDA implies an initial MAHF-filtering stage that implies the eigendecomposition. For the computation in 5, the use of all the elements of the eigendecomposition has a really high computational cost for

large N . To solve this limitation, Chebychev polynomial approximation by Huang et al. (2020); Hammond et al. (2011) can be applied in order to rewrite **Eq. 5** as a polynomial in \mathbf{L} . For the sake of concreteness, we have evaluated the time associated with each stage of the algorithm, implemented in Matlab® over a processor using this approximation for the Bunny cloud with $N = 8,146$. The net time for the computation of the heat kernel $K_t^{(g)}(\mathbf{p}_i, \mathbf{p}_j)$, $j = 0, \dots, N - 1$, $i = 0, \dots, N - 1$ sums up to $T_{K_t} = 40.7[s]$, the computation of the filtered normals $\mathbf{v}(\mathbf{p}_i)$ requires up to $T_v = 9.6[s]$, and the computation of the $L = 3$ or 4 iterations requires $T_L = 1.8[s]$. Overall, the iterative denoising algorithms require $T_{VIPDA} = 60.74$, which is comparable with state-of-the-art methods (e.g., the execution time on the same machine for the method in Dinesh et al., 2020 is about 70[s]).

Second, VIPDA is iterative, and it is not suited for parallelization. This observation stimulated the definition of an alternative version of the algorithm, namely VIPDAfast, boiling down to a single iteration and suitable for parallelization. This is achieved by a simplified application of the VIPDA key concept, that is, the adaptation of the size of the estimation neighborhood to the MAHF-filtered output. In the single iteration algorithm VIPDAfast, each point is straightforwardly estimated on a patch whose size is as a given function of the MAHF output at that point. Since all the estimates are obtained directly from the noisy sample, the algorithm may be parallelized on different subsets of points. The numerical simulation results will show that VIPDAfast, suited for parallelization, approximates the performance of the complete iterative VIPDA, especially on relatively smooth point clouds. VIPDAfast is expected to reduce the execution time in a way proportional to the number of available concurrent threads.



Finally, a remark on the noise model is in order. Indeed, the VIPDA at each iteration performs a local averaging, which tackles Gaussian noise and, in a suboptimal way, also impulsive noise. Still, the core of VIPDA allows 1) to adaptively select the neighborhood of the vertex to be used in the estimate and 2) to apply the correction to the noise component orthogonal to the mesh surface. These two principles can also be applied when the actual estimate is realized by different nonlinear operators tuned to the actual noise statistics by Ambike et al. (1994). Therefore, VIPDA can be extended to deal with different kinds of noise by replacing the average operator with a suitable nonlinear one, this is left for further study.

6 SIMULATION RESULTS

In this section, we present simulation results associated with the application of MAHF on synthetic and real PC, and we measure the performances of VIPDA. In particular, in **subsec.6.1**, we illustrate the MAHF behavior, also in comparison with CHF, and in **subsec.6.2** we assess the performance of VIPDA, also in comparison with state-of-the-art denoising algorithms.

6.1 MAHF-Based Point Cloud Filtering

In this subsection, we present some examples on the application of MAHF on different point clouds. In this article, we introduce a point cloud filtering method inspired by HVS, and we show its potential to

both geometric and texture PC data. The proposed class of filters presents a local anisotropic behavior that highlights directional components of the point cloud texture or geometry. The filter output can be leveraged as input to various adaptive processing tasks.

First, we consider the case of a point cloud obtained by equispaced sampling of a planar surface, over which a discontinuous signal is defined. This case is illustrated in **Figure 7A**), in which we see the point cloud in which a two-valued signal is defined; the signal is characterized by a discontinuity in the middle. Then, MAHF filtering (with $k = 1$) is applied. In **Figure 7B**), we report the square of the module of the related MAHF output. As expected, the MAHF highlights the vertices in correspondence with the signal discontinuity. In order to compare MAHF with CHF behavior, we take into account an analogous scenario for CHF filtering, namely we consider an image representing a discrete bidimensional step function, which is represented in **Figure 8A**. We apply the $k = 1$ CHF to the image, and we separately plot its module and phase in the panel **Figures 8B, C**, respectively. The output of the $k = 1$ MAHF and CHF filters highlights the areas in correspondence with the discontinuity of the signal. Thereby, we recognize that the $k = 1$ MAHF filter straightforwardly extends to the planar point cloud domain, the behavior observed applying the $k = 1$ CHF filter on image data. Indeed, we remark that the MAHF and CHF filters definitions in the point cloud domain and image domain, respectively, are analogous, and it is expected that the MAHF can retrieve structured discontinuities of the signals defined on a point cloud.

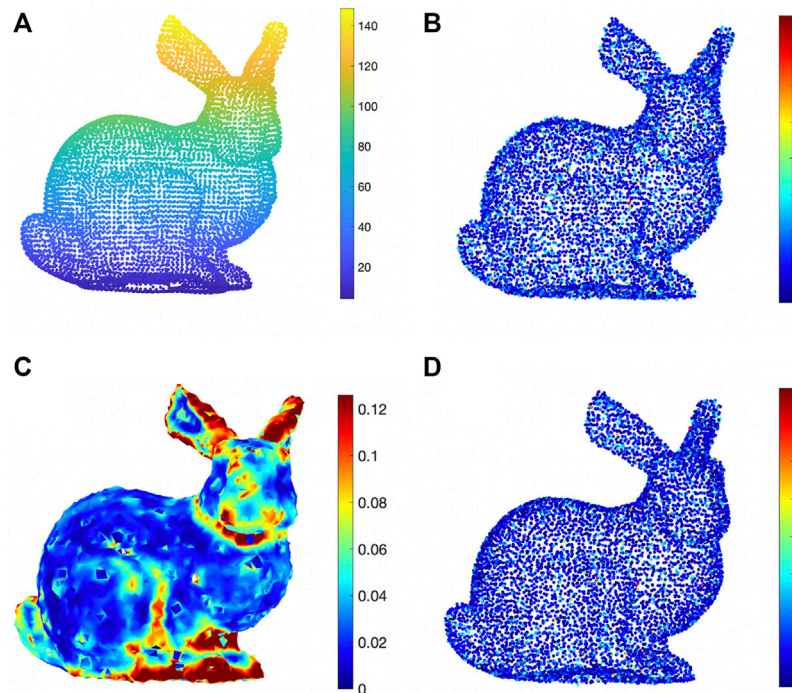


FIGURE 11 | Results of VIPDA at each step. We consider a point cloud related to the Stanford bunny (Turk and Levoy 1994) which is represented with its original coordinates \mathbf{p}_i in panel (A). Then, a Gaussian noise \mathbf{w}_i is added with SNR = 38, and the noisy points \mathbf{q}_i on the point cloud are represented in panel (B). The colors associated to the color bar correspond to the difference between noisy and original coordinates $\|\mathbf{p}_i - \mathbf{q}_i\|^2$. In panel (C), we report the output of the MAHF $\|\mathbf{v}(\mathbf{q}_i)\|^2$ applied to the estimated normals \mathbf{n}_{q_i} . In panel (D), we have results of VIPDA. We have $\hat{\mathbf{p}}_i$ points belonging to the denoised PC, and the colors relate to the difference between reconstructed and original coordinates $\|\hat{\mathbf{p}}_i - \mathbf{q}_i\|^2$.

TABLE 2 | Table of SNR with Stanford bunny point cloud.

SNR _{noisy}	SNR _(Dinesh et al. 2020)	SNR _{average}	SNR _{iter}	SNR _{iterPARA}
35	35.07	35.42	35.89	35.94
38	38.28	38.28	38.92	38.96
40	40.69	40.15	40.97	40.98
48	46.69	46.49	48.62	48.62

Highest values for each SNR are highlighted in bold.

After exemplifying the relation between MAHF and CHF, we apply MAHF to open access point clouds. For this study, we consider two point clouds belonging to the 8iVSLF dataset (d'Eon et al., 2017). Specifically, we consider the point clouds red and black and long dress (d'Eon et al., 2017), which are the 3D point clouds illustrated in the insets of panels A and B in Figure 9. Each point cloud vertex is associated to the red, green, and blue components of the surface color as seen using a multi-camera rig. The original point clouds red and black and long dress have been resampled to a number of points equal to $N = 9,622$ and $N = 9,378$, respectively.

First of all, we apply MAHF to the normals \mathbf{n}_{p_i} , $i = 0, \dots, N - 1$ to the point cloud surface², and we compute the square module of the

MAHF output, namely the estimated $\|\mathbf{v}(\mathbf{p}_i)\|^2$, $i = 0, \dots, N - 1$ in Figure 9. We graphically represent it in gray-scale pseudo-colors in Figure 9. The largest values of $\|\mathbf{v}(\mathbf{p}_i)\|^2$, $i = 0, \dots, N - 1$ are associated to the vertices characterized by curvature changes in each point cloud.

Then, for each point cloud, we analyze the filtering of the color information. At each vertex p_i , $i = 0, \dots, N - 1$, we compute the luminance given by the available RGB values, and we apply MAHF by considering the luminance as the real-valued input signal $s(\mathbf{p}_i)$, $i = 0, \dots, N - 1$ over the point cloud graph. The MAHF output $r(\mathbf{p}_i) = \sum_{j=0}^{N-1} \varphi^{(k)}(\mathbf{p}_i, \mathbf{p}_j) s(\mathbf{p}_j)$, $i = 0, \dots, N - 1$ is then computed. We present the square module $r(\mathbf{p}_i)$ of the MAHF filter output in Figure 10 for the two point clouds under study. In this case, this method is able to highlight luminance variations on the graph, and it spots out details in the images such as the arms in panel A or the feet in panel B of Figure 10.

6.2 VIPDA Performances

After illustrating the application of MAHF to point cloud filtering on shape and texture information by means of examples on real data, we address the assessment of VIPDA in this subsection.

To this aim, we first illustrate the application of VIPDA over synthetic data. Specifically, we consider the point cloud related to the Stanford bunny (Turk and Levoy 1994), resampled at $N = 8,146$. The noisy coordinates \mathbf{q}_i of the points on the PC are obtained as in Eq. 9. In the simulations, a Gaussian noise \mathbf{w}_i is added to the original coordinates \mathbf{p}_i . The intensity of the additive

²The point cloud normals are computed using the ©Matlab by the implementation of the method in Hoppe et al. (1992).

TABLE 3 | Table of SNR with fixed SNR of noisy data at 40 dB with different point clouds, i.e., Stanford bunny, long dress and red and black, and spherical meshes.

Point cloud	SNR _(Dinesh et al. 2020)	SNR _{average}	SNR _{VIPDA}	SNR _{VIPDAfast}
Stanford bunny	40.69	40.15	40.97	40.98
Red and black	40.03	40.49	40.48	40.50
Long dress	39.99	40.49	40.63	40.63
Sphere	39.97	39.49	40.46	40.46

Highest values for each SNR are highlighted in bold.

TABLE 4 | Table of SNR with fixed SNR of noisy data at 48 dB with several point clouds, i.e., Stanford bunny, long dress and red and black, and spherical meshes.

Point cloud	SNR _(Dinesh et al. 2020)	SNR _{average}	SNR _{VIPDA}	SNR _{VIPDAfast}
Stanford Bunny	46.69	46.49	48.62	48.62
Red and black	48.05	47.72	48.26	48.27
Long dress	48.03	47.70	48.58	48.57
Sphere	48.09	38.25	48.26	48.26

Highest values for each SNR are highlighted in bold.

noise is measured by the signal-to-noise ratio (SNR), which is a parameter varied in the following analyses, and it is computed as:

$$SNR_{noisy} = 10 \log_{10} \frac{\frac{1}{N} \sum_{i=0}^{N-1} \|\mathbf{p}_i\|^2}{\frac{1}{N} \sum_{i=0}^{N-1} \|\mathbf{p}_i - \mathbf{q}_i\|^2}. \quad (11)$$

We plot the original point cloud in **Figure 11A**, in which the pseudo-color is associated to the third coordinate of \mathbf{q}_i , $i = 0, \dots, N-1$ (coordinate w.r.t. the z -axis). In **Figure 11B**, we plot the noisy point cloud obtained for $SNR_{noisy} = 38\text{dB}$. The pseudo-colors of the point cloud represent the square module of the difference between the noisy coordinates $\|\mathbf{p}_i - \mathbf{q}_i\|^2$, $i = 0, \dots, N-1$ and the original ones computed as $\|\mathbf{p}_i - \mathbf{q}_i\|^2$. In this manner, the point color (represented in a color scale from blue for zero values and red for the maximum values) reflects how much each point is corrupted by the additive Gaussian noise.

Then, we apply the MAHF to the estimated normals $\mathbf{n}(\mathbf{q}_i)$, $i = 0, \dots, N-1$ of the noisy point cloud, and we compute the square value of the MAHF output at each vertex, namely $\|\mathbf{v}(\mathbf{q}_i)\|^2$, $i = 0, \dots, N-1$. The so-obtained values are illustrated in **Figure 11C**, as pseudo-colors at the vertices. We recognize that the largest values are observed in correspondence to vertices in areas of normal changes. These results exemplify that the MAHFs are able to capture the variability of the signals.

Finally, we apply VIPDA to the filtered point cloud, and we show the denoised point cloud in **Figure 11D**. Here, we have the $\hat{\mathbf{p}}_i$ points reconstructed by VIPDA, and we define the signal associated to each new point as the error computed between the original coordinates \mathbf{p} and the reconstructed ones as $\|\mathbf{p}_i - \hat{\mathbf{p}}_i\|^2$. From a visual analysis, we recognize that the points in D are closer to original ones in A with respect to absence of denoising in B; this effect is more visible at the boundary of the point cloud. A more quantitative result is provided by the following SNR computation:

Specifically, we design a signal-dependent feature graph Laplacian regularizer (SDFGLR) that assumes surface normals computed from point coordinates are piecewise smooth with respect to a signal-dependent graph Laplacian matrix.

Finally, we perform analyses based on the SNR and compare our method with different alternatives. In this direction, we first

consider the algorithm proposed in Dinesh et al., (2020), in which authors perform a graph Laplacian regularization that starts from the hypothesis that the normals to the surface at the point cloud vertices are smooth w.r.t graph Laplacian. For sake of comparison, we also study the average case, in which we analyze the PC results from the local average of its spatial coordinates. Finally, we consider the VIPDA and the VIPDAfast algorithms. In the simulations, $\mathcal{K}(\|\mathbf{v}(\mathbf{q}_i)\|^2)$ is set equal to a bi-level function, depending on whether $\|\mathbf{v}(\mathbf{q}_i)\|^2$ is above the threshold θ or not. We set $t = 10$ and $\alpha_0 = 0.9$ on all the data and $\theta = 0.06$, $\mathcal{K} \in \{3, 9\}$ on synthetic data and $\theta = 0.08$, $\mathcal{K} \in \{3, 15\}$ on real data.

In order to perform the computations, we first consider the Stanford bunny point cloud, and we select different levels of SNR, namely $SNR_{noisy} = 35, 38, 40, 48\text{dB}$. Then, we take into account different denoising algorithms and report the SNR achieved on the denoised point cloud in **Table 2**. For each method, we compute the SNR as the distance between the reconstructed coordinates and the original ones as

$$SNR = 10 \log_{10} \frac{\frac{1}{N} \sum_{i=1}^{N-1} \|\mathbf{p}_i\|^2}{\frac{1}{N} \sum_{i=0}^{N-1} \|\mathbf{p}_i - \hat{\mathbf{p}}_i\|^2}.$$

Our results show that our denoising method outperforms the alternatives. The method with the parallelization even increases the performances w.r.t. the iterative one. This is due to the particular nature of the point cloud, in which the flat areas and the high curvature areas are relatively easy to distinguish, and the method coarsely operating on the two point sets achieves the best results. This is more clearly highlighted on point clouds acquired on real objects as illustrated in the following: We consider the two other point clouds (Red and Black and Long Dress) for two fixed levels of noise with SNR at 40dB and 48dB. The results are, respectively, reported in **Tables 3, 4**. SNR values show that the proposed VIPDA, either in its original or fast version, performs better than state-of-the-art competitors in denoising point cloud signals corrupted by Gaussian noise. Finally, we consider a smooth point cloud, namely a sphere with $N = 900$ points. Also on this smooth point cloud, where the MAHF gives a uniform output and the patch size is fixed, VIPDA achieves an SNR

improvement. The correction by VIPDA is restrained to the normal direction and leads to a smooth surface. Thereby, VIPDA achieves a SNR improvement also on smooth surfaces, and an improvement due to the reduction of the normal noise component is observed even in the limit case of a planar mesh.

It is important to mention that the performance of all the methods, including the proposed method, degrades severely if the SNR decreases. This is due to the fact that, in correspondence with low SNR, the positions of the vertices are displaced such that the positions of the noisy points may even exchange with respect to the original ones, and this phenomenon is not recovered even though the MAHF on the signal is recomputed at each iteration. A possible solution to this would be to initially denoise the Laplacian associated to the point cloud graph by leveraging a spectral prior, as in Cattai et al. (2021), or by jointly exploiting the shape and texture information; this latter point is left for future studies.

To sum up, these findings demonstrate the potential of the proposed VIPDA approach for point cloud denoising and pave the way for designing new processing tools for signals defined over non-Euclidean domains.

7 CONCLUSION

This work has presented a novel point cloud denoising approach, the visually driven point cloud denoising algorithm (VIPDA). The proposed method differs from other competitors in linking the denoising with visually relevant features, as estimated by suitable anisotropic angular filters in the vertex domain.

REFERENCES

- Ambike, S., Ilow, J., and Hatzinakos, D. (1994). Detection for Binary Transmission in a Mixture of Gaussian Noise and Impulsive Noise Modeled as an Alpha-Stable Process. *IEEE Signal. Process. Lett.* 1, 55–57. doi:10.1109/97.295323
- Beghdadi, A., Larabi, M.-C., Bouzerdoum, A., and Iftekharuddin, K. M. (2013). A Survey of Perceptual Image Processing Methods. *Signal. Processing: Image Commun.* 28, 811–831. doi:10.1016/j.image.2013.06.003
- Belkin, M., Sun, J., and Wang, Y. (2009). “Constructing Laplace Operator from point Clouds in R^d ,” in Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, New York, January 4–6, 2009 (SIAM), 1031–1040.
- Campisi, P., and Scarano, G. (2002). A Multiresolution Approach for Texture Synthesis Using the Circular Harmonic Functions. *IEEE Trans. Image Process.* 11, 37–51. doi:10.1109/83.977881
- Cattai, T., Scarano, G., Corsi, M.-C., Bassett, D., De Vico Fallani, F., and Colonnese, S. (2021). Improving J-Divergence of Brain Connectivity States by Graph Laplacian Denoising. *IEEE Trans. Signal. Inf. Process. Over Networks* 7, 493–508. doi:10.1109/tsipn.2021.3100302
- Chen, H., Wei, M., Sun, Y., Xie, X., and Wang, J. (2019). Multi-patch Collaborative point Cloud Denoising via Low-Rank Recovery with Graph Constraint. *IEEE Trans. Vis. Comput. Graph* 26, 3255–3270. doi:10.1109/TVCG.2019.2920817
- Chen, H., Xu, F., Liu, W., Sun, D., Liu, P. X., Menhas, M. I., et al. (2021). 3d Reconstruction of Unstructured Objects Using Information from Multiple Sensors. *IEEE Sensors J.* 21–26963. doi:10.1109/jsen.2021.3121343.26951
- Colonnese, S., Campisi, P., Panci, G., and Scarano, G. (2004). Blind Image Deblurring Driven by Nonlinear Processing in the Edge Domain. *EURASIP J. Adv. Signal Process.* 2004, 1–14. doi:10.1155/s1110865704040132
- Colonnese, S., Randi, R., Rinauro, S., and Scarano, G. (2010). “Fast Image Interpolation Using Circular Harmonic Functions,” in 2010 2nd European Workshop on Visual Information Processing (EUVIP), Paris, France, 5–7 July 2010 (IEEE), 114–118. doi:10.1109/euvip.2010.5699119
- Colonnese, S., Rinauro, S., and Scarano, G. (2013). Bayesian Image Interpolation Using Markov Random fields Driven by Visually Relevant Image Features. *Signal. Processing: Image Commun.* 28, 967–983. doi:10.1016/j.image.2012.07.001
- Conti, F., Scarano, G., and Colonnese, S. (2021). “Multiscale Anisotropic Harmonic Filters on Non Euclidean Domains,” in 2021 29th European Signal Processing Conference (EUSIPCO) (Virtual Conference), (IEEE), 701–705.
- d'Eon, E., Harrison, B., Myers, T., and Chou, P. A. (2017). 8i Voxelized Full Bodies-A Voxelized point Cloud Dataset. *ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document WG11M40059/WG1M74006* 7, 8
- Dai, Y., Wen, C., Wu, H., Guo, Y., Chen, L., and Wang, C. (2021). “Indoor 3d Human Trajectory Reconstruction Using Surveillance Camera Videos and point Clouds,” in *IEEE Transactions on Circuits and Systems for Video Technology*. doi:10.1109/tcsvt.2021.3081591
- de Hoog, J., Ahmed, A. N., Anwar, A., Latré, S., and Hellinckx, P. (2021). “Quality-aware Compression of point Clouds with Google Draco,” in *Advances on P2P, Parallel, Grid, Cloud and Internet Computing. 3PGCIC 2021. Lecture Notes in Networks and Systems*. Editor L. Barolli (Cham: Springer) 343, 227–236. doi:10.1007/978-3-030-89899-1_23
- Dinesh, C., Cheung, G., and Bajic, I. V. (2020). Point Cloud Denoising via Feature Graph Laplacian Regularization. *IEEE Trans. Image Process.* 29, 4143–4158. doi:10.1109/tip.2020.2969052
- Diniz, R., Farias, M. Q., and Garcia-Freitas, P. (2021). “Color and Geometry Texture Descriptors for point-cloud Quality Assessment,” in *IEEE Signal Processing Letters*. doi:10.1109/lsp.2021.3088059

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found at: <https://mpeg-pcc.org/index.php/pcc-content-database/>.

AUTHOR CONTRIBUTIONS

TC and SC conceived and designed the analysis; AD collected the data, performed the analysis, and contributed to the algorithm definition; GS contributed to the analysis tools; and TC and SC wrote the manuscript.

- Erçelik, E., Yurtsever, E., and Knoll, A. (2021). 3d Object Detection with Multi-Frame Rgb-Lidar Feature Alignment. *IEEE Access* 9, 143138–143149. doi:10.1109/ACCESS.2021.3120261
- Friedlander, B. (1984). On the Cramer- Rao Bound for Time Delay and Doppler Estimation (Corresp.). *IEEE Trans. Inform. Theor.* 30, 575–580. doi:10.1109/tit.1984.1056901
- Hammond, D. K., Vandergheynst, P., and Gribonval, R. (2011). Wavelets on Graphs via Spectral Graph Theory. *Appl. Comput. Harmonic Anal.* 30, 129–150. doi:10.1016/j.acha.2010.04.005
- Hoppe, H., DeRose, T., Duchamp, T., McDonald, J., and Stuetzle, W. (1992). Surface Reconstruction from Unorganized Points. *ACM SIGGRAPH Comput. Graph.* 26, 71–78. doi:10.1145/142920.134011
- Hou, T., and Qin, H. (2012). Continuous and Discrete Mexican Hat Wavelet Transforms on Manifolds. *Graphical Models* 74, 221–232. doi:10.1016/j.gmod.2012.04.010
- Hu, W., Hu, Q., Wang, Z., and Gao, X. (2021a). Dynamic point Cloud Denoising via Manifold-To-Manifold Distance. *IEEE Trans. Image Process.* 30, 6168–6183. doi:10.1109/tip.2021.3092826
- Hu, W., Pang, J., Liu, X., Tian, D., Lin, C.-W., and Vetro, A. (2021b). “Graph Signal Processing for Geometric Data and beyond: Theory and Applications,” in *IEEE Transactions on Multimedia*. doi:10.1109/tmm.2021.3111440
- Huang, J., Choudhury, P. K., Yin, S., and Zhu, L. (2021). Real-time Road Curb and Lane Detection for Autonomous Driving Using Lidar point Clouds. *IEEE Access* 9, 144940–144951. doi:10.1109/access.2021.3120741
- Huang, S.-G., Lyu, I., Qiu, A., and Chung, M. K. (2020). Fast Polynomial Approximation of Heat Kernel Convolution on Manifolds and its Application to Brain Sulcal and Gyral Graph Pattern Analysis. *IEEE Trans. Med. Imaging* 39, 2201–2212. doi:10.1109/tmi.2020.2967451
- Huang, T., Li, R., Li, Y., Zhang, X., and Liao, H. (2021). Augmented Reality-Based Autostereoscopic Surgical Visualization System for Telesurgery. *Int. J. Comput. Assist. Radiol. Surg.* 16, 1985–1997. doi:10.1007/s11548-021-02463-5
- Irfan, M. A., and Magli, E. (2021a). Exploiting Color for Graph-Based 3d point Cloud Denoising. *J. Vis. Commun. Image Representation* 75, 103027. doi:10.1016/j.jvcir.2021.103027
- Irfan, M. A., and Magli, E. (2021b). Joint Geometry and Color point Cloud Denoising Based on Graph Wavelets. *IEEE Access* 9, 21149–21166. doi:10.1109/access.2021.3054171
- Jacovitti, G., and Neri, A. (2000). Multiresolution Circular Harmonic Decomposition. *IEEE Trans. Signal. Process.* 48, 3242–3247. doi:10.1109/78.875481
- Ji, T., Guo, Y., Wang, Q., Wang, X., and Li, P. (2021). Economy: Point Clouds-Based Energy-Efficient Autonomous Navigation for Uavs. *IEEE Trans. Netw. Sci. Eng.* 8 (4), 2885–2896. doi:10.1109/tNSE.2021.3049263
- Neri, A., and Jacovitti, G. (2004). Maximum Likelihood Localization of 2-d Patterns in the Gauss-Laguerre Transform Domain: Theoretic Framework and Preliminary Results. *IEEE Trans. Image Process.* 13, 72–86. doi:10.1109/tip.2003.818021
- Panci, G., Campisi, P., Colonnese, S., and Scarano, G. (2003). Multichannel Blind Image Deconvolution Using the Bussgang Algorithm: Spatial and Multiresolution Approaches. *IEEE Trans. Image Process.* 12, 1324–1337. doi:10.1109/tip.2003.818022
- Ramalho, E., Peixoto, E., and Medeiros, E. (2021). Silhouette 4d with Context Selection: Lossless Geometry Compression of Dynamic point Clouds. *IEEE Signal. Process. Lett.* 28, 1660–1664. doi:10.1109/lsp.2021.3102525
- Rist, C., Emmerichs, D., Enzweiler, M., and Gavrilu, D. (2021). “Semantic Scene Completion Using Local Deep Implicit Functions on Lidar Data,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi:10.1109/tpami.2021.3095302
- Sun, X., Wang, S., Wang, M., Cheng, S. S., and Liu, M. (2020). “An Advanced Lidar point Cloud Sequence Coding Scheme for Autonomous Driving,” in Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, October 12–16, 2020, 2793–2801. doi:10.1145/3394171.3413537
- Turk, G., and Levoy, M. (1994). “Zippered Polygon Meshes from Range Images,” in Proceedings of the 21st annual conference on Computer graphics and interactive techniques, Orlando, Florida, July 24–29, 1994. Computer Graphics Proceedings, Annual Conference Series (ACM SIGGRAPH), 311–318. doi:10.1145/192161.192241
- Wu, J., Li, L., Dong, W., Shi, G., Lin, W., and Kuo, C.-C. J. (2017). Enhanced Just Noticeable Difference Model for Images with Pattern Complexity. *IEEE Trans. Image Process.* 26, 2682–2693. doi:10.1109/tip.2017.2685682
- Xiong, J., Gao, H., Wang, M., Li, H., and Lin, W. (2021). “Occupancy Map Guided Fast Video-Based Dynamic point Cloud Coding,” in *IEEE Transactions on Circuits and Systems for Video Technology*.
- Yang, Q., Ma, Z., Xu, Y., Li, Z., and Sun, J. (2020). “Inferring point Cloud Quality via Graph Similarity,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Yu, K., Gorbachev, G., Eck, U., Pankratz, F., Navab, N., and Roth, D. (2021). Avatars for Teleconsultation: Effects of Avatar Embodiment Techniques on User Perception in 3d Asymmetric Telepresence. *IEEE Trans. Vis. Comput. Graphics* 27, 4129–4139. doi:10.1109/tvcg.2021.3106480
- Zhu, D., Chen, H., Wang, W., Xie, H., Cheng, G., Wei, M., et al. (2022). “Non-local Low-Rank point Cloud Denoising for 3d Measurement Surfaces,” in *IEEE Transactions on Instrumentation and Measurement*. doi:10.1109/tim.2021.3139686

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Cattai, Delfino, Scarano and Colonnese. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Multivariate Lipschitz Analysis of the Stability of Neural Networks

Kavya Gupta^{1,2*}, Fateh Kaakai², Beatrice Pesquet-Popescu², Jean-Christophe Pesquet¹ and Fragkiskos D. Malliaros¹

¹Inria, Centre de Vision Numérique, CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France, ²Air Mobility Solutions BL, Thales LAS, Rungis, France

The stability of neural networks with respect to adversarial perturbations has been extensively studied. One of the main strategies consist of quantifying the Lipschitz regularity of neural networks. In this paper, we introduce a multivariate Lipschitz constant-based stability analysis of fully connected neural networks allowing us to capture the influence of each input or group of inputs on the neural network stability. Our approach relies on a suitable re-normalization of the input space, with the objective to perform a more precise analysis than the one provided by a global Lipschitz constant. We investigate the mathematical properties of the proposed multivariate Lipschitz analysis and show its usefulness in better understanding the sensitivity of the neural network with regard to groups of inputs. We display the results of this analysis by a new representation designed for machine learning practitioners and safety engineers termed as a Lipschitz star. The Lipschitz star is a graphical and practical tool to analyze the sensitivity of a neural network model during its development, with regard to different combinations of inputs. By leveraging this tool, we show that it is possible to build robust-by-design models using spectral normalization techniques for controlling the stability of a neural network, given a safety Lipschitz target. Thanks to our multivariate Lipschitz analysis, we can also measure the efficiency of adversarial training in inference tasks. We perform experiments on various open access tabular datasets, and also on a real Thales Air Mobility industrial application subject to certification requirements.

Keywords: lipschitz, neural networks, stability, adversarial attack, sensitivity, safety, tabular data

OPEN ACCESS

Edited by:

Hagit Messer,
Tel Aviv University, Israel

Reviewed by:

Zhenyu Liao,
Huazhong University of Science and
Technology, China
Jonatan Ostrometzky,
Tel Aviv University, Israel

*Correspondence:

Kavya Gupta
kavya.gupta100@gmail.com

Specialty section:

This article was submitted to
Signal Processing Theory,
a section of the journal
Frontiers in Signal Processing

Received: 13 October 2021

Accepted: 31 January 2022

Published: 05 April 2022

Citation:

Gupta K, Kaakai F,
Pesquet-Popescu B, Pesquet J-C and
Malliaros FD (2022) Multivariate
Lipschitz Analysis of the Stability of
Neural Networks.
Front. Sig. Proc. 2:794469.
doi: 10.3389/frsip.2022.794469

1 INTRODUCTION

Artificial neural networks are at the core of recent advances in Artificial Intelligence. One of the main challenges faced today, especially by companies designing advanced industrial systems, is to ensure the safety of new generations of products using these technologies. Neural networks have been shown to be sensitive to adversarial perturbations (Szegedy et al., 2013). For example, changing a few pixels of an image may lead to misclassification of the image by a Deep Neural Network (DNN), which emphasizes the potential lack of stability of such architectures. DNNs being sensitive to adversarial examples, can thus be fooled, in an intentional manner (security issue) or in undeliberate/accidental manner (safety issue), which raises a major stability concern for safety-critical systems which need to be certified by an independent certification authority prior to any entry into production/operation. DNN-based solutions are hindered with such issue due to their complex nonlinear structure. Attempts towards verification of neural networks have been made for example in (Katz et al., 2017;

Weng et al., 2019). It has been proven in (Tsipras et al., 2018b) that there exists a trade-off between the prediction performance and the stability of neural networks.

In the last years, the number of works devoted to the stability issue of neural networks has grown in manifolds. In these works, the terms “stability”, “robustness” or “local robustness” are used interchangeably with the same meaning which is formally defined in this paper as the extent to which a neural network can continue to operate correctly despite small perturbations in its inputs. The stability criterion considered here highlights the fact that these small perturbations in the inputs do not produce high variations of the outputs. Many approaches have been proposed, some dedicated to specific architectures (e.g., networks using only ReLU activation functions) and grounded on more or less empirical techniques. We can break down broadly these techniques into three categories:

- Purely computational approaches which consist in attacking a neural network and observing its response to such attacks,
- methods based on (often clever) heuristics for testing/promoting the stability of a neural net,
- studies that aim at establishing mathematical proofs of stability.

These three kinds of strategies are useful for building and certifying effectively robust neural networks. However, the techniques based on mathematical proofs of stability are generally preferred by industrial safety experts since they enable a safe-by-design approach that is more efficient than a robustness verification activity done a posteriori with a necessarily bounded effort. Among the possible mathematical approaches, we focus in this article on those relying upon the analysis of the Lipschitz properties of neural networks. Such properties play a fundamental role in the understanding of the internal mechanisms governing these complex nonlinear systems. Besides, they make few assumptions on the type of non-linearities used and are thus valid for a wide range of networks. Nevertheless, they generate a number of challenges both from a theoretical and numerical standpoints.

Since DNNs are sensitive to small specific perturbations, providing a quantitative estimation of the stability of such architectures is of paramount importance for safe and secure product development in domains such as aeronautics, ground transportation, autonomous vehicles, energy, and healthcare. One metric to assess the stability of neural networks to adversarial perturbations is the Lipschitz constant, which upper bounds the ratio between output variations and input variations for a given metric. More generally, in deep learning theory, novel generalization bounds critically rely on the Lipschitz constant of the neural network (Bartlett et al., 2017). One of the main limitations of the Lipschitz constant, defined in either global or local context, is that it only provides a single parameter to quantify the robustness of a neural network. Such a single-parameter analysis does not facilitate the understanding of potential sources of instability. In particular, it may be insightful to identify the inputs which have the highest impact in terms of sensitivity. In the context of tabular data mining, the inputs often

have quite heterogeneous characteristics. Some of them are categorical data, often encoded in a specific way [e.g., one-hot encoder (Hancock and Khoshgoftaar, 2020)] and among them, one can usually distinguish those which are unsorted (like labels identifying countries) or those which are sorted (like severity scores in a disease). So, it may appear useful to analyze in a specific manner each type of inputs of a NN and even sometimes to exclude some of these inputs (e.g., unsorted categorical data for which the notion of small perturbation may be meaningless) from the performed sensitivity analysis.

The contributions of the work are summarized below:

- A multivariate analysis of the Lipschitz properties of NNs is performed by generating a set of partial Lipschitz constants. This opens a new dimension to studying the stability of NNs.
- Our sensitivity analysis allows us to capture the behaviour of an individual input or group of inputs.
- The results of this analysis are displayed by a new graphical representation termed as a Lipschitz star.
- Using the proposed analysis, we also study quantitatively the effect of spectral normalization constraint and adversarial training on the stability of NNs.
- We showcase our results on various open-source datasets along with a real industrial application in the domain of Air Traffic Management.

In the next section we give a detailed description of the state-of-the-art related to the quantification of the Lipschitz constant in neural networks. **Section 3** gives our proposed method pertaining to sensitivity of inputs and introduction to Lipschitz stars. **Section 4** provides an analytical evaluation for our approach with synthetic datasets. The next section gives detailed results on three open source datasets and a real safety critical industrial dataset. The last section concludes our paper.

2 OVERVIEW ON THE ESTIMATION OF THE LIPSCHITZ CONSTANT OF FEEDFORWARD NETWORKS

2.1 Theoretical Background

An m -layered feedforward network can be modelled by the following recursive equations:

$$(\forall i \in \{1, \dots, m\}) \quad x_i = T_i(x_{i-1}) = R_i(W_i x_{i-1} + b_i), \quad (1)$$

where, at the i th layer, $x_{i-1} \in \mathbb{R}^{N_{i-1}}$ designates the input vector, $x_i \in \mathbb{R}^{N_i}$ the output one, $W_i \in \mathbb{R}^{N_i \times N_{i-1}}$ is the weight matrix, $b_i \in \mathbb{R}^{N_i}$ is the bias vector, and $R_i: \mathbb{R}^{N_i} \rightarrow \mathbb{R}^{N_i}$ is the activation operator. This operator may consist of the application of basic nonlinear functions, e.g., ReLU or tanh, to each component of the input. Alternatively, it may consist of a softmax operation or group sorting operations which typically arise in max pooling. In this model, when the matrix W_i has a Toeplitz or a block-Toeplitz structure, a convolutive layer is obtained.

Since the seminal work in (Szegedy et al., 2013), it is known that instability in the outputs of the neural networks may arise. This issue, often referred to as the stability with respect to adversarial noise, tends to be more severe when the training set is small. However, it may even happen with large datasets such as ImageNet. As shown in (Goodfellow et al., 2015), the problem is mainly related to the choice of the weight matrices. One way of quantifying the stability of the system is to calculate a Lipschitz constant of the network.

A Lipschitz constant of a function T is an upper bound on the ratio between the variations of the output values and the variations of input arguments of a function T . Thus, it is a measure of sensitivity of the function with respect to input perturbations. This means that, if $\theta \in [0, +\infty]$ is such that, for every input $x \in \mathbb{R}^{N_0}$ and perturbation $z \in \mathbb{R}^{N_0}$,

$$\|T(x+z) - T(x)\| \leq \theta \|z\|, \quad (2)$$

then θ is a Lipschitz constant of T . Note that, the same notation is used here for the norms on \mathbb{R}^{N_0} and \mathbb{R}^{N_m} , but actually different norms can be used. If not specified, the standard Euclidean norm will be used. Another important remark which follows from the mean value inequality is that, if T is differentiable on \mathbb{R}^{N_0} , the optimal (i.e., smallest) Lipschitz constant is

$$\theta = \sup_{x \in \mathbb{R}^{N_0}} \|T'(x)\|_S = \sup_{x \in \mathbb{R}^{N_0}} \sup_{z \in \mathbb{R}^{N_0}} \frac{\|T'(x)z\|}{\|z\|}, \quad (3)$$

where $T'(x) \in \mathbb{R}^{N_m \times N_0}$ is the Jacobian matrix of T at x and $\|\cdot\|_S$ denotes the spectral matrix norm. Local definitions of the Lipschitz constant are also possible (Yang et al., 2020). In order to get more meaningful expressions of Lipschitz constants, an important assumption which will be made in this paper is that the operators $(R_i)_{1 \leq i \leq m}$ are nonexpansive, i.e., 1-Lipschitz. This assumption is satisfied for all the standard choices of activation operators.

The first upper-bound on the Lipschitz constant of a neural network was derived by analyzing the effect of each layer independently and considering a product of the resulting spectral norms (Goodfellow et al., 2015). This leads to the following Trivial Upper Bound:

$$\bar{\theta}_m = \|W_m\|_S \|W_{m-1}\|_S \dots \|W_1\|_S. \quad (4)$$

Although easy to compute, this upper bound turns out to be over-pessimistic. In (Virmaux and Scaman, 2018), the problem of computing the exact Lipschitz constant of a differentiable function is pointed out to be NP-hard. A first generic algorithm (AutoLip) for upper bounding the Lipschitz constant of any differentiable function is proposed. This bound however reduces to Eq. 4 for standard feedforward neural networks. Additionally, the authors proposed an algorithm, called SeqLip, for sequential neural networks, which shows significant improvement over AutoLip. A sequential neural network is a network for which the activation operators are separable in the sense that, for every $i \in (1, \dots, m)$,

$$(\forall x_i = (\xi_{i,k})_{1 \leq k \leq N_i} \in \mathbb{R}^{N_i}) \quad R_i(x) = (\rho_i(\xi_{i,k}))_{1 \leq k \leq N_i}, \quad (5)$$

where the activation function $\rho_i: \mathbb{R} \rightarrow \mathbb{R}^1$. In (Virmaux and Scaman, 2018), it is assumed that the functions $(\rho_i)_{1 \leq i \leq m}$ are differentiable, increasing, and their derivative are upper bounded by one. It can be deduced that a Lipschitz constant of the network is

$$\vartheta_m = \sup_{\Lambda_1 \in \mathcal{D}_{N_1}(\{0,1\}), \dots, \Lambda_{m-1} \in \mathcal{D}_{N_{m-1}}(\{0,1\})} \|W_m \Lambda_{m-1} \dots \Lambda_1 W_1\|_S, \quad (6)$$

where $\mathcal{D}_N(I)$ designates the set of diagonal matrices of dimension $N \times N$ with diagonal values in $I \subset \mathbb{R}$. This bound simplifies as

$$\vartheta_m = \sup_{\Lambda_1 \in \mathcal{D}_{N_1}(\{0,1\}), \dots, \Lambda_{m-1} \in \mathcal{D}_{N_{m-1}}(\{0,1\})} \|W_m \Lambda_{m-1} \dots \Lambda_1 W_1\|_S, \quad (7)$$

which shows that 2^{N_i} values of the diagonal elements of matrix Λ_i have to be tested at each layer $i \in (1, \dots, m)$, so that the global complexity amounts to $2^{N_1 + \dots + N_{m-1}}$ and thus grows exponentially as a function of the number of neurons. Estimating the Lipschitz constant using this method is intractable even for medium-size networks; thus, the authors use a greedy algorithm to compute a bound, which may under-approximate the Lipschitz constant. This does not provide true upper bounds.

In Combettes and Pesquet (2020b) various bounds on the Lipschitz constant of a feedforward network are derived by assuming that, for every $i \in (1, \dots, m)$ the activation operator R_i is α_i -averaged with $\alpha_i \in [0, 1]$. We recall that this means that there exists a non-expansive (i.e., 1-Lipschitz) operator Q_i such that $R_i = (1-\alpha_i)Id + \alpha_i Q_i$. The following inequality is then satisfied:

$$(\forall (x, y) \in \mathbb{R}^{N_i}) \quad \|R_i(x) - R_i(y)\|^2 \leq \|x - y\|^2 - \frac{1 - \alpha_i}{\alpha_i} \|x - R_i(x) - y + R_i(y)\|^2. \quad (8)$$

We thus see that the smaller α_i , the more “stable” R_i is. In the limit case when $\alpha_i = 1$, R_i is non-expansive and, when $\alpha_i = 1/2$, R_i is said to be firmly nonexpansive. An important subclass of firmly nonexpansive operators is the class of proximity operators of convex functions which are proper and lower-semicontinuous. Let $\Gamma_0(\mathbb{R}^N)$ be the class of such functions defined from \mathbb{R}^N to $] -\infty, +\infty]$. The proximity operator of a function $f \in \Gamma_0(\mathbb{R}^N)$, at some point $x \in \mathbb{R}^N$, is the unique vector denoted by $\text{prox}_f(x)$ such that

$$\text{prox}_f(x) = \underset{p \in \mathbb{R}^N}{\operatorname{argmin}} \frac{1}{2} \|p - x\|^2 + f(p). \quad (9)$$

The proximity operator is a fundamental tool in convex optimization. As shown in (Combettes and Pesquet, 2020a), the point is that most of the activation functions (e.g., sigmoid, ReLU, leaky ReLU, ELU) currently used in neural networks are the proximity operators of some proper lower-semicontinuous convex functions. This property is also satisfied by activation operators which are not separable, like softmax or the squashing function used in capsule networks. The few

¹More generally, a function $\rho_{i,k}$ can be applied to each component $\xi_{i,k}$ but this situation rarely happens in standard neural networks.

activation operators which are not proximity operators (e.g., convex combinations of a max pooling and an average pooling) can be viewed as over-relaxations of proximity operators and correspond to a value of the averaging parameter greater than 1/2.

Based on these averaging assumptions, a first estimation of the Lipschitz constant is given by

$$\theta_m = \beta_{m;\emptyset} \|W_m \circ \dots \circ W_1\| + \sum_{k=1}^{m-1} \sum_{(j_1, \dots, j_k) \in \mathbb{J}_{m,k}} \beta_{m;\{j_1, \dots, j_k\}} \sigma_{m;\{j_1, \dots, j_k\}}, \quad (10)$$

where

$$(\forall \mathbb{J} \subset \{1, \dots, m-1\}) \quad \beta_{m;\mathbb{J}} = \left(\prod_{j \in \mathbb{J}} \alpha_j \right) \prod_{j \in \{1, \dots, m-1\} \setminus \mathbb{J}} (1 - \alpha_j), \quad (11)$$

for every $k \in (1, \dots, m-1)$,

$$\mathbb{J}_{m,k} = \begin{cases} \{(j_1, \dots, j_k) \in \mathbb{N}^k \mid 1 \leq j_1 < \dots < j_k \leq m-1\}, & \text{if } k > 1; \\ \{1, \dots, m-1\}, & \text{if } k = 1 \end{cases} \quad (12)$$

and for every $(j_1, \dots, j_k) \in \mathbb{J}_{m,k}$,

$$\sigma_{m;\{j_1, \dots, j_k\}} = \|W_m \dots W_{j_k+1}\|_S \|W_{j_k} \dots W_{j_{k-1}+1}\|_S \dots \|W_{j_1} \dots W_1\|_S \quad (13)$$

When, for every $i \in (1, \dots, m-1)$, R_i is firmly nonexpansive, the expression simplifies as

$$\theta_m = \frac{1}{2^{m-1}} \left(\|W_m \dots W_1\|_S + \sum_{k=1}^{m-1} \sum_{(j_1, \dots, j_k) \in \mathbb{J}_{m,k}} \sigma_{m;\{j_1, \dots, j_k\}} \right) \quad (14)$$

If, for every $i \in (1, \dots, m-1)$, R_i is separable², a second estimation is provided which reads

$$\vartheta_m = \sup_{\substack{\Lambda_1 \in \mathcal{D}_{N_1}(\{2\alpha_1-1, 1\}), \\ \vdots \\ \Lambda_{m-1} \in \mathcal{D}_{N_{m-1}}(\{2\alpha_{m-1}-1, 1\})}} \|W_m \Lambda_{m-1} \dots \Lambda_1 W_1\|_S \quad (15)$$

We thus see that, when $\alpha_1 = \dots = \alpha_{m-1} = 1/2$, we recover **Eq. 7** without making any assumption on the differentiability of the activation functions. This estimation is more accurate than the previous one in the sense that

$$\|W_m \dots W_1\|_S \leq \vartheta_m \leq \theta_m. \quad (16)$$

It is proved in (Combettes and Pesquet, 2020b) that, if the network is with non-negative weights, that is $(\forall i \in \{1, \dots, m\}) W_i \in [0, +\infty^{N_i \times N_i}]$, the lower bound in **Eq. 16** is attained, i.e.,

$$\vartheta_m = \|W_m \dots W_1\|_S. \quad (17)$$

Another interesting result which is established in (Combettes and Pesquet, 2020b) is that similar results hold if other norms

than the Euclidean norm are used to quantify the perturbations on the input and the output. For example, for a given $i \in (1, \dots, m)$, for every $p \in (1, +\infty)$, we can define the following norm:

$$(\forall x_i = (\xi_{i,k})_{1 \leq k \leq N_i} \in \mathbb{R}^{N_i})$$

$$\|x\|_p = \begin{cases} \sum_{k=1}^{N_i} |\xi_{i,k}|^{p/p}, & \text{if } p < +\infty \\ \sup_{1 \leq k \leq N_i} |\xi_{i,k}|, & \text{if } p = +\infty. \end{cases} \quad (18)$$

If $(p, q) \in (1, +\infty)^2$, the input space \mathbb{R}^{N_0} is equipped with the norm $\|\cdot\|_p$, and the output space \mathbb{R}^{N_m} is equipped with the norm $\|\cdot\|_q$, a Lipschitz constant for a network with separable activation operators is

$$\vartheta_m = \sup_{\substack{\Lambda_1 \in \mathcal{D}_{N_1}(\{2\alpha_1-1, 1\}), \\ \vdots \\ \Lambda_{m-1} \in \mathcal{D}_{N_{m-1}}(\{2\alpha_{m-1}-1, 1\})}} \|W_m \Lambda_{m-1} \dots \Lambda_1 W_1\|_{p,q} \quad (19)$$

$$= \sup_{\substack{\Lambda_1 \in \mathcal{D}_{N_1}(\{2\alpha_1-1, 1\}), \\ \vdots \\ \Lambda_{m-1} \in \mathcal{D}_{N_{m-1}}(\{2\alpha_{m-1}-1, 1\})}} \|W_m \Lambda_{m-1} \dots \Lambda_1 W_1\|_{p,q} \quad (20)$$

where $\|\cdot\|_{p,q}$ is the subordinate $L_{p,q}$ matrix norm induced by the two previous norms. The ability to use norms other than the Euclidean one may be sometimes more meaningful in practice (especially for the ℓ_1 or the sup norm). However, computing such a subordinate norm is not always easy (Lewis, 2010).

2.2 SDP-Based Approach

The work in (Fazlyab et al., 2019) focuses on neural networks using separable activation operators. It assumes that the activation function ρ_i used at a layer $i \in (1, \dots, m)$ is slope-bounded, i.e., there exist nonnegative parameters Υ_{\min} and Υ_{\max} such that

$$(\forall (\xi, \xi') \in \mathbb{R}^2) \quad \xi \neq \xi' \Rightarrow \Upsilon_{\min} \leq \frac{\rho_i(\xi) - \rho_i(\xi')}{\xi - \xi'} \leq \Upsilon_{\max}.$$

As said by the authors, most activation functions satisfy this inequality with $\Upsilon_{\min} = 0$ and $\Upsilon_{\max} = 1$. In other words, the above inequality means that ρ_i is an increasing function and nonexpansive. But a known result (Combettes and Pesquet, 2008, Proposition 1.4) states that a function ρ_i satisfies these properties if and only if it is the proximity operator of some proper lower-semicontinuous convex function. So it turns out that we recover similar assumptions to those made in (Combettes and Pesquet, 2020a).

Let us thus assume that $\Upsilon_{\min} = 0$, $\Upsilon_{\max} = 1$, and $m \geq 2$. A known property is that R_i is firmly nonexpansive if and only if

$$(\forall (x, y) \in (\mathbb{R}^{N_i})^2) \quad (x - y)^\top (R_i(x) - R_i(y)) \geq \|R_i(x) - R_i(y)\|^2. \quad (21)$$

The point is that, if R_i is a separable operator, this inequality holds in a more general metric associated with a matrix

$$Q_i = \text{Diag}(q_{i,1,1}, \dots, q_{i,N_i,N_i}), \quad (22)$$

²The result remains valid if different scalar activation functions are used in a given layer.

where $(\forall k \in \{1, \dots, N_i\}^2) q_{i,k,k} \geq 0$. In the following, the set of such matrices $(Q_i)_{1 \leq i \leq m-1}$ will be denoted by \mathcal{Q} . This means that

$$(\forall (x, y) \in (\mathbb{R}^{N_i})^2) \quad (x - y)^\top Q_i (R_i(x) - R_i(y)) \geq (R_i(x) - R_i(y))^\top Q_i (R_i(x) - R_i(y)). \quad (23)$$

For every $(x_i, y_i) \in (\mathbb{R}^{N_i})^2$, let $x_i = R_i(W_i x_{i-1} + b_i)$ and $y_i = R_i(W_i y_{i-1} + b_i)$. It follows from **Eq. 23** that

$$(W_i(x_{i-1} - y_{i-1}))^\top Q_i (x_i - y_i) \geq (x_i - y_i)^\top Q_i (x_i - y_i). \quad (24)$$

Summing for the first $m-1$ layers yields

$$\sum_{i=1}^{m-1} (W_i(x_{i-1} - y_{i-1}))^\top Q_i (x_i - y_i) \geq \sum_{i=1}^{m-1} (x_i - y_i)^\top Q_i (x_i - y_i). \quad (25)$$

On the other hand, $\vartheta_m > 0$ is a Lipschitz constant of the neural network T if

$$\vartheta_m^2 \|x_0 - y_0\|^2 \geq \|W_m(x_{m-1} - y_{m-1})\|^2. \quad (26)$$

For the latter inequality to hold, it is thus sufficient to ensure that

$$\begin{aligned} \vartheta_m^2 \|x_0 - y_0\|^2 - \|W_m(x_{m-1} - y_{m-1})\|^2 &\geq 2 \\ &\times \sum_{i=1}^{m-1} (W_i(x_{i-1} - y_{i-1}))^\top Q_i (x_i - y_i) - 2 \\ &\times \sum_{i=1}^{m-1} (x_i - y_i)^\top Q_i (x_i - y_i). \end{aligned} \quad (27)$$

This inequality can be rewritten in matrix form as

$$\begin{bmatrix} x_0 - y_0 \\ \vdots \\ x_{m-1} - y_{m-1} \end{bmatrix}^\top M(\rho_m, Q_1, \dots, Q_{m-1}) \begin{bmatrix} x_0 - y_0 \\ \vdots \\ x_{m-1} - y_{m-1} \end{bmatrix} \geq 0 \quad (28)$$

with $\rho_m = \vartheta_m^2$ and

$$M(\rho_m, Q_1, \dots, Q_{m-1}) = \begin{bmatrix} \rho_m \text{Id}_{N_0} & -W_1^\top Q_1 & & 0 \\ -Q_1 W_1 & 0 & \ddots & \\ & \ddots & \ddots & \\ 0 & & -Q_{m-1} W_{m-1} & 2Q_{m-1} - W_m^\top W_m \end{bmatrix}. \quad (29)$$

In the case of a network having just one hidden layer, which is mainly investigated in (Fazlyab et al., 2019), the above matrix reduces to

$$M(\rho_2, Q_1) = \begin{bmatrix} \rho_2 \text{Id}_{N_0} & -W_1^\top Q_1 \\ -Q_1 W_1 & 2Q_1 - W_2^\top W_2 \end{bmatrix}. \quad (30)$$

Condition **Eq. 28** is satisfied, for every (x_0, \dots, x_{m-1}) and (y_0, \dots, y_{m-1}) if and only if

$$M(\rho_m, Q_1, \dots, Q_{m-1}) \geq 0. \quad (31)$$

It is actually sufficient that this positive semidefiniteness constraint be satisfied for any matrices $(Q_1, \dots, Q_{m-1}) \in \mathcal{Q}$ for $\sqrt{\rho_m}$ to be a Lipschitz constant. The smallest possible value of the resulting constant can be obtained by solving the following Semidefinite Programming (SDP) problem:

$$\underset{(\rho_m, Q_1, \dots, Q_{m-1}) \in C}{\text{minimize}} \quad \rho_m, \quad (32)$$

where C is the closed convex set

$$C = \{(\rho_m, Q_1, \dots, Q_{m-1}) \in [0, +\infty[\times \mathcal{Q} \mid (31) \text{ holds}\}. \quad (33)$$

Although there exists efficient SDP solvers, the method remains computationally intensive. A solution to reduce its computational complexity at the expense of a lower accuracy consists of restricting the optimization of the metric matrices Q_1, \dots, Q_{m-1} to a subset of \mathcal{Q} .

One limitation of this method is that it is tailored to the use of the Euclidean norm.

Remark 1. In (Fazlyab et al., 2019), it is claimed that **Eq. 23** is valid for every metric matrix

$$Q_i = \sum_{k=1}^{N_i} q_{i,k,k} e_k e_k^\top + \sum_{1 \leq k < \ell \leq N_i} q_{i,k,\ell} (e_k - e_\ell)(e_k - e_\ell)^\top, \quad (34)$$

where $(e_k)_{1 \leq k \leq N_i}$ is the canonical basis of \mathbb{R}^{N_i} and $(\forall (k, \ell) \in \{1, \dots, N_i\}^2)$ with $k \leq \ell$, $q_{i,k,\ell} \geq 0$. Unfortunately, this turns out to be incorrect. The erroneous statement comes from a flaw in the deduction of Lemma 1 from Lemma 2 in (Fazlyab et al., 2019). A counterexample was recently provided in (Pauli et al., 2022).

2.3 Polynomial Optimization Based Approach

The approach in (Latorre et al., 2020) applies to neural networks having a single output (i.e., $N_m = 1$)³. The authors mention that their approach is restricted to differentiable activation functions, but it is actually valid for any separable firmly nonexpansive activation operators. Indeed, when $N_m = 1$, the Lipschitz constant in **Eq. 19** reduces to

$$\vartheta_m = \sup_{\substack{\Lambda_1 \in \mathcal{D}_{N_1}([0,1]), \\ \vdots \\ \Lambda_{m-1} \in \mathcal{D}_{N_{m-1}}([0,1])}} \|W_1^\top \Lambda_1 \dots \Lambda_{m-1} W_m^\top\|_{p^*}, \quad (35)$$

where $p^* \in (1, +\infty)$ is the dual exponent of p (such that $1/p + 1/p^* = 1$). Recall that $p \in (1, +\infty)$ is the exponent of the ℓ_p -norm equipping the input space. This shows that ϑ_m is equal to

$$\vartheta_m = \sup \left\{ \Phi(x, \lambda_1, \dots, \lambda_{m-1}) \mid \|x\|_p \leq 1, (\lambda_i)_{1 \leq i \leq m-1} \in [0, 1]^{N_1 + \dots + N_{m-1}} \right\}, \quad (36)$$

³This can be extended to multiple output network, if the output space is equipped with the $\ell_{+\infty}$ norm.

TABLE 1 | Comparison of state-of-the-art Lipschitz estimation approaches vs the proposed one.

Method	Properties	Sensitivity of inputs
Naive upper Bound [Goodfellow et al. (2014)]	spectral bound, loose bound, univariate	No
SDPLip [Fazlyab et al. (2019)]	ℓ_2 norm, more scalable to broad networks, univariate	No
CPLip [Combettes and Pesquet (2020b)]	$\ell_p \in (1, +\infty)$, not scalable to broad networks, univariate	No
LipOpt-k [Latorre et al. (2020)]	$\ell_p \in (1, +\infty)$, univariate	No
Proposed	scalable to broad networks, multivariate	Yes

where, for every $x \in \mathbb{R}^{N_0}$ and $(\lambda_i)_{1 \leq i \leq m} \in \mathbb{R}^{N_1 + \dots + N_{m-1}}$,

$$\Phi(x, \lambda_1, \dots, \lambda_{m-1}) = x^\top W_1^\top \text{Diag}(\lambda_1) \dots \text{Diag}(\lambda_{m-1}) W_m^\top. \quad (37)$$

Function Φ is a multivariate polynomial of the components of its vector arguments. Therefore, if the unit ball associated with the ℓ_p norm can be described via polynomial inequalities, which happens when $p \in \mathbb{N} \setminus \{0\}$ and $p = +\infty$, then finding ϑ_m turns out to be a polynomial constrained optimization problem. Solving such an optimization problem can be achieved by solving a hierarchy of convex problems. However, the size of the hierarchy tends to grow fast and if the order of the hierarchy is truncated to a too small value, the delivered result becomes inaccurate. Leveraging the sparsity properties that might exist for the weight matrices may be helpful numerically. Note that, the approach is further improved in (Chen et al., 2020) by using Lasserre's hierarchy.

A comparison of the state-of-the-art and proposed approach is presented in **Table 1**.

3 WEIGHTED LIPSCHITZ CONSTANTS FOR SENSITIVITY ANALYSIS

To extend the theoretical results presented above on the evaluation of neural network stability through their Lipschitz regularity, we present in this section a new approach based on a suitable weighting operation performed in the computation of Lipschitz constants. This enables a multivariate sensitivity analysis of the neural network stability for individual inputs or groups of inputs. We will start by motivating this weighting from a statistical standpoint. Then we will define it in a more precise manner, before discussing its resulting mathematical properties.

3.1 Statistical Motivations

For tractability, assume that the perturbation at the network input is a realization of a zero-mean Gaussian distributed random vector z with $N_0 \times N_0$ covariance matrix $\Sigma > 0$. Then, its density upper level sets are defined as

$$C_\eta = \{z \in \mathbb{R}^{N_0} \mid z^\top \Sigma^{-1} z \leq \eta\}, \quad (38)$$

for every $\eta \in]0, +\infty$. The set C_η defines an ellipsoid where the probability density takes its highest values. More precisely, the probability for z to belong to this set is independent of Σ (**Supplementary Appendix S1**) and is equal to.

$$P(z \in C_\eta) = \frac{\gamma(N_0/2, \eta/2)}{\Gamma(N_0/2)}, \quad (39)$$

where Γ is the gamma function and γ the lower (unnormalized) incomplete gamma function.

On the other hand, let us assume that the maximum standard deviation σ_{\max} of the components of z (i.e., square root of the maximum diagonal element of matrix Σ) is small enough. If we suppose that the network T is differentiable in the neighborhood of a given input $x \in \mathbb{R}^{N_0}$, as the input perturbation is small enough, we can approximate the network output by the following expansion:

$$T(x+z) \simeq T(x) + T'(x)z. \quad (40)$$

Let us focus our attention on perturbations in C_η . By doing so, we impose some norm-bounded condition, which may appear more realistic for adversarial perturbations. Then, we will be interested in calculating

$$\sup_{z \in C_\eta} \|T(x+z) - T(x)\| \simeq \sup_{z \in C_\eta} \|T'(x)z\|. \quad (41)$$

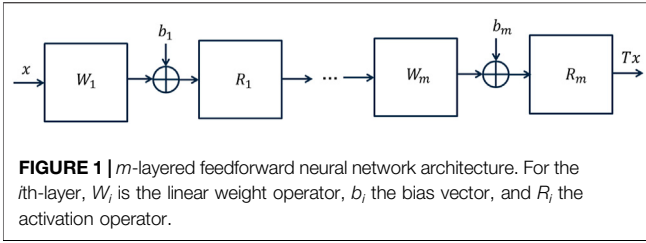
By making the variable change $z' = z/\sqrt{\eta}$ and using **Eq. 38**,

$$\begin{aligned} \sup_{z \in C_\eta} \|T(x+z) - T(x)\| &\simeq \sqrt{\eta} \sup_{z' \in C_1} \|T'(x)z'\| \\ &= \sqrt{\eta} \sup_{\substack{z' \in \mathbb{R}^{N_0} \\ z' \neq 0}} \frac{\|T'(x)z'\|}{\|z'\|_{\Sigma^{-1}}} \\ &= \sqrt{\eta} \sigma_{\max} \sup_{\substack{z' \in \mathbb{R}^{N_0} \\ z' \neq 0}} \frac{\|T'(x)z'\|}{\|z'\|_{\Omega^{-1}}}, \end{aligned} \quad (42)$$

where $\Omega = \Sigma/\sigma_{\max}^2$ and $\|\cdot\|_{\Omega^{-1}} = \sqrt{(\cdot)^\top \Omega^{-1} (\cdot)}$. This suggests that, in this context, the suitable subordinate matrix norm for computing the Lipschitz constant in **Eq. 3** is obtained by weighting the Euclidean norm in the input space with Ω^{-1} . We can also deduce from **Eq. 42**, by setting $z'' = \Omega^{-1/2}z'$, that

$$\begin{aligned} \sup_{z \in C_\eta} \|T(x+z) - T(x)\| &\simeq \sqrt{\eta} \sigma_{\max} \sup_{\substack{z'' \in \mathbb{R}^{N_0} \\ z'' \neq 0}} \frac{\|T'(x)\Omega^{1/2}z''\|}{\|z''\|} \\ &= \sqrt{\eta} \|T'(x)\Sigma^{1/2}\|_S. \end{aligned} \quad (43)$$

On the other hand, based on the first-order approximation in **Eq. 40**, $T(x+z)$ is approximately Gaussian with mean $T(x)$ and covariance matrix $T'(x)\Sigma T'(x)^\top$. As $\|T'(x)\Sigma T'(x)^\top\|_S = \|T'(x)\Sigma^{1/2}\|_S^2$, we see that another insightful interpretation of **Eq. 43** is that, up to the scaling factor $\sqrt{\eta}$, it approximately delivers the square root of the



spectral norm of the covariance matrix of the output perturbations.

3.2 New Definition of a Weighted Lipschitz constant

Based on the previous motivations, we propose to employ a weighted norm to define a Lipschitz constant of the network as follows:

Definition 1. Let Ω be an $N_0 \times N_0$ symmetric positive definite real-valued matrix. We say that $\{\theta\}_{\Omega}$ is an Ω -weighted norm Lipschitz constant of T as described in Figure 1 if

$$(\forall (x, z) \in (\mathbb{R}^{N_0})^2) \quad \|T(x + z) - T(z)\| \leq \theta_m^\Omega \|z\|_{\Omega^{-1}}. \quad (44)$$

The above definition can be extended to non Euclidean norms by making use of exponents $(p, q) \in (1, +\infty)^2$ and by replacing inequality Eq. 44 with

$$(\forall (x, z) \in (\mathbb{R}^{N_0})^2) \quad \|T(x + z) - T(z)\|_q \leq \theta_m^\Omega \|\Omega^{-1/2} z\|_p. \quad (45)$$

By changes of variable, this inequality can also be rewritten as.

$$(\forall (x', z') \in (\mathbb{R}^{N_0})^2) \quad \|T(\Omega^{1/2}(x' + z')) - T(\Omega^{1/2}z')\|_q \leq \theta_m^\Omega \|z'\|_p. \quad (46)$$

Therefore, we see that calculating θ_m^Ω is equivalent to derive a Lipschitz constant of the network T where an additional first linear layer $\Omega^{1/2}$ has been added. Throughout the rest of this section, it will be assumed that, for every $i \in (1, \dots, m-1)$ the activation operator R_i is separable and α_i -averaged. It then follows from Eq. 20 that an Ω -weighted norm Lipschitz constant of T is

$$\theta_m^\Omega = \sup_{\substack{\Lambda_1 \in \mathcal{D}_{N_1}(\{2\alpha_1-1, 1\}), \\ \Lambda_{m-1} \in \mathcal{D}_{N_{m-1}}(\{2\alpha_{m-1}-1, 1\})}} \|W_m \Lambda_{m-1} \dots \Lambda_1 W_1 \Omega^{1/2}\|_{p,q}. \quad (47)$$

Although all our derivations were based on the fact that Ω is positive definite, from the latter expression we see that, by continuous extension, θ_m^Ω can be defined when Ω is a singular matrix.

3.3 Sensitivity with Respect to a Group of Inputs

In this section, we will be interested in a specific family of weighted norms associated with the set of matrices

$$\{\Omega_{\epsilon, \mathbb{K}} \mid \emptyset \neq \mathbb{K} \subset \{1, \dots, N_0\}, \epsilon \in]0, 1]\},$$

defined, for every nonempty subset \mathbb{K} of $(1, \dots, N_0)$ and for every $\epsilon \in]0, 1]$, as

$$\Omega_{\epsilon, \mathbb{K}} = \text{Diag}(\sigma_{\epsilon, \mathbb{K}, 1}^2, \dots, \sigma_{\epsilon, \mathbb{K}, N_0}^2), \quad (48)$$

where

$$(\forall \ell \in \{1, \dots, N_0\}) \quad \sigma_{\epsilon, \mathbb{K}, \ell} = \begin{cases} 1 & \text{if } \ell \in \mathbb{K} \\ \epsilon & \text{otherwise.} \end{cases} \quad (49)$$

If we come back to the statistical interpretation in Section 3.1, $\Omega_{\epsilon, \mathbb{K}}$ is then (up to a positive scale factor) the covariance matrix of a Gaussian random vector Z with independent components⁴. The components with indices in \mathbb{K} have a given variance σ_{\max}^2 while the others have variance $\epsilon^2 \sigma_{\max}^2$. Such a matrix thus provides a natural way of putting emphasis on the group of inputs with indices in \mathbb{K} . Thus, variables $\theta_m^{\Omega_{\epsilon, \mathbb{K}}}$ will be termed partial Lipschitz constants in the following.

The next proposition lists the main properties related to the use of such weighted norms for calculating Lipschitz constants. The proofs of these results are given in **Supplementary Appendix S2**.

Proposition 1. Let $(p, q) \in (1, +\infty)^2$. For every nonempty subset \mathbb{K} of $(1, \dots, N_0)$ and for every $\epsilon \in]0, 1]$, let $\Omega_{\epsilon, \mathbb{K}}$ be defined as above and let $\theta_m^{\Omega_{\epsilon, \mathbb{K}}}$ be defined by (47). Let \mathbb{K}_0 and \mathbb{K}_1 be nonempty subsets of $(1, \dots, N_0)$. Then the following hold:

- 1) As $\epsilon \rightarrow 0$, $\theta_m^{\Omega_{\epsilon, \mathbb{K}_0}}$ converges to the Lipschitz constant of a network where all the inputs with indices out of \mathbb{K}_0 are kept constant.
- 2) $\theta_m^{\Omega_{1, \mathbb{K}_0}}$ is equal to the global Lipschitz constant θ_m defined by Eq. 20.
- 3) Let $(\epsilon, \epsilon') \in]0, 1]^2$. If $\Omega_{\epsilon, \mathbb{K}_0} \leq \Omega_{\epsilon', \mathbb{K}_1}$, then $\theta_m^{\Omega_{\epsilon, \mathbb{K}_0}} \leq \theta_m^{\Omega_{\epsilon', \mathbb{K}_1}}$.
- 4) Function $\theta_m^{\Omega_{\epsilon, \mathbb{K}_0}} :]0, 1] \rightarrow [0, +\infty[: \epsilon \mapsto \theta_m^{\Omega_{\epsilon, \mathbb{K}_0}}$ is monotonically increasing.
- 5) Let $\epsilon \in]0, 1]$. If $\mathbb{K}_0 \subset \mathbb{K}_1$, then $\theta_m^{\Omega_{\epsilon, \mathbb{K}_0}} \leq \theta_m^{\Omega_{\epsilon, \mathbb{K}_1}}$.
- 6) Let $\epsilon \in]0, 1]$, let $K \in \mathbb{N} \setminus \{0\}$, and let

$$\omega_{K, \epsilon} = \left(\frac{N_0 - 1}{K - 1} \right) \left(1 + \left(\frac{N_0}{K} - 1 \right) \epsilon \right). \quad (50)$$

We have

$$\max_{\substack{\mathbb{K} \subset \{1, \dots, N_0\} \\ \text{card} \mathbb{K} = K}} \theta_m^{\Omega_{\epsilon, \mathbb{K}}} \leq \theta_m \leq \frac{1}{\omega_{K, \epsilon}} \sum_{\substack{\mathbb{K} \subset \{1, \dots, N_0\} \\ \text{card} \mathbb{K} = K}} \theta_m^{\Omega_{\epsilon, \mathbb{K}}}. \quad (51)$$

- 7) Let $\epsilon \in]0, 1]$, let \mathcal{P} be a partition of $(1, \dots, N_0)$, and let.

$$\omega_{\mathcal{P}, \epsilon} = 1 + (\text{card} \mathcal{P} - 1) \epsilon.$$

We have

$$\max_{\mathbb{K} \in \mathcal{P}} \theta_m^{\Omega_{\epsilon, \mathbb{K}}} \leq \theta_m \leq \frac{1}{\omega_{\mathcal{P}, \epsilon}} \sum_{\mathbb{K} \in \mathcal{P}} \theta_m^{\Omega_{\epsilon, \mathbb{K}}}. \quad (52)$$

⁴Recall that this interpretation is valid when $p = 2$ in Eq. 47.

- 8) Let \mathbb{K}_2 be such that $\mathbb{K}_1 \cap \mathbb{K}_2 \neq \emptyset$ and $\mathbb{K}_1 \cup \mathbb{K}_2 = \mathbb{K}_0$. Let $p^* \in (1, +\infty)$ be such that $1/p + 1/p^* = 1$. Then

$$\vartheta_m^{\Omega_{\epsilon, \mathbb{K}_0}} \leq \left(\left(\vartheta_m^{\Omega_{\epsilon, \mathbb{K}_1}} \right)^{p^*} + \left(\vartheta_m^{\Omega_{\epsilon, \mathbb{K}_2}} \right)^{p^*} \right)^{1/p^*} + o(\epsilon). \quad (53)$$

Let us comment on these results. According to Property (i) in the limit case when $\epsilon \rightarrow 0$, only the inputs with indices in \mathbb{K}_0 are used in the computation of the associated Lipschitz constant. In turn, Property (ii) states that, when $\epsilon = 1$, we recover the classical expression of a Lipschitz constant where the perturbations on all the inputs are taken into account. In addition, based on Property (iv), the evolution of $\vartheta_m^{\Omega_{\epsilon, \mathbb{K}_0}}$ when ϵ varies from 1 to 0 provides a way of assessing how the group of inputs indexed by \mathbb{K}_0 contributes to the overall Lipschitz behaviour of the network. Although one would expect that summing the Lipschitz constants obtained for each group of inputs would yield the global Lipschitz constant, Properties (vi) and (vii) show that this does not hold in general whatever the way the entries are split (possibly overlapping groups of given size K or disjoint groups of arbitrary size). Instead, after suitable normalization, such sums provide upper bounds on ϑ_m . Furthermore, it follows from (2), Eqs 51, 52 that the difference between these normalized sums and ϑ_m tends to vanish when ϵ increases.

Note that, when looking at the sensitivity with respect to individual inputs, i.e., when the considered set of indices are singletons, both (6) (with $K = 1$) and (7) (with $\mathcal{P} = \{\{k\} \mid k \in \{1, \dots, N_0\}\}$) lead to the same inequality

$$\max_{k \in \{1, \dots, N_0\}} \vartheta_m^{\Omega_{\epsilon, \{k\}}} \leq \vartheta_m \leq \frac{1}{1 + (N_0 - 1)\epsilon} \sum_{k=1}^{N_0} \vartheta_m^{\Omega_{\epsilon, \{k\}}}. \quad (54)$$

4 VALIDATION ON SYNTHETIC DATA

4.1 Context

To highlight the need for advanced sensitivity analysis tools in the design of neural networks, we first study simple synthetic examples of polynomial systems for which we can calculate explicitly the partial Lipschitz constants. We generate input-output data for the defined systems, and train a fully connected model using a standard training, i.e., without any constraints. We compare this approach with a training subject to a spectral norm constraint on the layers.

Spectral Normalization: For safety critical tasks, Lipschitz constant and performance targets can be specified as engineering requirements, prior to network training. A Lipschitz target can be defined by a safety analysis of the acceptable perturbations for each output knowing the input range and it constitutes a current practice in many industries. Imposing this Lipschitz target can be done either by controlling the Lipschitz constant for each layer or for the whole network depending on the application at hand. Such a work for controlling the Lipschitz constant has been presented in (Serrurier et al., 2021) using Hinge regularization. In our experiments, we train networks while using a spectral normalization technique (Miyato et al., 2018) which has been proved to be effective in controlling Lipschitz

properties in GANs. Given an m layer fully connected architecture and a Lipschitz target L , we can constrain the spectral norm of each layer to be less than $\sqrt[m]{L}$. According to Eq. 4, this ensures that the upper bound on the global Lipschitz constant is less than L .

For each training, we study the effect of input variables on the stability of the networks. As proposed in Section 3.3, for a given group of inputs with indices in \mathbb{K} , we will quantify the partial Lipschitz constant $\vartheta_m^{\Omega_{\epsilon, \mathbb{K}}}$. The obtained value of $\vartheta_m^{\Omega_{\epsilon, \mathbb{K}}}$ allows us to evaluate how the corresponding group of variables may potentially affect the stability of the network. For simplicity, in this section, we will focus on the limit case when $\epsilon = 0$ (see the last remark in Section 3.2).

Partial Lipschitz constant values $\vartheta_m^{\Omega_{0, \mathbb{K}}}$, for all possible choices for \mathbb{K} , are computed using the numerical method described in Section 2.2 and compared with the theoretical values derived in the following subsection. More details on the models are also provided in these sections.

4.2 Polynomial Systems

We consider regression problems where the data is synthesized by a second-order multivariate polynomial. The system to be modelled is thus described by the following function:

$$\begin{aligned} & (\forall (\xi_1, \dots, \xi_{N_0}) \in \mathbb{R}^{N_0}) \quad f(\xi_1, \dots, \xi_{N_0}) \\ &= \sum_{k=1}^{N_0} a_k \xi_k + \sum_{k=1}^{N_0} \sum_{l=1}^{N_0} b_{k,l} \xi_k \xi_l, \end{aligned} \quad (55)$$

where $(a_k)_{k \in N_0}$ and $(b_{k,l})_{1 \leq k, l \leq N_0}$ are the real-valued polynomial coefficients. Note that, such a polynomial system is generally not Lipschitz-continuous. The Lipschitz-continuity property only holds on every compact set. Subsequently, we will thus study this system on the hypercube $[-M, M]^{N_0}$ with $M > 0$.

The explicit values of the partial Lipschitz constant on this domain can be derived as follows. We first calculate the gradient of f

$$\nabla f(\xi_1, \dots, \xi_{N_0}) = (\partial_k f(\xi_1, \dots, \xi_{N_0}))_{1 \leq k \leq N_0}, \quad (56)$$

where, for every $k \in (1, \dots, N_0)$, $\partial_k f$ denotes the partial derivative w.r.t. the k -th variable given by

$$\partial_k f(\xi_1, \dots, \xi_{N_0}) = a_k + \sum_{l=1}^{N_0} (b_{k,l} + b_{l,k}) \xi_l. \quad (57)$$

For every $\mathbb{K} \subset \{1, \dots, N_0\}$, the partial Lipschitz constant $\vartheta_m^{\Omega_{0, \mathbb{K}}}$ of the polynomial system (restricted to $[-M, M]^{N_0}$) w.r.t. the group of variables with indices in \mathbb{K} is then equal to.

$$\vartheta_m^{\Omega_{0, \mathbb{K}}} = \sup_{(\xi_1, \dots, \xi_{N_0}) \in [-M, M]^{N_0}} \sqrt{\lambda_{\Omega_{0, \mathbb{K}}}(\xi_1, \dots, \xi_{N_0})}, \quad (58)$$

where, for every diagonal matrix $\Lambda = \text{Diag}(\epsilon_1^2, \dots, \epsilon_{N_0}^2)$ with $(\epsilon_1, \dots, \epsilon_{N_0}) \in [0, +\infty)^{N_0}$,

$$\begin{aligned} \lambda_{\Lambda}(\xi_1, \dots, \xi_{N_0}) &= \|(\nabla f(\xi_1, \dots, \xi_{N_0}))^\top \Lambda^{1/2}\|^2 \\ &= \sum_{k=1}^{N_0} \epsilon_k (\partial_k f(\xi_1, \dots, \xi_{N_0}))^2. \end{aligned} \quad (59)$$

TABLE 2 | Comparison of Lipschitz constant values when $\gamma = 0$. Test performance for standard training: NMSE = 0.007, NMAE = 0.005, for spectral normalization: NMSE = 0.011, NMAE = 0.009.

Partial LC	Analytical	Standard	Spectral normalized
$\Theta_{(1)}$	1	133.9	6.75
$\Theta_{(2)}$	100	211.7	76.3
$\Theta_{(3)}$	100	299.7	136.0
$\Theta_{(1,2)}$	100.0	229.0	102.2
$\Theta_{(1,3)}$	100.0	303.1	136.0
$\Theta_{(2,3)}$	122.5	314.2	141.2
$\Theta_{(1,2,3)}$	141.4	315.3	141.2

Since the partial derivatives in Eq. 57 are affine functions of the variables $(\xi_1, \dots, \xi_{N_0})$, λ_Λ is a convex function. We deduce that the supremum in Eq. 58 is attained when $\xi_1 = \pm M, \dots, \xi_{N_0} = \pm M$, so that $\overset{\circ}{\vartheta}_{\Omega_{0,\mathbb{K}}}$ can be computed by looking for the maximum of a finite number of values.

4.3 Numerical Results

In our numerical experiments, we consider a toy example corresponding to $N_0 = 3$ and

$$(\forall (\xi_1, \xi_2, \xi_3) \in \mathbb{R}^3) \quad f(\xi_1, \xi_2, \xi_3) = \xi_1 + 100\xi_3 - \xi_2^2 + \gamma\xi_1\xi_3, \quad (60)$$

where $\gamma \in [0, +\infty]$. We deduce from Eq. 59 that

$$\lambda_\Lambda(\xi_1, \xi_2, \xi_3) = \varepsilon_1(1 + \gamma\xi_3)^2 + 4\varepsilon_2\xi_2^2 + \varepsilon_3(100 + \gamma\xi_1)^2 \quad (61)$$

and, consequently,

$$\begin{aligned} \sup_{(\xi_1, \xi_2, \xi_3) \in [-M, M]^3} \lambda_\Lambda(\xi_1, \xi_2, \xi_3) &= \varepsilon_1(1 + \gamma M)^2 + 4\varepsilon_2 M^2 \\ &\quad + \varepsilon_3(100 + \gamma M)^2. \end{aligned} \quad (62)$$

By looking at the seven possible binary values of $(\varepsilon_1, \varepsilon_2, \varepsilon_3) \neq (0, 0, 0)$, we thus calculate the Lipschitz constant of f with respect to each group of inputs. For example,

- if $\varepsilon_1 = 1, \varepsilon_2 = 0, \varepsilon_3 = 0$, we calculate $\overset{\circ}{\vartheta}_{\Omega_{0,\mathbb{K}}}$ with $\mathbb{K} = \{1\}$, i.e., evaluate the sensitivity w.r.t. the first variable
- if $\varepsilon_1 = \varepsilon_2 = 1, \varepsilon_3 = 0$, we calculate $\overset{\circ}{\vartheta}_{\Omega_{0,\mathbb{K}}}$ with $\mathbb{K} = \{1, 2\}$, i.e., evaluate the joint sensitivity w.r.t. the first and second variables;
- if $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = 1$, we calculate $\overset{\circ}{\vartheta}_{\Omega_{0,\mathbb{K}}}$ with $\mathbb{K} = \{1, 2, 3\}$, i.e., evaluate the sensitivity w.r.t. all the variables (global Lipschitz constant).

These Lipschitz constants allow us to evaluate the intrinsic dynamics of the system, that is how it responds when its inputs vary.

Our interest will be now to evaluate how this dynamics is modified when the system is modelled by a neural network. To do so, three systems are studied by choosing $\gamma \in (0, 1/10, 1)$ and $M = 50$. We generate 5,000 data samples from each system, the input values being drawn independently from a random uniform distribution. While training the neural networks, the dataset is

TABLE 3 | Comparison of Lipschitz constant values when $\gamma = 1/10$. Test performance for standard training: NMSE = 0.006, NMAE = 0.005, for spectral normalization: NMSE = 0.009, NMAE = 0.007.

Partial LC	Analytical	Standard	Spectral normalized
$\Theta_{(1)}$	6	138.7	10.1
$\Theta_{(2)}$	100	219.2	90.0
$\Theta_{(3)}$	105	302.7	138.9
$\Theta_{(1,2)}$	100.2	231.6	108.1
$\Theta_{(1,3)}$	105.2	306.3	139.0
$\Theta_{(2,3)}$	145	316.4	147.2
$\Theta_{(1,2,3)}$	145.1	316.5	147.2

TABLE 4 | Comparison of Lipschitz constant values when $\gamma = 1$. Test performance for standard training: NMSE = 0.006, MAE = 0.005, for spectral normalization: NMSE = 0.014, NMAE = 0.009.

Partial LC	Analytical	Standard	Spectral normalized
$\Theta_{(1)}$	51	274.7	59.5
$\Theta_{(2)}$	100	298.9	80.3
$\Theta_{(3)}$	150	388.7	183.7
$\Theta_{(1,2)}$	112.6	337.0	119.4
$\Theta_{(1,3)}$	158.4	392.2	183.7
$\Theta_{(2,3)}$	180.3	400.1	188.9
$\Theta_{(1,2,3)}$	187.4	400.5	189.0

divided with a ratio of 4:1 into training and testing samples. The input is normalized using its mean and standard deviation, while the output is max-normalized. We build neural networks for approximating the systems using two hidden layers ($m = 3$) with a number of hidden neurons equal to 30 in each layer and ReLU activation functions. The training loss is the mean square error.

For different values of γ , we report the values of the partial Lipschitz constants in Tables 2, 3, 4. The variable $\theta_{\mathbb{K}}$ corresponds to $\overset{\circ}{\vartheta}_{\Omega_{0,\mathbb{K}}}$ for the analytical value we derived from previous formulas, whereas it corresponds to the Lipschitz constant $\vartheta_{\Omega_{0,\mathbb{K}}}^{\Omega_{0,\mathbb{K}}}$, when computed for the neural network trained either in a standard manner or with a spectral normalization constraint. The value of L used in the spectral normalization was adjusted to obtain a similar global Lipschitz constant to the polynomial system. In the caption, we also indicate the accuracy in terms of normalized mean square error (NMSE) and normalized mean absolute error (NMAE). These values are slightly higher for constrained training, but remain quite small.

Comments on the results:

- In general, ξ_3 impacts the output of this system the most, and (ξ_2, ξ_3) mainly account for the global dynamics of the system.
- With standard training, we see that there exists a significant increase of the sensitivity with respect to the input variations, so making the neural network vulnerable to adversarial perturbations.
- By using spectral normalization, it is possible to constrain the global Lipschitz constant of the system to be close to the analytical global value while keeping a good accuracy. One

TABLE 5 | Input and output variables description for the Thales Air Mobility industrial application dataset.

	Variable	Name	Type
Input	0	Speed	Continuous
	1	Flight distance	
	2	Departure delay	
	3	Initial ETE	
	4	Latitude origin	
	5	Longitude origin	
	6	Altitude origin	
	7	Latitude destination	
	8	Longitude destination	
	9	Altitude destination	
Output	10	Arrival time slot	7 slots (categorical)
	11	Departure time slot	7 slots (categorical)
	12	Aircraft category	6 classes (categorical)
	13	Airline company	19 classes (categorical)
	3	Refinement ETE	continuous

may however notice an increase of the sensitivity to ξ_1 and ξ_3 , and a decrease of the sensitivity to ξ_2 with respect to the original system.

- For all the three models, the values obtained with neural networks follow the same trend, for different groups of inputs, as those observed with the analytical values.
- Although the Lipschitz constant of the neural networks is computed on the whole space and the one of the system on $(-50,50)^3$, our Lipschitz estimates appear to be consistent without resorting to a local analysis.

These observations emphasize the importance of controlling the Lipschitz constant of neural network models through specific training strategies. In addition, we see that evaluating the Lipschitz constant with respect to groups of inputs allow us to have a better understanding of the behaviour of the models.

In this section, we have discussed the proposed method for synthetic datasets. In the next section, the sensitivity analysis will be made on widely used open source datasets and an industrial dataset.

5 APPLICATION ON DIFFERENT USE CASES

5.1 Datasets and Network Description

We study four regression problems involving tabular datasets to showcase our proposed multivariate analysis of the stability of neural networks. Tabular data take advantage of heterogeneous sources of information coming from different sensors or data collection processes. We apply our methods on widely used tabular datasets: 1) Combined Cycle Power Plant dataset⁵ which has 4 attributes with 9,568 instances; 2) Auto MPG dataset⁶ consists of 398 instances with 7 attributes; 3) Boston

Housing dataset⁷ consists of 506 instances with 13 attributes. For Combined Power Plant and Auto MPG datasets, we solve a regression problem with a single output, whereas for Boston Housing dataset we consider a two-output regression problem with “price” and “ptratio” as the output variables. The attributes in the dataset are a combination of continuous and categorical. The datasets are divided with a ratio of 4:1 between training and test data.

Thales Air Mobility industrial application represents the prediction of the Estimated Time En-route (ETE), meaning the time spent by an aircraft between the take-off and landing, considering a number of variables as described in **Table 5**. The application is important in air traffic flow management, which is an activity area where safety is critical. The purpose of the proposed sensitivity analysis is thus to help engineers in building safe by design models complying with given safety stability targets. The dataset consists of 2,219,097 training, 739,639 validation, and 739,891 test samples.

For all the models, we build fully connected networks with ReLU⁸ activation function on all the hidden layers, except the last one. The models are trained on Keras with Tensorflow backend. The initializers are set to Glorot uniform. The network architecture of the different models, number of layers, and neurons are tabulated in **Table 6**. Combined Cycle Power Plant dataset with (10, 6) network architecture is trained with two hidden layers having 10 and 6 hidden neurons, respectively. For Thales Air Mobility industrial application $[10 \times (30)]$ implies that the neural network has 10 hidden layers with 30 neurons each.

5.2 Sensitivity Analysis with Respect to Each Input

In this section we study the effect of input variables on the stability of the networks. More specifically, we study the effect of input variations on the stability of the networks by quantifying $\vartheta_m^{\Omega, \mathbb{K}}$ with $\epsilon \in [0, 1]$, for various choices of \mathbb{K} , instead of a global Lipschitz constant accounting for the influence of the whole set of inputs. The obtained value of $\vartheta_m^{\Omega, \mathbb{K}}$ allows us to evaluate how the corresponding group of variables may potentially affect the stability of the network. By performing this analysis for several choices of \mathbb{K} , we thus generate a multivariate analysis of the Lipschitz regularity of the network.

As shown by Proposition 1, varying the ϵ parameter is also insightful since it allows us to measure how the network behaves when input perturbations are gradually more concentrated on a given subset of inputs.

Although our approach can be applied to groups of inputs, for simplicity in this section, we will focus on the case when the sets \mathbb{K} reduce to singletons. In this context, we propose a new representation for displaying the results of the Lipschitz

⁷<https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html>.

⁸We present the results only for ReLU, but we tested our approach with other activation functions such as tanh as well and found the trends in sensitivity of inputs to be similar.

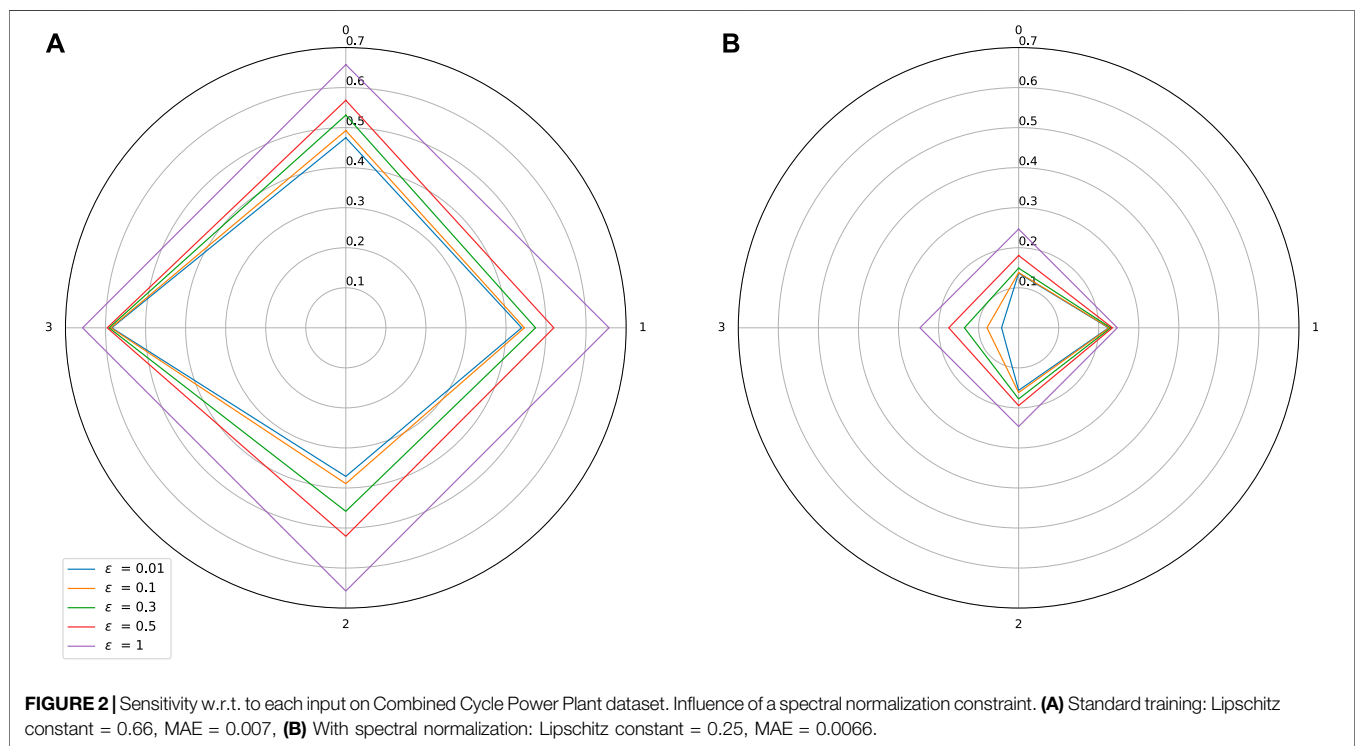
⁵<https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant>.

⁶<https://archive.ics.uci.edu/ml/datasets/auto+mpg>.

TABLE 6 | Network Architecture and training setup for different datasets.

Dataset	Hidden layers and neurons	Epochs	Optimizer	Learning rate
Combined cycle power plant	(10, 6)	100	Adam	0.01
Auto MPG	(16, 8)	1,000	RMSprop	0.001
Boston housing	(10, 5)	500	RMSprop	0.001
Thales air mobility app.	$[10 \times (30)]$	100	Adam	0.01

The input attributes are normalized by removing their mean and scaling to unit variance.



analysis of a neural network. More precisely, we plot the values of $(\vartheta_m^{\Omega_{\epsilon, (k)}})_{1 \leq k \leq N_0}$ on a star or radar chart where each branch of the star corresponds to the index k of an input. For each value of ϵ , a new plot is obtained which is displayed in a specific color. Note that, according to Proposition 3(iv), the plots generated for different ϵ values cannot cross. When $\epsilon = 1$, we obtain an “isotropic” representation whose “radius” corresponds to the global Lipschitz constant ϑ_m of the network. This representation is called a Lipschitz star. All the results of our analysis will be displayed with this representation.

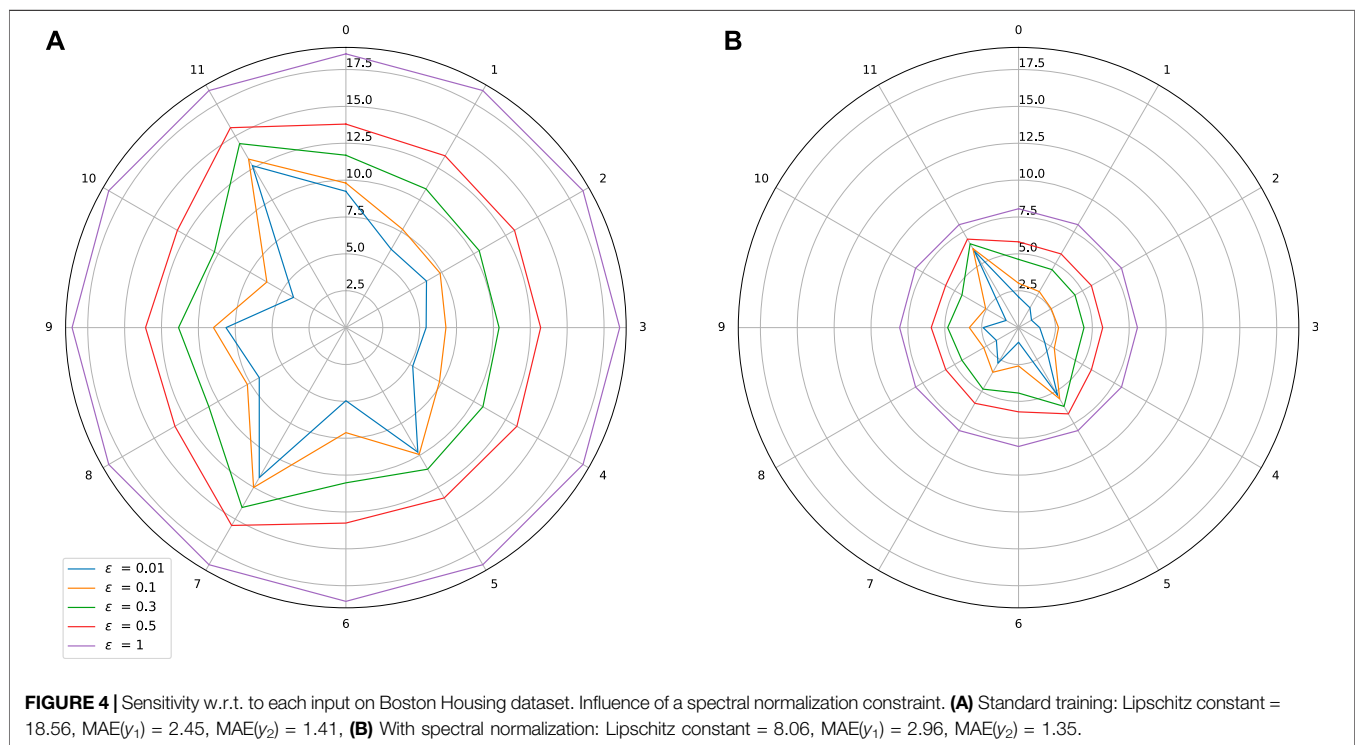
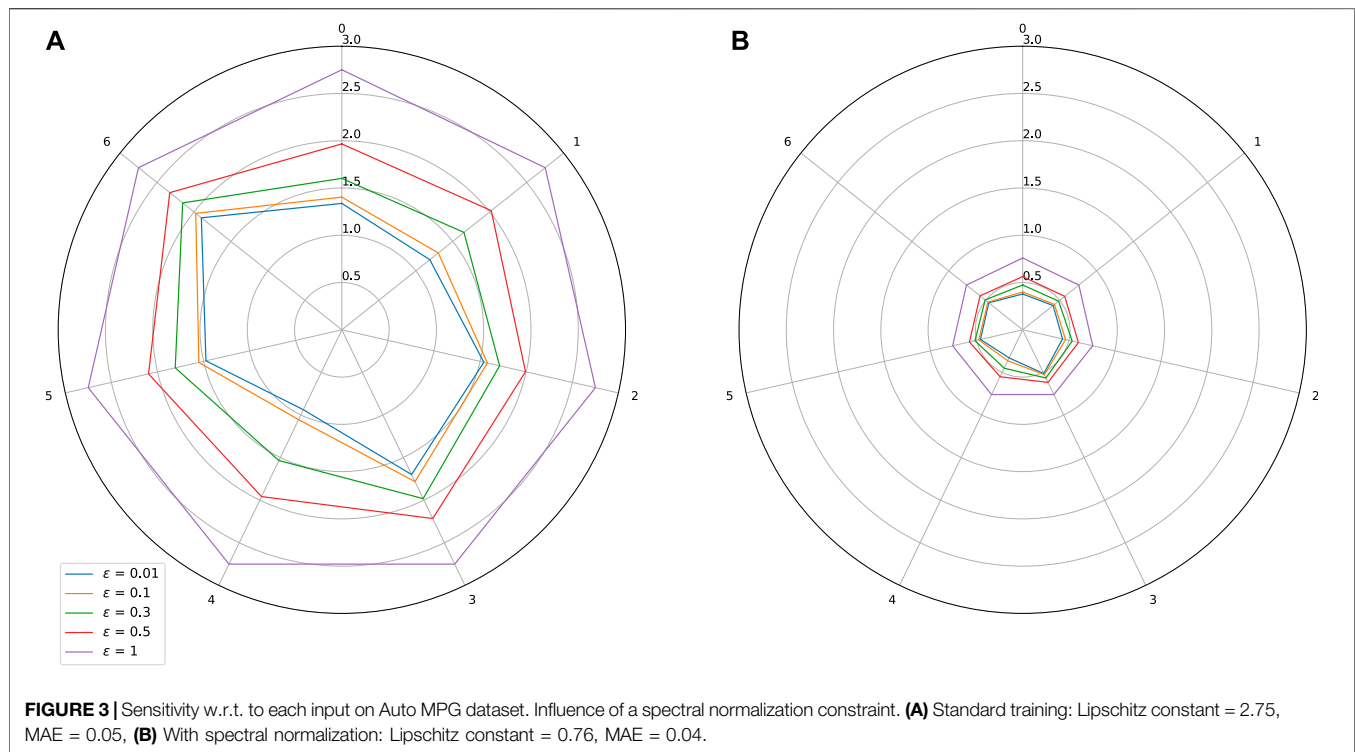
For each dataset, we first perform a standard training when designing the network. To facilitate comparisons, the Lipschitz star of the network trained in such standard manner is presented as the first subplot of all the figures in the paper. Next, we show the variation in terms of input sensitivity, when 1) a Lipschitz target is imposed, and 2) when an adversarial training of the networks is performed. The network architecture remains unchanged, for all our experiments and each dataset, as indicated in Section 5.1. All the Lipschitz constants for each value of ϵ are calculated using LipSDP-Neuron (Fazlyab et al.,

2019). Since an increased stability may come at the price of a loss of accuracy (Tsipras et al., 2018a), we also report the performance of the networks on test datasets in terms of MAE (Mean Absolute Error) for each of the Lipschitz star plot.

5.3 Effect of Training With Specified Lipschitz Target

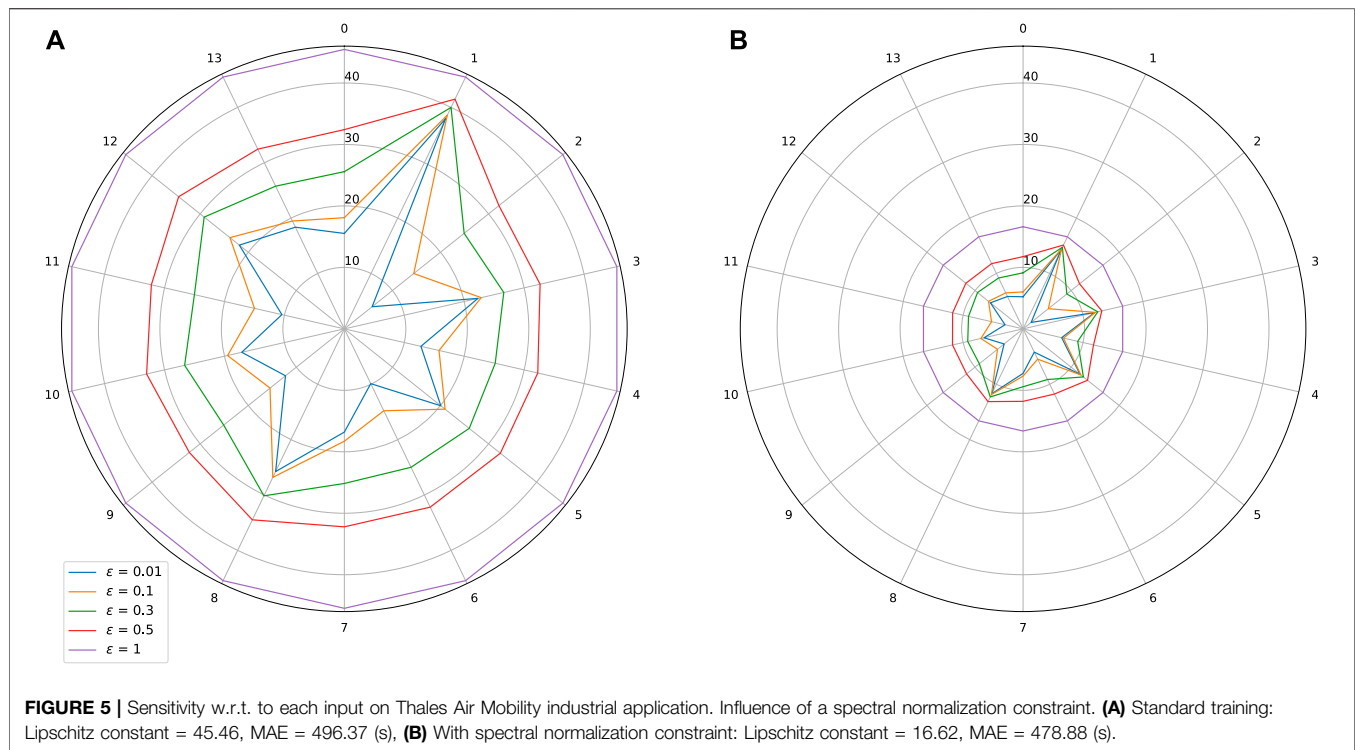
Spectral norm constrained training is performed as explained in Section 4.1. The results are shown for our three datasets in Figures 2–5. On these plots, we can observe a shrinkage of the Lipschitz stars following the reduction of the target Lipschitz value. Interestingly, improving stability does not affect significantly the performance of the networks. Let us comment on the last use case in light of the obtained results.

Comments on the Thales Air Mobility industrial application From the star plots, it is clear that the various variables have a quite different effect on the Lipschitz behavior of the network. This is an expected outcome since these variables carry a different amount of



information captured by learning. From **Figure 5** we observe that variables 1—Flight Distance and 3—Initial ETE play a prominent role, while variables 5—Longitude

Origin, and 8—Longitude Destination are also sensitive. Some plausible explanations for these facts are mentioned below.



- **Flight distance:** The impact of a change of this input can be significant since because of air traffic management separation rules, the commercial aircrafts cannot freely increase their speed to minimize the impact of a longer flight distance.
- **Initial ETE:** Modifying this input is equivalent to changing the initial conditions, which will have a significant impact. It is possible, in the worst case scenario, to accumulate other perturbations coming from other coupled inputs and parameters (e.g., weather conditions) and this is probably the reason why the partial Lipschitz constant is very high, and close to the global Lipschitz constant.
- **Longitude origin and destination parameters:** These parameters are related to different continents and even countries of the origin and destination airports and probably with different qualities of air traffic equipment.

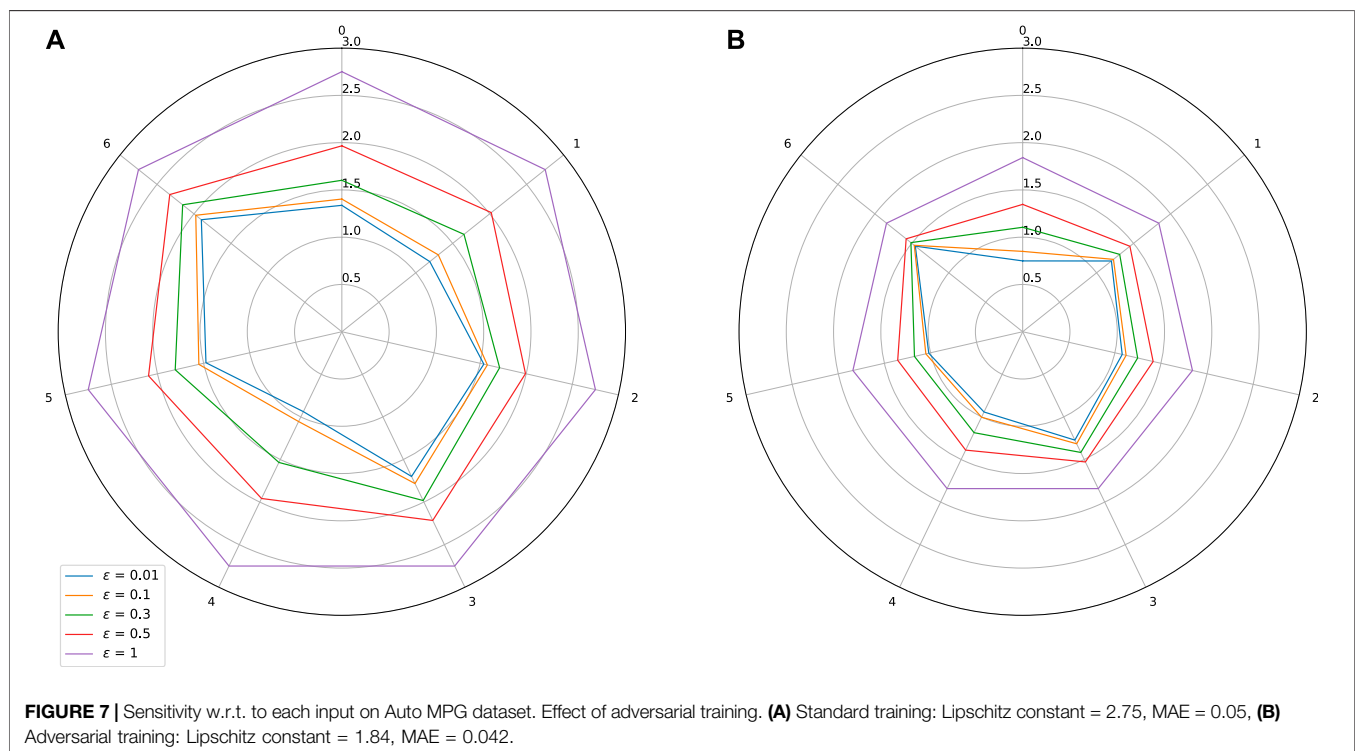
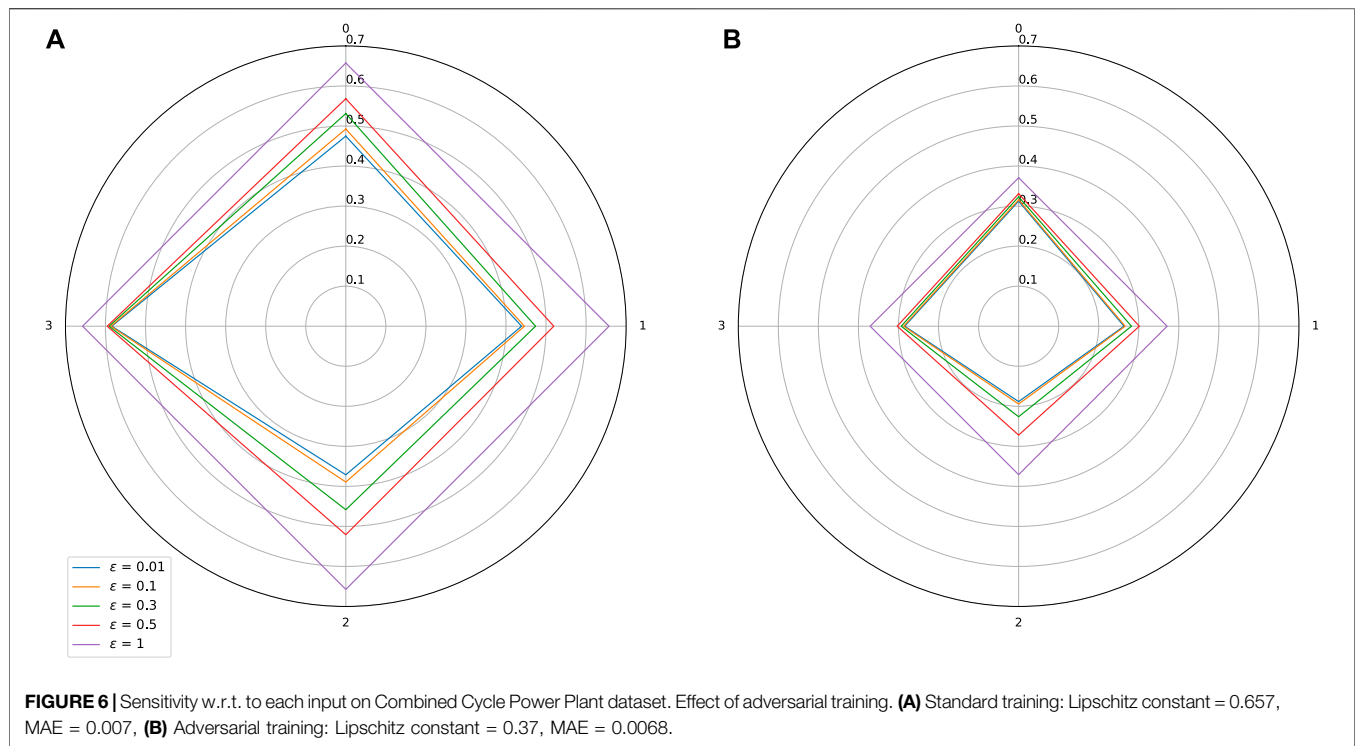
5.4 Effect of Adversarial Training

Generating adversarial attacks and performing adversarial training constitute popular methods in designing robust neural networks. However, these techniques have received less attention for regression tasks, since most of the works deal with classification tasks (Goodfellow et al., 2015; Kurakin et al., 2018; Eykholt et al., 2018). Also, most of the existing works in the deep learning literature are for standard signal/image processing problems, whereas there are only few works handling tabular data (Zhang et al., 2016; Ke et al., 2018). One noticeable exception is (Ballet et al., 2019) which investigates problems related to adversarial attacks for classification tasks involving tabular data. Since our applications are related to regression problems for which few existing works are directly

applicable, we designed a specific adversarial training method. More specifically, for a given amplitude of the adversarial noise and for each sample in the training set, we generate the worst attack based on the spectral properties of the Jacobian of the network, computed by backpropagation at this point. At each epoch of the adversarial training procedure, we solve the underlying minmax problem (Tu et al., 2019). More details on the generation of adversarial attacks for regression attacks can be found in (Gupta et al., 2021).

The generated adversarial attacks from the trained model at the previous epoch are successively concatenated to the training set for the next training epoch, much like in standard adversarial training practices using FGSM (Goodfellow et al., 2015) and Deepfool (Moosavi-Dezfooli et al., 2016) attacks. While generating adversarial attacks on tabular data, some of the variables may be more susceptible to attacks than others. The authors of (Ballet et al., 2019) take care of this aspect by using a feature importance vector. They also only attack the continuous variables, disregarding categorical ones while generating attacks. For the Power plant and Boston Housing datasets, we attack all the four input variables, while on the MPG dataset, we attack only the continuous variables. For the industrial dataset, we generate attacks for the five most sensitive input variables. We also tried attacking all the variables of the dataset but this was not observed to be more efficient. The results in form of Lipschitz star are given in Figures 6–9.

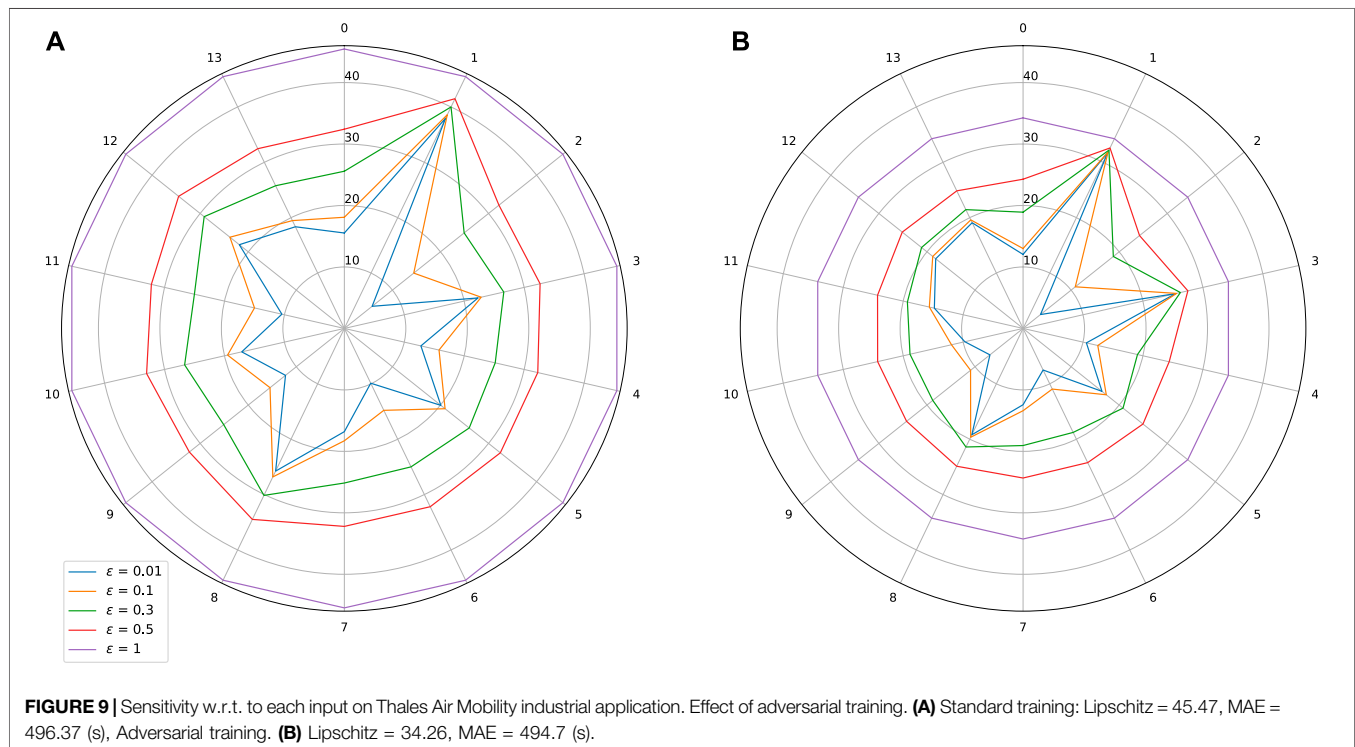
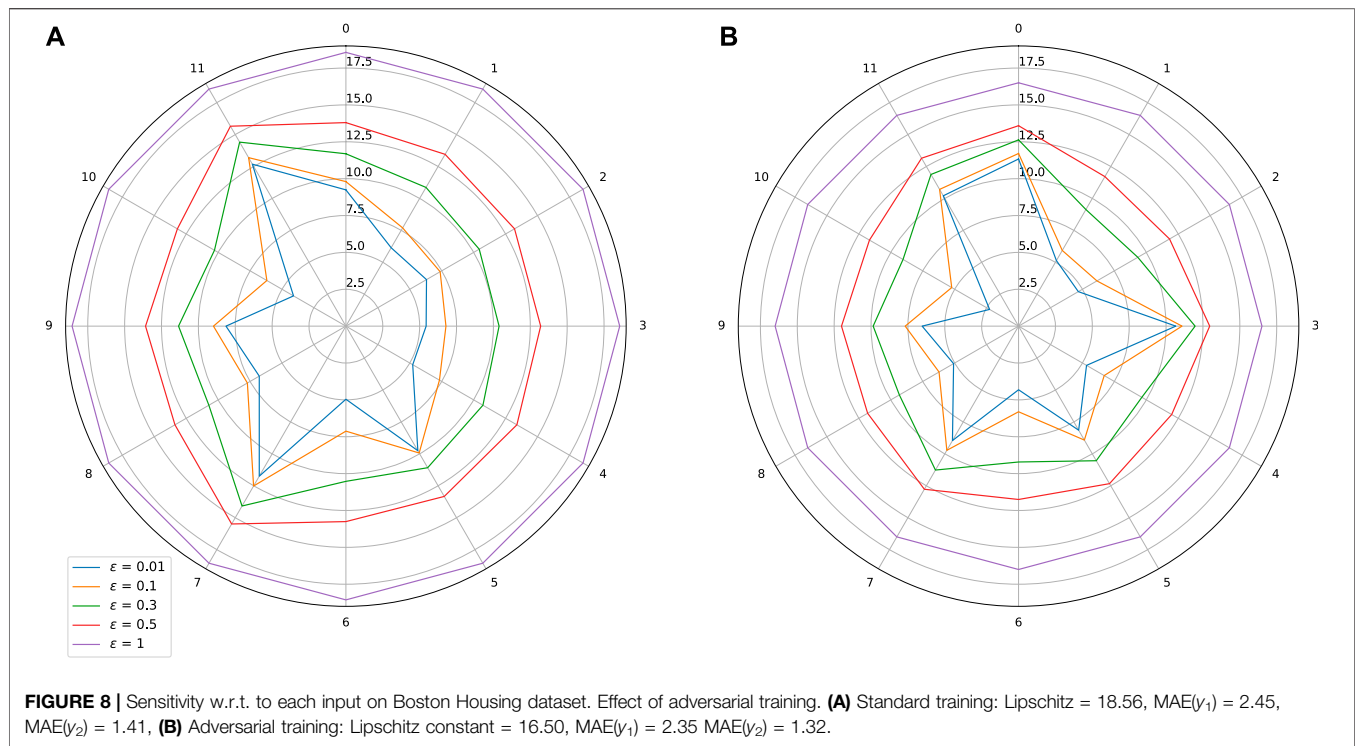
As expected, adversarial training leads to a shrinkage of the star plots, which indicates a better control on the stability of the trained models, while also improving slightly the MAE. In the test we did, we observe however that our adversarial training procedure is globally less efficient than the spectral normalization technique.



5.5 Sensitivity w.r.t. Pair of Variables

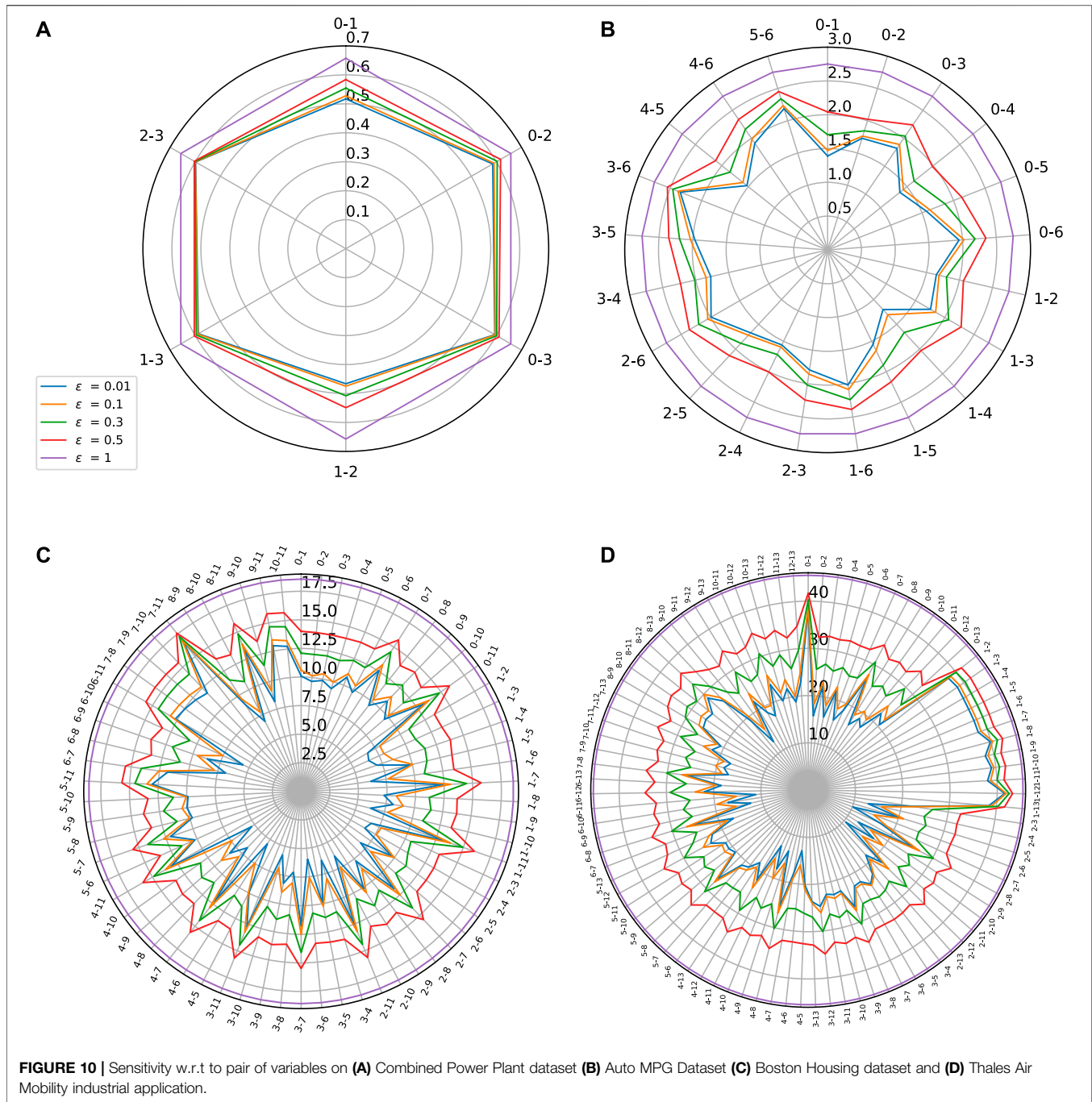
We now consider the case when the set \mathbb{K} contains pairs of elements. We first show the corresponding Partial Lipschitz

constants using a Lipschitz star representation in **Figure 10**, for the different datasets we have discussed in the article. Vertices in the Lipschitz star represent the



obtained Lipschitz constant value $\vartheta_m^{\Omega_{\epsilon, \mathbb{K}}}$ for all possible combinations of pair of variables with varying values of ϵ , i.e., it represents the sensitivity w.r.t. to that particular pair.

As shown by **Figure 10**, this Lipschitz star representation can be useful for displaying the influence of groups of variables instead of single ones. This may be of high interest when the number of inputs is large, especially if they can be grouped into



variables belonging to a given class having a specific physical meaning (e.g., electrical variables versus mechanical ones). Such Lipschitz star representation might however not be very insightful for identifying the coupling that may exist between the variables within a given group. For example, it may happen that, considered together, two variables yield an increased sensitivity than the sensitivity of each of them individually. The reason why we need to find a better way for highlighting these coupling effects is related to Proposition 3(v) which states that, for every $\epsilon \in]0, 1]$ and $(k, \ell) \in \{1, \dots, N_0\}^2$,

$$\max\{\vartheta_m^{\Omega_{\epsilon, [k]}}, \vartheta_m^{\Omega_{\epsilon, [\ell]}}\} \leq \vartheta_m^{\Omega_{\epsilon, [k, \ell]}}. \quad (63)$$

This property means that, when considering a pair of inputs, the one with the highest partial Lipschitz constant will “dominate” the other. To circumvent this difficulty and make our analysis more interpretable, we can think of normalizing the Lipschitz constant in a suitable manner. Such a strategy is a common practice in statistics when, for example, the covariance of a pair of variables is normalized by the product of their standard deviations to define their correlation factor. Once

TABLE 7 | Second order normalized coupling matrix with $\epsilon = 0.001$ on **(A)** Combined Power Plant Dataset **(B)** Auto MPG Dataset **(C)** Boston Housing Dataset and **(D)** Thales Air Mobility industrial application.

Variable	1	2	3
0	0.22	0.57	0.04
1	—	0.15	0.04
2	—	—	0.06

Variable	1	2	3	4	5	6
0	0.1	0.29	0.18	0.06	0.16	0.05
1	—	0.17	0.08	0.03	0.12	0.15
2	—	—	0.14	0.03	0.20	0.11
3	—	—	—	0.11	0.39	0.56
4	—	—	—	—	0.08	0
5	—	—	—	—	—	0.34

Variable	1	2	3	4	5	6	7	8	9	10	11
0	0.22	0.11	0.16	0.03	0.25	0.12	0.17	0.09	0.43	0.11	0.10
1	—	0.17	0.16	0.37	0.00	0.00	0.14	0.18	0	0.05	0.11
2	—	—	0.17	0.05	0	0.23	0	0.35	0.061	0.02	0.00
3	—	—	—	0.35	0.07	0.21	0	0.12	0.02	0.16	0.01
4	—	—	—	—	0.05	0.11	0.11	0.01	0.04	0.06	0.08
5	—	—	—	—	—	0.11	0.07	0.01	0.08	0.04	0.07
6	—	—	—	—	—	—	0.01	0	0.16	0.35	0
7	—	—	—	—	—	—	—	0.27	0.1	0.03	0.76
8	—	—	—	—	—	—	—	—	0.27	0.04	0.14
9	—	—	—	—	—	—	—	—	—	0.02	0.06
10	—	—	—	—	—	—	—	—	—	—	0.01

Variable	1	2	3	4	5	6	7	8	9	10	11	12
0	0.01	0.03	0.01	0.03	0.21	0.03	0.23	0.09	0.21	0.27	0.06	0.28
1	—	0	0.03	0.01	0	0	0	0.12	0.07	0	0.24	0.03
2	—	—	0	0.04	0.01	0.05	0.02	0	0.01	0.02	0	0.01
3	—	—	—	0.01	0.19	0	0.03	0.08	0.06	0.01	0.26	0.15
4	—	—	—	—	0	0.02	0.17	0	0.01	0.06	0.01	0.01
5	—	—	—	—	—	0.06	0.13	0.11	0.13	0.07	0.19	0.27
6	—	—	—	—	—	—	0.03	0.01	0.02	0.19	0.02	0.01
7	—	—	—	—	—	—	—	0.01	0.04	0.09	0.02	0.32
8	—	—	—	—	—	—	—	—	0.07	0.02	0.29	0.03
9	—	—	—	—	—	—	—	—	0.01	0.06	0.02	0.03
10	—	—	—	—	—	—	—	—	—	0	0.21	0.16
11	—	—	—	—	—	—	—	—	—	—	0.00	0.07
12	—	—	—	—	—	—	—	—	—	—	—	0.12

again, we can take advantage of the properties established in Proposition 3 to provide us a guideline to perform this normalization. In addition to Eq. 63, according to Property (viii),

$$\begin{aligned} \mathfrak{g}_m^{\Omega_{\epsilon, [k, \ell]}} &\leq \left(\left(\mathfrak{g}_m^{\Omega_{\epsilon, [k]}} \right)^{p^*} + \left(\mathfrak{g}_m^{\Omega_{\epsilon, [\ell]}} \right)^{p^*} \right)^{1/p^*} + o(\epsilon) \\ &\leq 2^{1/p^*} \max \{ \mathfrak{g}_m^{\Omega_{\epsilon, [k]}}, \mathfrak{g}_m^{\Omega_{\epsilon, [\ell]}} \} + o(\epsilon). \end{aligned} \quad (64)$$

The two previous inequalities suggest to normalize the Lipschitz constant for pairs of inputs by defining

$$\tilde{\mathfrak{g}}_m^{\Omega_{\epsilon, [k, \ell]}} = \frac{1}{2^{1/p^*} - 1} \left(\frac{\mathfrak{g}_m^{\Omega_{\epsilon, [k, \ell]}}}{\max \{ \mathfrak{g}_m^{\Omega_{\epsilon, [k]}}, \mathfrak{g}_m^{\Omega_{\epsilon, [\ell]}} \}} - 1 \right). \quad (65)$$

Indeed, when ϵ is close to zero, Eqs 63–65 show that $\tilde{\mathfrak{g}}_m^{\Omega_{\epsilon, [k, \ell]}} \in [0, 1]$. Note that, for the diagonal terms, $\tilde{\mathfrak{g}}_m^{\Omega_{\epsilon, [k, k]}} = 0$.

The higher $\tilde{\mathfrak{g}}_m^{\Omega_{\epsilon, [k, \ell]}}$, the higher the gain in sensitivity due to the coupling between k and ℓ . The normalized values for the different datasets are reported in Table 7.

5.6 Interpretation of the Results

We summarize some important observations/properties concerning the stability of the NNs which can be drawn from training on different datasets and leveraging the quantitative tools we have proposed in this article.

a) Combined Power Plant Dataset

- “3—Exhaust Vacuum” is the most sensitive variable out of the four variables.
- We observe for any variable coupled with “3” gives a higher partial Lipschitz constant.

- From **Table 7A**, we see that the effect is mostly caused by the sensitivity of “3” and there is no gain when coupled with other variables. Hence, “3” dominates the overall sensitivity of the NN.
 - On the other hand, we observe that, “0” when coupled with “1” and “2” becomes more sensitive as evidenced by the gain in **Table 7A**.
- b) Auto MPG Dataset
- Variable “6—Origin” and “3—Weight” are the most sensitive variables.
 - The values of partial Lipschitz constant peak when the other variables are coupled with “3” or “6”.
 - From **Table 7B**, we see that most of the values coupled with either “3” or “6” are close to zero, except when “3” and “6” are coupled together. Also, we see an exception when “5” is coupled with either “3” or “6”. This suggests that altogether “3”, “5”, and “6” have a higher impact on the stability of the network.
- c) Boston Housing Dataset
- Variable “7—DIS” and “11—LSTAT” are the most sensitive variables.
 - We observe a high partial Lipschitz constant when coupling any variables with “7” or “11”.
 - From **Table 7C**, we see that all the values for both “7” and “11” coupled with other variables are close to zero, except when “7” and “11” are jointly considered. Hence, “7” and “11” dominate the sensitivity of the NN.
 - We observe from the table of normalized values, that “2–9” have a higher impact on the sensitivity of the NN when coupled. Similar observation can be made for pairs “2–8”, “1–4”, “3–4”.
- d) Thales Air Mobility industrial application
- Variable “1—Flight distance”, “3—Initial ETE”, and “8—Longitude Destination” are the most sensitive variables.
 - We see peaks in the partial Lipschitz constant values when these highly sensitive variables are coupled with other variables.
 - But when analyzing the normalized tables, it becomes clear that the gain is mostly due to these sensitive variables.
 - We also observe from **Table 7D**, an increased sensitivity of “0” when coupled with other variables “5”, “7”, “10”, “11”, and “13”.

6 CONCLUSION

We have proposed a new multivariate analysis of the Lipschitz regularity of a neural network. Our approach, whose theoretical foundations are given in **Section 3**, allows the sensitivity with respect to any group of inputs to be

highlighted. We have introduced a new “Lipschitz star” representation which is helpful to display how each input or group of inputs contributes to the global Lipschitz behaviour of a network. The use of these tools has been illustrated on four regression use cases involving tabular data. The improvements brought by two robust training methods (training subject to Lipschitz bounds and adversarial training) have been measured. More generally the proposed methodology is applicable to various machine learning tasks to build “safe-by-design” models where heterogeneous/multimodal/multimodal data can be used.

DATA AVAILABILITY STATEMENT

The industrial dataset presented in this article is not readily available because the dataset is internal to Thales. Further inquiries should be directed to kavya.gupta100@gmail.com. All other datasets are readily available from the following: <https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant>; <https://archive.ics.uci.edu/ml/datasets/auto+mpg>; <https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html>.

AUTHOR CONTRIBUTIONS

KG—Doctoral student handling the processing of the datasets, coding of the tools proposed, optimization of the results and writing of the article. FK—Thales advisor of the doctoral student. Responsible for procuring industrial datasets and technical advises on the experimentation and better utilization of tools, editing of the article. BP—Thales advisor of the doctoral student. Responsible for procuring industrial datasets and technical advises on the experimentation and better utilization of tools, editing of the article. J-CP—Academic advisor responsible mathematical proofs of the work presented in the article and writing of the article. FM—Academic advisor responsible for editing the article.

FUNDING

Doctoral thesis of KG is funded by l’Association Nationale de la Recherche et de la Technologie (ANRT) and Thales LAS France under CIFRE convention. Part of this work was supported by the ANR Research and Teaching Chair in Artificial Intelligence, BRIDGEABLE.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frsip.2022.794469/full#supplementary-material>

REFERENCES

- Ballet, V., Renard, X., Aigrain, J., Laugel, T., Frossard, P., and Detryniecki, M. (2019). Imperceptible Adversarial Attacks on Tabular Data. *NeurIPS 2019 Workshop on Robust AI in Financial Services: Data, Fairness, Explainability, Trustworthiness and Privacy (Robust AI in FS 2019)*. Available at <https://hal.archives-ouvertes.fr/hal-03002526>.
- Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. (2017). "Spectrally-normalized Margin Bounds for Neural Networks," in *Advances in Neural Information Processing Systems*, 6240–6249.
- Chen, T., Lasserre, J.-B., Magron, V., and Pauwels, E. (2020). Semialgebraic Optimization for Lipschitz Constants of ReLU Networks. *Adv. Neural Inf. Process. Syst.* 33, 19189–19200.
- Combettes, P. L., and Pesquet, J.-C. (2020a). Deep Neural Network Structures Solving Variational Inequalities. *Set-Valued Variational Anal.* 28, 1–28. doi:10.1007/s11228-019-00526-z
- Combettes, P. L., and Pesquet, J.-C. (2020b). Lipschitz Certificates for Layered Network Structures Driven by Averaged Activation Operators. *SIAM J. Math. Data Sci.* 2, 529–557. doi:10.1137/19m1272780
- Combettes, P. L., and Pesquet, J.-C. (2008). Proximal Thresholding Algorithm for Minimization over Orthonormal Bases. *SIAM J. Optim.* 18, 1351–1376. doi:10.1137/060669498
- Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., et al. (2018). "Robust Physical-World Attacks on Deep Learning Visual Classification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1625–1634. doi:10.1109/cvpr.2018.00175
- Fazlyab, M., Robey, A., Hassani, H., Morari, M., and Pappas, G. (2019). "Efficient and Accurate Estimation of Lipschitz Constants for Deep Neural Networks," in *Advances in Neural Information Processing Systems*, 11423–11434.
- Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *International Conference on Learning Representations*. Available at <http://arxiv.org/abs/1412.6572>.
- Gupta, K., Pesquet, J.-C., Pesquet-Popescu, B., Malliaros, F., and Kaakai, F. (2021). An Adversarial Attacker for Neural Networks in Regression Problems. *IJCAI Workshop on Artificial Intelligence Safety (AI Safety)*. Available at http://ceur-ws.org/Vol-2916/paper_17.pdf.
- Hancock, J. T., and Khoshgoftaar, T. M. (2020). Survey on Categorical Data for Neural Networks. *J. Big Data* 7, 1–41. doi:10.1186/s40537-020-00305-w
- Katz, G., Barrett, C., Dill, D. L., Julian, K., and Kochenderfer, M. J. (2017). "Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks," in *International Conference on Computer Aided Verification (Springer)*, 97–117. doi:10.1007/978-3-319-63387-9_5
- Ke, G., Zhang, J., Xu, Z., Bian, J., and Liu, T.-Y. (2018). *TabNN: A Universal Neural Network Solution for Tabular Data*.
- Kurakin, A., Goodfellow, I. J., and Bengio, S. (2018). Adversarial Examples in the Physical World. *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC, 99–112.
- Latorre, F., Rolland, P. T. Y., and Cevher, V. (2020). Lipschitz Constant Estimation of Neural Networks via Sparse Polynomial Optimization. *8th International Conference on Learning Representations*.
- Lewis, A. D. (2010). A Top Nine List: Most Popular Induced Matrix Norms. Available at <https://mast.queensu.ca/~andrew/notes/pdf/2010a.pdf>.
- Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral Normalization for Generative Adversarial Networks. *International Conference on Learning Representations*. Available at <https://openreview.net/forum?id=BIQRgziT->.
- Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). "Deepfool: a Simple and Accurate Method to Fool Deep Neural Networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (IEEE), 2574–2582. doi:10.1109/cvpr.2016.282
- Pauli, P., Koch, A., Berberich, J., Kohler, P., and Allgower, F. (2022). Training Robust Neural Networks Using Lipschitz Bounds. *IEEE Control. Syst. Lett.* 6, 121–126. doi:10.1109/LCSYS.2021.3050444
- Serrurier, M., Mamelet, F., González-Sanz, A., Boissin, T., Loubes, J.-M., and del Barrio, E. (2021). "Achieving Robustness in Classification Using Optimal Transport with Hinge Regularization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (IEEE), 505–514. doi:10.1109/cvpr46437.2021.00057
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., et al. (2013). "Intriguing Properties of Neural Networks," in 2nd International Conference on Learning Representations, Banff, AB, April 14–16, 2014. Available at <https://dblp.org/rec/journals/corr/SzegedyZSBEGF13.bib>.
- Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. (2018a). Robustness May Be at Odds with Accuracy. *International Conference on Learning Representations*. Available at <https://openreview.net/forum?id=SyxAb30cY7>.
- Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. (2018b). There Is No Free Lunch in Adversarial Robustness (But There Are Unexpected Benefits). *arXiv preprint arXiv:1805.12152* 2 (3).
- Tu, Z., Zhang, J., and Tao, D. (2019). Theoretical Analysis of Adversarial Learning: A Minimax Approach. *Advances in Neural Information Processing Systems* 32.
- Virmaux, A., and Scaman, K. (2018). "Lipschitz Regularity of Deep Neural Networks: Analysis and Efficient Estimation," in *Advances in Neural Information Processing Systems*, 3835–3844.
- Weng, L., Chen, P.-Y., Nguyen, L., Squillante, M., Boopathy, A., Oseledets, I., et al. (2019). "Proven: Verifying Robustness of Neural Networks with a Probabilistic Approach," in *International Conference on Machine Learning (Long Beach, CA: PMLR)*, 6727–6736.
- Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R. R., and Chaudhuri, K. (2020). A Closer look at Accuracy vs. Robustness. *Advances in Neural Information Processing Systems*. 33, 8588–8601.
- Zhang, W., Du, T., and Wang, J. (2016). "Deep Learning over Multi-Field Categorical Data," in *European Conference on Information Retrieval (Springer)*, 45–57. doi:10.1007/978-3-319-30671-1_4

Conflict of Interest: KG, FK and BP-P were employed by the company Thales LAS France.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Gupta, Kaakai, Pesquet-Popescu, Pesquet and Malliaros. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Optimization of Network Throughput of Joint Radar Communication System Using Stochastic Geometry

Shobha Sundar Ram*, Shubhi Singhal and Gourab Ghatak

Indraprastha Institute of Information Technology Delhi, New Delhi, India

OPEN ACCESS

Edited by:

Monica Bugallo,
Stony Brook University, United States

Reviewed by:

Fabrizio Santi,
Sapienza University of Rome, Italy
Milica Pejanovic-Djurisic,
University of Montenegro,
Montenegro

*Correspondence:

Shobha Sundar Ram
shobha@iitd.ac.in

Specialty section:

This article was submitted to
Radar Signal Processing,
a section of the journal
Frontiers in Signal Processing

Received: 14 December 2021

Accepted: 22 March 2022

Published: 28 April 2022

Citation:

Ram SS, Singhal S and Ghatak G
(2022) Optimization of Network
Throughput of Joint Radar
Communication System Using
Stochastic Geometry.
Front. Sig. Proc. 2:835743.
doi: 10.3389/frsip.2022.835743

Recently joint radar communication (JRC) systems have gained considerable interest for several applications such as vehicular communications, indoor localization and activity recognition, covert military communications, and satellite based remote sensing. In these frameworks, bistatic/passive radar deployments with directional beams explore the angular search space and identify mobile users/radar targets. Subsequently, directional communication links are established with these mobile users. Consequently, JRC parameters such as the time trade-off between the radar exploration and communication service tasks have direct implications on the network throughput. Using tools from stochastic geometry (SG), we derive several system design and planning insights for deploying such networks and demonstrate how efficient radar detection can augment the communication throughput in a JRC system. Specifically, we provide a generalized analytical framework to maximize the network throughput by optimizing JRC parameters such as the exploration/exploitation duty cycle, the radar bandwidth, the transmit power and the pulse repetition interval. The analysis is further extended to monostatic radar conditions, which is a special case in our framework. The theoretical results are experimentally validated through Monte Carlo simulations. Our analysis highlights that for a larger bistatic range, a lower operating bandwidth and a higher duty cycle must be employed to maximize the network throughput. Furthermore, we demonstrate how a reduced success in radar detection due to higher clutter density deteriorates the overall network throughput. Finally, we show a peak reliability of 70% of the JRC link metrics for a single bistatic transceiver configuration.

Keywords: joint radar communication, stochastic geometry, throughput, bistatic radar, explore/exploit

1 INTRODUCTION

Over the last decade, joint radar communication (JRC) frameworks are being researched and developed for numerous applications at microwave and millimeter wave (mmWave) frequencies Liu et al. (2020). Through the integration of sensing and communication functionalities on a common platform, JRC based connected systems offer the advantages of increased spectral efficiency through shared spectrum and reduced hardware costs. The most common applications are WiFi/WLAN based indoor detection of humans Falcone et al. (2012); Storrer et al. (2021); Tan et al. (2016); Li et al. (2020); Alloulah and Huang (2019); Yildirim et al. (2021), radar enhanced vehicular communications Ali et al. (2020); Kumari et al. (2017); Dokhanchi et al. (2019); Duggal et al. (2020), covert communications supported by radar based localization Kellett et al. (2019); Hu et al.

(2019) and radar remote sensing based on global navigation satellite systems (GNSS) Zavorotny et al. (2014). All of these systems consist of a dual functional (radar-communication) transmitter and either a standalone or integrated radar/communications receiver. When the radar receiver is not co-located with the transmitter, the system constitutes a passive/bistatic radar framework. This is the most common scenario in sub-6 GHz indoor localization systems where the WiFi access point serves as both a radar and communication transmitter and humans activities are sensed for intrusion detection, surveillance, or assisted living. The bistatic scenario is also encountered in GNSS based remote sensing where the ground reflected satellite signals are analyzed, at a passive radar receiver, to estimate land and water surface properties Zavorotny et al. (2014). JRC based systems are also being researched for next generation intelligent transportation services where one of the main objectives is to share environment information for collision avoidance, and pedestrian detection eventually leading to autonomous driving. MmWave communication protocols such as IEEE 802.11 ad/ay characterized by high wide bandwidths and low latency have been identified for vehicular-to-everything (V2X) communications Nitsche et al. (2014); Zhou et al. (2018). However, due to the severe propagation loss at mmWave carrier frequencies, they are meant to operate in short range line-of-sight (LOS) conditions with highly directional beams realized through digital beamforming. In high mobility environments, beam training will result in considerable overhead and significant deterioration of latency. Hence, the integration of the radar functionality within the existing millimeter wave communication frameworks is being explored for rapid beam alignment Kumari et al. (2017); Dokhanchi et al. (2019); Duggal et al. (2019); Grossi et al. (2021). The wide bandwidth supported by the mmWave signals along with the channel estimation capabilities within the packet preamble are uniquely suited for radar remote sensing operations. To summarize, we divide the integrated sensing and communication systems into two broad categories. In the first category, the communication transmitter serves as an opportunistic illuminator whose parameters cannot be modified for maximizing a passive radar receiver's detection performance. The second category is where a dual functional system is implemented with optimized design parameters - such as antennas, transmit waveform and signal processing algorithms—for enhanced radar detection performance without deterioration in the communication metrics Hassanien et al. (2016); Mishra et al. (2019); Ma et al. (2021). In this work, we consider the second category and focus on the time resource management between the radar and communication functionalities for maximizing communication network throughput. A preliminary work on the detection metrics of a bistatic radar was presented in Ram and Ghatak (2022). Here, we consider a generalized passive/bistatic radar framework that can be used to model the JRC application scenarios described above and analyze the communication network throughput performance as a function of radar detection metrics. The monostatic radar scenario is considered as a limiting case of the bistatic radar and the corresponding results are obtained as a corollary.

Prior works have tackled the time resource management for multi-functional radars Miranda et al. (2007). In Grossi et al. (2017), the radar dwell time was optimized for maximum target detection for a constant false alarm rate. In Ghatak et al. (2021), the time resource management between the localization and communication functionalities was determined as a function of the density of base station deployment. During the radar/localization phase, the transmitter must scan the angular search space and determine the number and location of the mobile users. Then these users must be served during the remaining duration through directional/pencil beams. The exploration and service process must be repeated periodically due to the motion of the mobile user. Now, if the angular beamwidth of the search beams are very narrow, then they will take longer to cover the search space (for a fixed dwell time) and this will result in reduced communication service time. However, the radar link quality will be higher due to the improved gain and result in a larger number of targets being detected. Hence, the overall network throughput is a function of the explore/exploit time management. In this paper, we use stochastic geometry (SG) based formulations to optimize the network throughput as a function of the explore/exploit duty cycle.

SG tools were originally applied to communication problems in cellular networks, mmWave systems, and vehicular networks Chiu et al. (2013); Andrews et al. (2011); Bai and Heath (2014); Thornburg et al. (2016); Ghatak et al. (2018). In all of these scenarios, there is considerable variation in the strength and spatial distribution of the base stations. More recently, they have been used in diverse radar scenarios to study the radar detection performance under interference and clutter conditions Al-Hourani et al. (2017); Munari et al. (2018); Ren et al. (2018); Park and Heath (2018); Fang et al. (2020). These works have considered the significant diversity in the spatial distributions and density of radars. SG offers a mathematical framework to analyze performance metrics of spatial stochastic processes that approximate to Poisson point process distributions without the requirement of computationally expensive system simulation studies or laborious field measurements. Based on the mathematical analysis, insights are obtained of the impact of design parameters on system level performances. In our problem related to JRC, there can be considerable variation in the position of the dual functional base station transmitter, the radar receiver and the communication end users who are the primary radar targets. Additionally, the JRC will encounter reflections from undesired targets/clutter in the environment. We model the discrete clutter scatterers in the bistatic radar environment as a homogeneous Poisson point process (PPP) similar to Chen et al. (2012); Ram et al. (2020, 2021). This generalized framework allows us to regard each specific JRC deployment, not as an individual case, but as a specific instance of an overall spatial stochastic process. Further, the target parameters such as the position and radar cross-section are also modelled as random variables. Using SG we quantify the mean number of mobile users that can be detected by the radar provided the statistics of the target and clutter conditions are known and subsequently determine the network throughput. Then we use the theorem to optimize system parameters such as the explore/exploit duty

cycle, transmitted power, radar bandwidth and pulse repetition interval for maximum network throughput. Our results are validated through Monte Carlo simulations carried out in the short range bistatic radar framework.

Our paper is organized as follows. In the following section, we present the system model of the JRC with the bistatic radar framework and describe the explore/exploit time management scheme. In **section 4**, we provide the theorem for deriving the network throughput as a function of the bistatic radar parameters. In **section 5**, we offer the key system parameter insights that are drawn from the theorem as well as the Monte Carlo simulation based experimental validation. Finally, we conclude the paper with a discussion on the strengths and limitations of the proposed analytical framework along with directions for future work.

Notation: In this paper, all the random variables are indicated with bold font and constants and realizations of a random variable are indicated with regular font.

2 SYSTEM MODEL

We consider a joint radar-communication (JRC) framework with a single base station (BS), multiple mobile users (MU) and a single passive radar receiver (RX) as shown in **Figure 1A**. The BS serves as a dual functional transmitter that supports both radar and communication functionalities in a time division manner as shown in **Figure 1B**. During the T_{search} interval, the BS serves as the radar transmitter or opportunistic illuminator and along with the RX, forms a bistatic radar whose objective is to localize the multiple MU in the presence of clutter/undesirable targets. During this interval, the BS transmits a uniform pulse stream of τ pulse width and T_{PRI} pulse repetition interval, through a directional and reconfigurable antenna system with gain G_{tx} and beamwidth $\Delta\theta_{tx}$. The radar must scan the entire angular search space within T_{search} to find the maximum number of MU. If the duration of an antenna beam is fixed at T_{beam} (based on hardware parameters such as circuit switching speed for electronic scanning or Doppler frequency resolution requirements), then the number of beams that can be searched within T_{search} is given by

$$n_{beam} = \frac{\Omega}{\Delta\theta_{tx}} = \frac{T_{search}}{T_{beam}}, \quad (1)$$

where Ω is the angular search space. In our problem formulation, we set $\Omega = 2\pi$ to correspond to the entire azimuth angle extent. During the remaining duration of T_{serve} , directional communication links are assumed to be established between the BS and the detected MUs. Thus the beam alignment for communication during T_{serve} is based on radar enabled localization during T_{search} . Since the position of the MU does not remain fixed with time, the process of beam alignment is repeated for every $T = T_{search} + T_{serve}$ as shown in the figure. An important tuning parameter in the above JRC framework is the duty cycle $\epsilon = \frac{T_{search}}{T}$. From (1), it is evident that $\Delta\theta_{tx} = \frac{\Omega T_{beam}}{\epsilon T} = \frac{1}{B_0 \epsilon}$. Here, B_0 is a constant and equal to $\frac{T}{\Omega T_{beam}}$.

Note that when the beams become broader, the gain of the radar links become poorer. As a result of the deterioration in the radar link metrics due to larger $\Delta\theta_{tx}$, the detection performance becomes poorer and fewer MU (η) are likely to be detected in the search space. Thus η is directly proportional to ϵ . On the other hand, the network throughput (Y) of the system is defined as

$$Y = \eta(\epsilon)(1 - \epsilon)D, \quad (2)$$

where $(1 - \epsilon)$ is the duty cycle of the communication service time ($\frac{T_{search}}{T}$). Here, we assume that the communication resources such as spectrum are available to all the η detected MU and all the MU are characterized by identical data rates D . The objective of our work is to present a theorem to optimize the duty cycle ϵ for maximum Y under the assumption that the noise, MU and clutter statistics are known and fixed during the radar processing time. These conditions are generally met for microwave or millimeter-wave systems Billingsley (2002); Ruuskanen et al. (2003). The theoretical framework is derived for a generalized bistatic JRC framework where inferences for monostatic conditions are derived from limiting conditions.

Next, we discuss the planar bistatic radar geometry that we have considered based on the north-referenced system described in Jackson (1986). We assume that the BS is located in the Cartesian coordinates $(-\frac{L}{2}, 0)$ while the passive receiver, RX, is assumed to be omnidirectional and located at $(+\frac{L}{2}, 0)$. High gain transmission links from the BS support high quality communication link metrics. The gain of the passive RX antenna is intentionally kept low so that the common search space of the bistatic radar transmitter and receiver does not become too narrow which would then have to be supported by very time consuming and complicated beam scanning operations. Note that the geometry considered here is specifically suited to model short surveillance based JRC systems (such as indoor/outdoor wireless communication systems). It does not model the bistatic GNSS-R scenario where both the transmitter and receiver are characterized by high gain antennas; and a three-dimensional geometry would have to be considered. The baseline length between the bistatic radar transmitter and receiver is L . The two-dimensional space is assumed to be populated by multiple scatterers - some MU (m) and the remaining discrete clutter (c) scatterers. In real world conditions, there can be significant variation in the number and spatial distribution of the point scatterers (both MU and clutter) in the radar channel. Further, the positions of scatterers are independent of each other. The Poisson point process is a *completely random* process since it has the property that each point is stochastically independent to all the other points in the process. Consequently, we consider the distribution of scatterers as an independent Poisson point processes (PPP: Φ)—wherein each instance is assumed to be a realization (ϕ) of a spatial stochastic process. We specifically consider a homogeneous PPP wherein the number of the scatterers in each realization follows a Poisson distribution and the positions of these scatterers follow a uniform distribution. Some prior works where discrete scatterers have been modelled as a PPP are Chen et al. (2012); Ram et al. (2020, 2021). We assume that the mean spatial densities of the

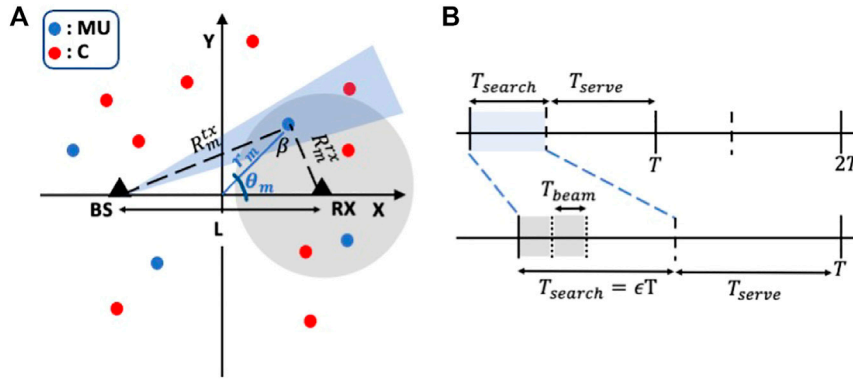


FIGURE 1 | (A) Illustration of the joint radar-communication (JRC) scenario. The base station (BS) at $(\frac{L}{2}, 0)$ and indicated by a triangle is a dual functional transmitter that supports both radar and comm. functionalities with a directional and reconfigurable antenna system of $\Delta\theta_{tx}$ beamwidth. An omnidirectional receiver (RX) at $(-\frac{L}{2}, 0)$ forms the passive/bistatic radar receiver. The channel consists of mobile users (MU) at (r_m, θ_m) at distances, R_m^{tx} and R_m^{rx} , from BS and MU respectively indicated by blue dots; and undesirable clutter scatterers indicated by red dots. The bistatic radar angle is β . **(B)** Timing diagram of the JRC framework where each T consists of $T_{search} = \epsilon T$ when the BS scans the angular search space for MU using n_{beam} of T_{beam} duration. During the remaining T_{serve} duration, directional beam links are established between BS and MU based on the localization by the radar during T_{search} .

MU and clutter scatterers are ρ_m and ρ_c respectively where $\rho_m \ll \rho_c$. The position of an MU/clutter scatterer is specified in polar coordinates (r_i, θ_i) , $i \in m, c$ where r_i is the distance from the origin and θ_i is the angle with the positive X axis. The distance from BS and RX are R_i^{tx} and R_i^{rx} respectively and the bistatic range (κ_i) is specified by the geometric mean of both the one-way propagation distances ($\kappa_i = \sqrt{R_i^{tx} R_i^{rx}}$). In bistatic radar geometry, the contours of constant κ_i for a fixed L are called Cassini ovals Willis (2005). Two regions are identified: the first is the *cosite* region when $L \leq 2\kappa_i$ and the contours appear as concentric ovals for different κ_i ; and the second is when $L > 2\kappa_i$ and the oval splits into two circles centered around BS and RX. In our work, we assume that cosite conditions prevail and that the bistatic angle at MU is β . Note that when L is zero, $\beta = 0$ and the system becomes a monostatic radar scenario. Here, the Cassini ovals become concentric circles for different values of $R_i^{tx} = R_i^{rx} = \kappa_i$.

Classically, radar detection metrics and the radar operating curve are obtained from binary hypothesis testing derived from the Neyman-Pearson (NP) theorem Kay (1998). The probability of detection, \mathcal{P}_d , is the probability that a radar received signal (along with noise and clutter) is above a predefined threshold while the probability of false alarm, \mathcal{P}_{fa} , is the probability that the noise and clutter are above the threshold. For a fixed \mathcal{P}_{fa} , the \mathcal{P}_d is directly proportional to the SCNR. For very simple scenarios (pulse radars in the absence of clutter), the relationship between \mathcal{P}_d , \mathcal{P}_{fa} and SCNR are given by the Albersheim's equation Skolnik (1980) while in more complex scenarios, the relationships have to be derived from extensive measurements. In Ram and Ghatak (2022), we presented a metric called the radar detection coverage probability (\mathcal{P}_{DC}^{Bi}) to indicate the likelihood of a radar target being detected by a bistatic radar based on the signal-to-clutter-and-noise ratio (SCNR). The metric is analogous to wireless detection coverage probability which is widely studied in communication systems to study the network coverage in wireless links Andrews et al. (2011). We prefer the \mathcal{P}_{DC}^{Bi} metric is to \mathcal{P}_d and \mathcal{P}_{fa} since it offers physics based insights into system

performance and because of its tractable problem formulation. Specifically, we use \mathcal{P}_{DC}^{Bi} to estimate the mean number of detected MU (η) as a function of ϵ and optimize the network throughput (Υ). An extended discussion on the derivations of \mathcal{P}_d and \mathcal{P}_{fa} metrics are provided in the appendix of previous work on monostatic radar in Ram et al. (2021). If the transmitted power from BS is P_{tx} and the bistatic radar cross-section (RCS) of the MU, σ_m , is a random variable, then the received signal at RX, S , is given by the Friis radar range equation as

$$S(\kappa_m) = P_{tx} G_{tx}(\theta_m) \sigma_m \mathcal{H}(\kappa_m), \quad (3)$$

where $\mathcal{H}(\kappa_m)$ is the two-way propagation factor. In line-of-sight (LOS) conditions this is

$$\mathcal{H}(\kappa_m) = \frac{\lambda^2}{(4\pi)^3 (R^{tx} R^{rx})^2} = \frac{H_0}{\kappa_m^4}, \quad (4)$$

where λ is the wavelength of the radar. In the above expression, the gain of RX is assumed to be 1 since it is an omnidirectional antenna. We assume that the gain of the BS is uniform within the main lobe and is inversely proportional to the beam width: $G_{tx} = \frac{G_0}{\Delta\theta_{tx}}$ where G_0 is the constant of proportionality that accounts for antenna inefficiencies including impedance mismatch, dielectric and conductor efficiencies. If we assume that the MU is within the mainlobe of the radar, then using (1), **Equation 3** can be written as

$$S(\kappa_m) = \frac{P_{tx} G_0 \sigma_m \mathcal{H}(\kappa_m)}{\Delta\theta_{tx}} = P_{tx} G_0 B_0 \epsilon \sigma_m \mathcal{H}(\kappa_m). \quad (5)$$

In (3) and (5), we have assumed that only a single MU is within a radar resolution cell, A_c . In the real world, a single radar resolution cell may consist of one or more targets. However, there is no way for the radar operator to distinguish or count the targets that are within a single cell. Hence, it will always be counted/considered as a single target. The amplitude of the target signal will however fluctuate due to interference from the

multiple points within the cell and this fluctuation is captured with the Swerling models. Further, in the above discussion, we assume a single tone pulse radar of bandwidth BW . However, the system insights can be equally applied to other wide bandwidth signals as well. The clutter returns, C , at the radar receiver is given by

$$C(\kappa_m) = \sum_{c \in \Phi \cap A_c(\kappa_m)} P_{tx} G_{tx}(\theta_c) \sigma_c \mathcal{H}(\kappa_c). \quad (6)$$

In the above expression, we specifically only consider those clutter scatterers that fall within the same resolution cell, A_c , as the MU. We use the generalized Weibull model Sekine et al. (1990) to describe the distribution of the RCS (σ_c) of the clutter points. For a given noise of the radar receiver, $N_s = K_B T_s BW$ where K_B , T_s and BW are the Boltzmann constant, system noise temperature and bandwidth respectively, the signal to clutter and noise ratio is given by $SCNR(\kappa_m) = \frac{S(\kappa_m)}{C(\kappa_m) + N_s}$.

3 ESTIMATION OF NETWORK THROUGHPUT OF JOINT RADAR-COMMUNICATION

In this section we present the analytical framework to estimate the network throughput of the communication framework as a function of the explore/exploit duty cycle (ϵ). We use the \mathcal{P}_{DC}^{Bi} metric defined in Ram and Ghatak (2022) to estimate, η , the number of MU detected by the radar during the search interval $T_{search} = \epsilon T$ that will be subsequently served during T_{serve} . Theorem 1. The network throughput (Υ) for an explore/exploit duty cycle (ϵ) for a passive/bistatic radar based JRC system is given by

$$\Upsilon = \mathcal{P}_{DC}^{Bi} \left(2\pi\kappa_m - \frac{3\pi L^2}{8\kappa_m} \right) \frac{\rho_m c \tau}{2\sqrt{1 - \frac{L^2}{4\kappa_m^2}}} (1 - \epsilon) D \quad (7)$$

where

$$\mathcal{P}_{DC}^{Bi} = \exp \left(\frac{-\gamma N_s \kappa_m^4}{\sigma_{mavg} P_{tx} G_0 B_0 \epsilon H_0} + \frac{-\gamma \rho_c c \tau \kappa_m^2 \sigma_{cavg}}{B_0 \epsilon (\kappa_m + \sqrt{\kappa_m^2 - L^2}) (\sigma_{mavg} + \gamma \sigma_{cavg})} \right) \quad (8)$$

Proof. For an MU at bistatic range κ_m , the $SCNR$ is a function of several random variables such as the MU cross-section, the position of MU, the number and spatial distribution of the discrete clutter scatterers and their RCS as shown below

$$\begin{aligned} SCNR(\kappa_m) &= \frac{P_{tx} G_0 B_0 \epsilon \sigma_m \mathcal{H}(\kappa_m)}{\sum_{c \in \Phi \cap A_c(\kappa_m)} P_{tx} G_0 B_0 \epsilon \sigma_c \mathcal{H}(\kappa_c) + N_s} \\ &= \frac{\sigma_m}{\sum_{c \in \Phi \cap A_c(\kappa_m)} \frac{\sigma_c \mathcal{H}(\kappa_c)}{\mathcal{H}(\kappa_m)} + \frac{N_s}{P_{tx} G_0 B_0 \epsilon \mathcal{H}(\kappa_m)}}. \end{aligned} \quad (9)$$

We define the bistatic radar detection coverage probability (\mathcal{P}_{DC}^{Bi}) as the probability that the $SCNR$ is above a predefined threshold, γ . Therefore,

$$\begin{aligned} \mathcal{P}_{DC}^{Bi} &= \mathcal{P}(SCNR(\kappa_m) \geq \gamma) \\ &= \mathcal{P} \left(\sigma_m \geq \sum_{c \in \Phi \cap A_c(\kappa_m)} \frac{\gamma \sigma_c \kappa_m^4}{\kappa_c^4} + \frac{\gamma N_s \kappa_m^4}{P_{tx} G_0 B_0 \epsilon H_0} \right). \end{aligned} \quad (10)$$

The bistatic RCS, σ_m , has been shown to demonstrate similar statistics as monostatic RCS Skolnik (1961). In this work, we consider the MU to have Swerling-1 characteristics, which corresponds to mobile users such as vehicles and humans Raynal et al. (2011a,b), as shown below

$$\mathcal{P}(\sigma_m) = \frac{1}{\sigma_{mavg}} \exp \left(-\frac{\sigma_m}{\sigma_{mavg}} \right), \quad (11)$$

where, σ_{mavg} is the average radar cross-section. Hence, (10) can be expanded to

$$\begin{aligned} \mathcal{P}_{DC}^{Bi} &= \exp \left(\sum_{c \in \Phi \cap A_c(\kappa_m)} \frac{-\gamma \sigma_c}{\sigma_{mavg}} - \frac{\gamma N_s \kappa_m^4}{\sigma_{mavg} P_{tx} G_0 B_0 \epsilon H_0} \right) \\ &= \exp \left(\frac{-\gamma N_s \kappa_m^4}{\sigma_{mavg} P_{tx} G_0 B_0 \epsilon H_0} \right) I(\kappa_m). \end{aligned} \quad (12)$$

In the above expression, \mathcal{P}_{DC}^{Bi} consists of two terms. The first term consists entirely of constants and demonstrates the radar detection performance as a function of the signal-to-noise ratio (SNR). The second term, $I(\kappa_m)$, shows the effect of the signal-to-clutter ratio (SCR). Since, we are specifically considering the clutter points that fall within the same resolution cell, A_c , as the MU we can assume that $\mathcal{H}(\kappa_c) \approx \mathcal{H}(\kappa_m)$ in (10). We provide further insights into this path loss approximation in our later sections. Finally, the exponent of sum of terms can be written as a product of exponents. Hence, $I(\kappa_m)$ is

$$I(\kappa_m) = \mathbb{E}_{\sigma_c, c} \left[\prod_{c \in \Phi \cap A_c(\kappa_m)} \exp \left(\frac{-\gamma \sigma_c}{\sigma_{mavg}} \right) \right], \quad (13)$$

where \mathbb{E} is the expectation operator with respect to the clutter scatterers and their corresponding cross-section. The probability generating functional (PGFL) of a homogeneous PPP Haenggi (2012) based on stochastic geometry formulations is given as

$$I = \exp \left(-\mathbb{E}_{\sigma_c, c} \left[\iint_{\mathbf{r}_c, \phi_c} \rho_c \left(1 - \exp \left(\frac{-\gamma \sigma_c}{\sigma_{mavg}} \right) \right) d(\vec{x}_c) \right] \right), \quad (14)$$

where ρ_c is the mean spatial density of the clutter scatterers. The integral specifically considers the clutter scatterers that fall within the same resolution cell as the MU. Bistatic radar literature identifies three types of resolution cells—the range resolution cell, the beamwidth resolution cell and the Doppler resolution cell. In our study, the main objective of the radar is to perform range-azimuth based localization. Hence, we consider the range resolution cell, which based on Willis (2005); Moyer et al. (1989), corresponds to

$$A_c(\kappa_m) = \frac{c \tau R^{tx}(\theta_m) \Delta \theta_{tx}}{2 \cos^2(\beta(\theta_m)/2)} = \frac{c \tau R^{tx}(\theta_m)}{B_0 \epsilon \left(1 + \sqrt{1 - \sin^2 \beta(\theta_m)} \right)}, \quad (15)$$

for a pulse width of τ . In the above expression, note that the size of A_c varies as a function of constant κ_m and the random variable θ_m . Prior studies show that $\sin \beta$ takes on the value of $\sin \beta_{max}$ with a very high probability when $R_m^{tx} \approx \kappa_m$ Ram and Ghatak (2022). Based on bistatic geometry $\sin \beta_{max} = \sqrt{\frac{L^2}{\kappa_m^2} - \frac{L^4}{\kappa_m^4}} \approx \frac{L}{\kappa_m}$ when $\kappa_m > L$. Therefore, (15) reduces to

$$A_c \approx \frac{c\tau\kappa_m^2}{B_0\epsilon(\kappa_m + \sqrt{\kappa_m^2 - L^2})} \quad (16)$$

If we assume that the clutter statistics are uniform within A_c , then the integral in (14) can be further reduced to

$$\begin{aligned} I &= \exp\left(-\mathbb{E}_{\sigma_c}\left[\left(1 - \exp\left(\frac{-\gamma\sigma_c}{\sigma_{mavg}}\right)\right)\rho_c A_c\right]\right) \\ &= \exp\left(-\mathbb{E}_{\sigma_c}\left[\left(1 - \exp\left(\frac{-\gamma\sigma_c}{\sigma_{mavg}}\right)\right)\frac{\rho_c c\tau\kappa_m^2}{B_0\epsilon(\kappa_m + \sqrt{\kappa_m^2 - L^2})}\right]\right) \end{aligned} \quad (17)$$

If we define $J(\kappa_m) = \frac{\rho_c c\tau\kappa_m^2}{B_0\epsilon(\kappa_m + \sqrt{\kappa_m^2 - L^2})}$ as a constant independent of σ_c , then it can be pulled out of the integral for computing the expectation as shown below

$$I(\kappa_m) = \exp\left(-J(\kappa_m) \int_0^\infty \left(1 - \exp\left(\frac{-\gamma\sigma_c}{\sigma_{mavg}}\right)\right) \mathcal{P}(\sigma_c) d\sigma_c\right). \quad (18)$$

In our work, we specifically consider the contributions from discrete/point clutter responses that arise from direct and multipath reflections from the surrounding environment. We model the radar cross-section of these scatterers using the generalized Weibull model shown in

$$\mathcal{P}(\sigma_c) = \frac{\alpha}{\sigma_{cavg}} \left(\frac{\sigma_c}{\sigma_{cavg}}\right)^{\alpha-1} \exp\left(-\left(\frac{\sigma_c}{\sigma_{cavg}}\right)^\alpha\right), \quad (19)$$

where σ_{cavg} is the average bistatic radar cross-section and α is the corresponding shape parameter. The Weibull distribution has been widely used to model clutter due to its tractable formulation and its adaptability to different environment conditions Sekine et al. (1990). When the scenario is characterized by few dominant scatterers, α is near one and corresponds to the exponential distribution. On the other hand, when there are multiple scatterers of similar strengths, then α tends to two which corresponds to the Rayleigh distribution. The actual value of α in any real world scenario is determined through empirical studies. $I(\kappa_m)$ in (18) can be numerically evaluated for any value of α . But for $\alpha = 1$, the expression becomes

$$I(\kappa_m) = \exp\left(-\frac{\gamma J(\kappa_m) \sigma_{cavg}}{\sigma_{mavg} + \gamma \sigma_{cavg}}\right). \quad (20)$$

Substituting (20) in (12), we obtain

$$\mathcal{P}_{DC}^{Bi} = \exp\left(\frac{-\gamma N_s \kappa_m^4}{\sigma_{mavg} P_{tx} G_0 B_0 \epsilon H_0} + \frac{-\gamma \rho_c c\tau \kappa_m^2 \sigma_{cavg}}{B_0 \epsilon (\kappa_m + \sqrt{\kappa_m^2 - L^2}) (\sigma_{mavg} + \gamma \sigma_{cavg})}\right). \quad (21)$$

The above expression shows the probability that a MU at κ_m is detected by the bistatic radar based on its SCNR. If we assume a uniform spatial distribution, ρ_m , of the MU in Cartesian space, then the mean number of MU that can be detected within the

total radar field-of-view at κ_m bistatic range from the radar will be given by

$$\eta = \mathcal{P}_{DC}^{Bi}(\kappa_m) \rho_m \mathcal{C}(\kappa_m) \delta r, \quad (22)$$

where $\mathcal{C}(\kappa_m)$ is the circumference of a Cassini oval and $\delta r = \frac{c\tau}{2\cos(\beta/2)}$ is the range resolution of the radar. The parametric equation for the Cassini oval is given in

$$\left(r_m^2 + \frac{L^2}{4}\right)^2 - r_m^2 L^2 \cos^2 \theta_m = \kappa_m^4. \quad (23)$$

Hence, the circumference $\mathcal{C}(\kappa_m)$ can be computed from

$$\begin{aligned} \mathcal{C}(\kappa_m) &= \int_0^{2\pi} r_m(\theta_m) d\theta_m \\ &= \frac{L}{2} \int_0^{2\pi} \left[\cos 2\theta_m \pm \left(\frac{16\kappa_m^4}{L^4} - \sin^2 \theta_m\right)^{1/2} \right]^{1/2} d\theta_m \approx 2\pi\kappa_m \\ &\quad - \frac{3\pi L^2}{8\kappa_m}. \end{aligned} \quad (24)$$

When $\kappa_m > L$, the estimation of (24) can be approximated to the expression shown above. Note that for very large values of $\kappa_m \gg L$, the scenario approaches monostatic conditions. Here, the oval approximates to a circle of circumference $2\pi\kappa_m$. Also, as mentioned before β can be approximated to β_{max} . Hence $\cos(\beta_{max}/2) \approx \sqrt{1 - \frac{L^2}{4\kappa_m^2}}$. Therefore, the mean number of detected MU is

$$\eta = \mathcal{P}_{DC}^{Bi} \left(2\pi\kappa_m - \frac{3\pi L^2}{8\kappa_m}\right) \frac{\rho_m c\tau}{2\sqrt{1 - \frac{L^2}{4\kappa_m^2}}}, \quad (25)$$

and the resulting network throughput for the communication links that are set up with detected MUs is

$$\Upsilon = \mathcal{P}_{DC}^{Bi} \left(2\pi\kappa_m - \frac{3\pi L^2}{8\kappa_m}\right) \frac{\rho_m c\tau}{2\sqrt{1 - \frac{L^2}{4\kappa_m^2}}} (1 - \epsilon) D. \quad (26)$$

4 OPTIMIZATION OF JOINT RADAR-COMMUNICATION SYSTEM PARAMETERS FOR MAXIMIZATION OF NETWORK THROUGHPUT

In this section, we discuss the corollaries from the theorem presented in the previous section. Based on these inferences, we present how JRC parameters such as ϵ , τ , $\Delta\theta_{tx}$ and T_{PRI} can be optimized for maximum throughput. The results presented in this section are experimentally validated using Monte Carlo simulations. For the simulations, we assume that the bistatic radar transmitter (BS) and receiver (RX) are located at $(\pm \frac{L}{2}, 0)$ respectively as shown in Figure 2. We consider a $[200\text{ m} \times 200\text{ m}]$ region of interest. Radar, MU and clutter parameters such as P_{tx} , L , $\Delta\theta_{tx}$, N_s , σ_{mavg} , κ , σ_{cavg} and ρ_c are kept fixed and summarized in Table.1. In each realization of the Monte Carlo simulation, the

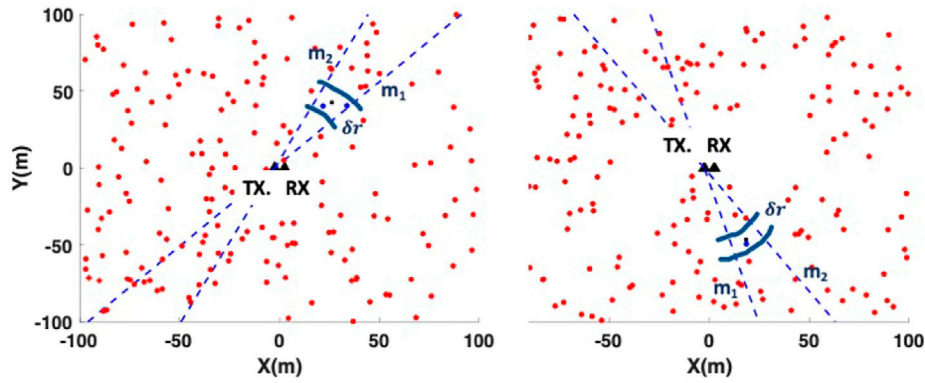


FIGURE 2 | Two realizations of Monte Carlo simulations with bistatic radar transmitter (BS) and receiver (RX) indicated by triangles. The BS is characterized by narrow beam indicated by dashed blue lines with slopes m_1 and m_2 while RX is omnidirectional. Target is indicated by black dot while clutter scatterers inside and outside the radar resolution cell are indicated by blue and red dots respectively.

MU's polar coordinate position, θ_m is drawn from a uniform distribution from $(0, 2\pi)$ and r_m is computed for a fixed κ_m . The RCS of the MU is drawn from the exponential distribution corresponding to the Swerling-1 model. The mean number of discrete clutter scatterers is equal to ρ_c times the area of the region of interest. The number of clutter scatterers are different for each realization and drawn from a Poisson distribution. The positions of the clutter scatterers are based on a uniform distribution in the two-dimensional Cartesian space while the RCS of each discrete scatterer is drawn from the Weibull model. We compute the SCNR based on the returns from the MU and the clutter scatterers estimated with the Friis bistatic radar range equation. Note that we only consider those point clutter that fall within the BS mainlobe and within δr proximity of the two-way distance of the radar and MU. In other words, they must lie within the radar range limited resolution cell. To do so, we compute the slope of the line joining the scatterer and BS (m_0). Then we compute $m_1 = m_0 + \tan(\Delta\theta_{tx}/2)$ and $m_2 = m_0 - \tan(\Delta\theta_{tx}/2)$ based on the radar BS beamwidth ($\Delta\theta_{tx}$). The scatterer is within the radar beamwidth provided the product of the differences $(m_1 - m_0)$ and $(m_2 - m_0)$ is negative. Then we check if the absolute difference of the two-way path lengths of MU ($R_m^{tx} + R_m^{rx}$) and point clutter ($R_c^{tx} + R_c^{rx}$) is within the range resolution δr . If the resulting SCNR is above the predefined threshold γ , then we assume that the target is detected. The results over a large number of realizations are used to compute the \mathcal{P}_{DC}^{Bi} for the results presented in **Figures 3–8** (a) in this section. Note that the Monte Carlo simulations are useful to test some key assumptions made in SG based analysis such as the path loss approximation of the point clutter within the radar range limited resolution cell to the path loss of the MU.

4.1 Explore/Exploit Duty Cycle (ϵ)

In the JRC framework, a key parameter is $\epsilon = \frac{T_{search}}{T}$, the duty cycle, of the system. When ϵ is high, there is longer time for radar localization (T_{search}) but less time for communication service (T_{serve}) and vice versa. As a result, the radar beams can be narrow while scanning the angular search space. This results in weaker

detection performance due to poorer gain. The Theorem (7) shows the dependence of throughput Υ on ϵ which can be written as

$$\Upsilon(\epsilon) = A_0 e^{-a/\epsilon} (1 - \epsilon), \quad (27)$$

where

$$a = \frac{-\gamma N_s \kappa_m^4}{\sigma_{mavg} P_{tx} G_0 B_0 H_0} + \frac{-\gamma \rho_c c \tau \kappa_m^2 \sigma_{cavg}}{B_0 (\kappa_m + \sqrt{\kappa_m^2 - L^2}) (\sigma_{mavg} + \gamma \sigma_{cavg})} \quad (28)$$

and

$$A_0 = \left(2\pi \kappa_m - \frac{3\pi L^2}{8\kappa_m} \right) \frac{\rho_m c \tau D}{2\sqrt{1 - \frac{L^2}{4\kappa_m^2}}}. \quad (29)$$

We find the optimized $\tilde{\epsilon}$ for maximum throughput by equating the first derivative of Υ to zero.

Corollary 1.1. The optimum explore/exploit duty cycle ($\tilde{\epsilon}$) for maximum throughput is given by

$$\tilde{\epsilon} = \frac{\sqrt{a^2 + 4a} - a}{2} \quad (30)$$

TABLE 1 | Radar, target and clutter parameters used in the stochastic geometry formulations and Monte Carlo simulations.

Parameter	Symbol	Values
Baselength	L	5 m
Transmitted power	P_{tx}	1 mW
Total time	$T_{search} + T_{serve}$	1 s
Dwell time	T_{beam}	5 ms
Pulse width	$\tau = \frac{1}{BW}$	1 ns
Noise temperature (Kelvin)	T_s	300 K
Gain constant	G_0	1
Threshold	γ	1
Mean clutter RCS	σ_{cavg}	1 m ²
Clutter density	ρ_c	0.01 /m ²
Mean MU RCS	σ_{mavg}	1 m ²

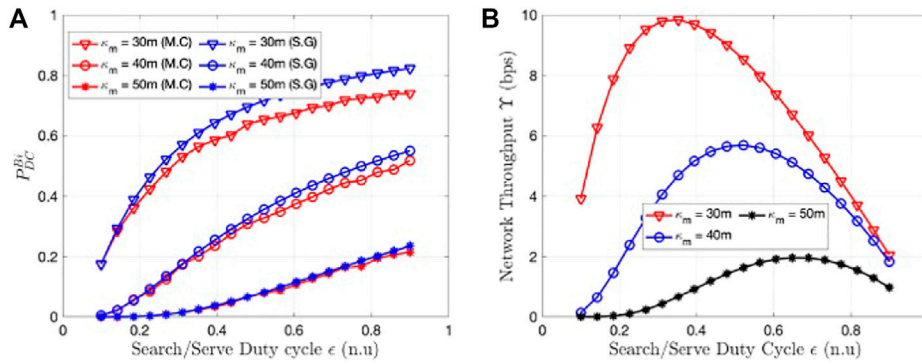


FIGURE 3 | (A) Detection coverage (P_{DC}^{Bi}) and **(B)** network throughput (Y) as a function of explore/exploit duty cycle (ϵ) for parametric bistatic range (κ).

The above case shows that the duty cycle is a function of the SCNR of the JRC system (shown in a in (28)). **Figure 3** shows the variation of P_{DC}^{Bi} and Y with respect to ϵ for different values of κ_m . The view graph, **Figure 3A**, shows that P_{DC}^{Bi} improves with increase in ϵ . In other words, when we have longer search time, we can use finer beams to search for the MU and thus have a greater likelihood of detecting them. However, the same is not true for the throughput (Y) shown in **Figure 3B**. An increase in ϵ initially improves the Y but subsequently causes a deterioration due to the reduction in communication service time. The optimum $\tilde{\epsilon}$ in the view graph matches the estimate from corollary (30). Since the above metric is shown to be a function of κ_m , it becomes difficult for a system operator to vary ϵ according to the position of the MU. Instead, we recommend that the above tuning is carried out for the maximum bistatic range of the JRC system which is determined based on the pulse repetition frequency. The selection of the PRF is discussed in subsection 5.4. Note that in the above view graphs, the results obtained from Monte Carlo system simulations closely match the results derived from the SG based analysis.

4.2 Signal-to-Noise Ratio Vs Signal-to-Clutter Ratio

Next, we discuss the effects of noise and clutter on the performance of the JRC. As pointed out earlier, there are two terms within the P_{DC}^{Bi} in (7) and (8). The first term captures the effect of the SNR on the JRC performance while the second term captures the effect of the SCR. **Figure 4** shows the effect of increasing the transmitted power P_{tx} on P_{DC}^{Bi} and Y . The results show that P_{DC}^{Bi} and Y increase initially with increase in power but subsequently, the performance saturates because the clutter returns also increase proportionately with increase in P_{tx} . On the other hand, when we consider the radar bandwidth which is the reciprocal of the pulse width ($BW = \frac{1}{\tau}$), we observe that there is an optimum BW for maximum Y in **Figure 5B**. This is because when BW is increased, the range resolution decreases and correspondingly the clutter resolution cell size. As a result, fewer clutter

scatterers contribute to the SCNR. But, on the other hand, the radar noise ($N_s = K_B T_s BW$) also increases which results in poorer quality radar links.

Corollary 1.2. *The optimum bandwidth \widetilde{BW} for maximum throughput Y is obtained by the derivation of (8) with respect to BW and is given by*

$$\widetilde{BW} = \left(\frac{\rho_c c \sigma_{c_{avg}} \sigma_{m_{avg}} P_{tx} G_0 H_0}{\kappa_m^2 K_B T_s (\kappa_m + \sqrt{\kappa_m^2 - L^2}) (\sigma_{m_{avg}} + \gamma \sigma_{c_{avg}})} \right)^{1/2} \quad (31)$$

The Monte Carlo results in **Figure 5A** show good agreement with SG results especially for higher values of wider BW . At low narrow BW , the errors due to the path loss approximation between the point clutter and the MU become more evident. However, in real world scenarios, microwave/millimeter JRC systems are developed specifically for high wide bandwidth waveforms for obtaining fine range resolution of the MU. Next we study the impact of clutter density and clutter RCS in **Figure 6** and **Figure 7**. When the clutter density is low (ρ_c approaches zero), we observe that P_{DC}^{Bi} decays at the fourth power of κ_m as shown in **Figure 6** and the throughput is entirely a function of the SNR. For large values of κ_m , the system is dominated by the effects of clutter rather than noise. We observe that the throughput increases initially with increase in κ_m due to the increase in number of MU within the area of interest and then subsequently the throughput falls due to the deterioration in the radar link metrics.

The effect of $\sigma_{c_{avg}}$ is less significant on P_{DC}^{Bi} and Y as both curves are flat in **Figures 7A,B**. On the other hand, the performances are far more sensitive to $\sigma_{m_{avg}}$.

4.3 Monostatic Conditions

A monostatic radar is a specific case of bistatic radar where the baseline length, L , and bistatic angle, β , are zero. Here, the one-way propagation distance from the transmitter and receiver to the target are equal. Hence, a monostatic radar can be assumed to be at the origin with the bistatic range κ_m equal to the polar distance r_m . We can, then directly, derive the radar detection coverage

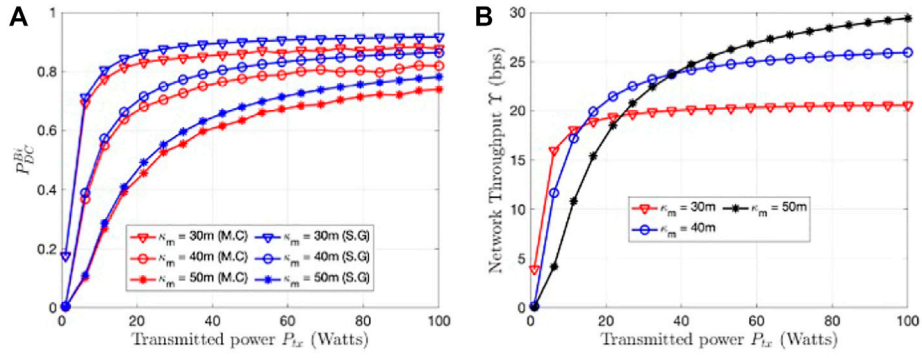


FIGURE 4 | (A) Detection coverage (P_{DC}^{Bi}) and **(B)** network throughput (Y) as a function of transmitted power (P_{tx}) for parametric bistatic range (k_m).

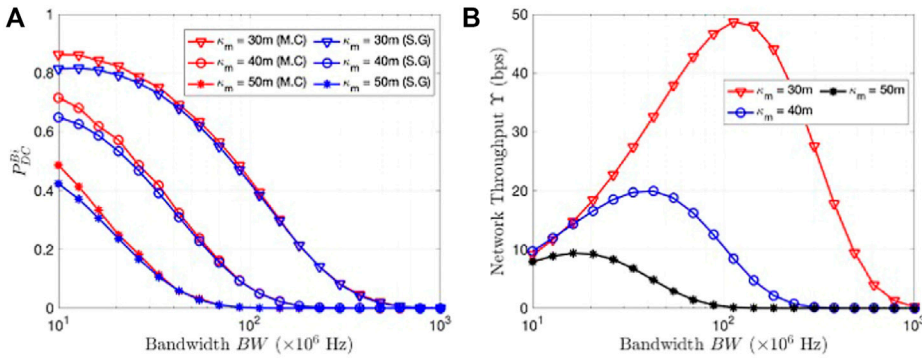


FIGURE 5 | (A) Detection coverage (P_{DC}^{Bi}) and **(B)** network throughput (Y) as a function of bandwidth (BW) for parametric bistatic range (k_m).

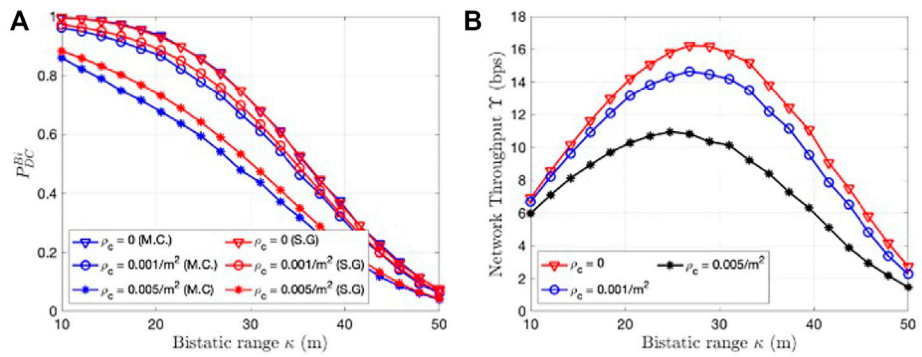


FIGURE 6 | (A) Detection coverage (P_{DC}^{Bi}) and **(B)** network throughput (Y) as a function of bistatic range (k_m) for parametric clutter density (ρ_c).

metric and throughput for this scenario from the bistatic case by making the corresponding substitutions to (7) and (8) and derive the following corollary.

Corollary 1.3. The radar detection coverage metric (\mathcal{P}_{DC}^{Mono}) and network throughput (Y) for a explore/exploit duty cycle (ϵ) for a monostatic radar based JRC system is given by

$$Y = \mathcal{P}_{DC}^{Mono} \pi r_m \rho_m c \tau (1 - \epsilon) D \quad (32)$$

where

$$\mathcal{P}_{DC}^{Mono} = \exp \left(\frac{-\gamma N_s r_m^4}{\sigma_{m_{avg}} P_{tx} G_0 B_0 \epsilon H_0} + \frac{-\gamma \rho_c c \tau r_m \sigma_{c_{avg}}}{2 B_0 \epsilon (\sigma_{m_{avg}} + \gamma \sigma_{c_{avg}})} \right) \quad (33)$$

The corollary again shows that the detection performance in the case of the monostatic radar can be studied through the SNR (the first term within the exponent of (33)) and the SCR

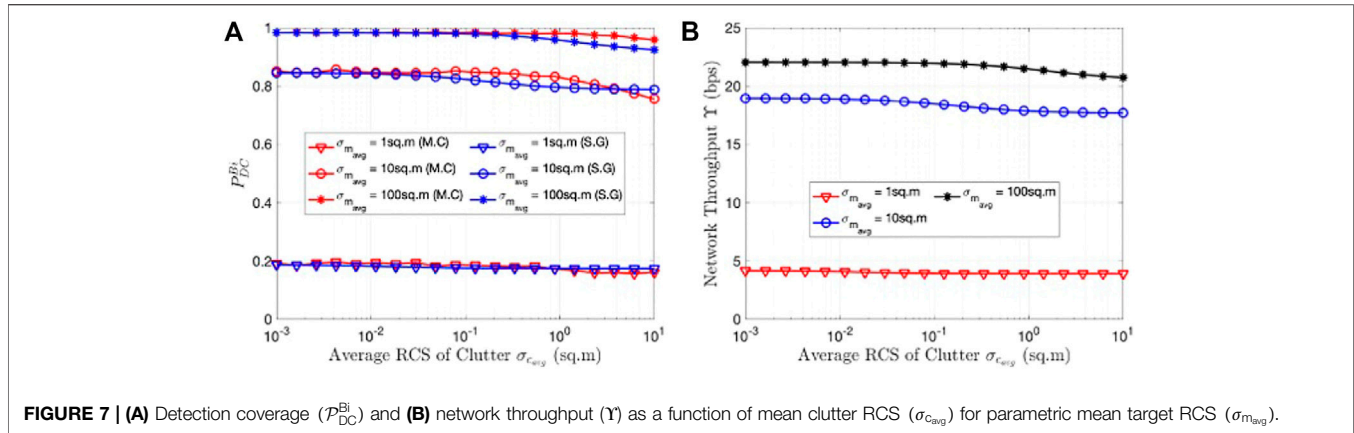


FIGURE 7 | (A) Detection coverage (P_{DC}^{Bi}) and **(B)** network throughput (Y) as a function of mean clutter RCS (σ_{cavg}) for parametric mean target RCS (σ_{mavg}).

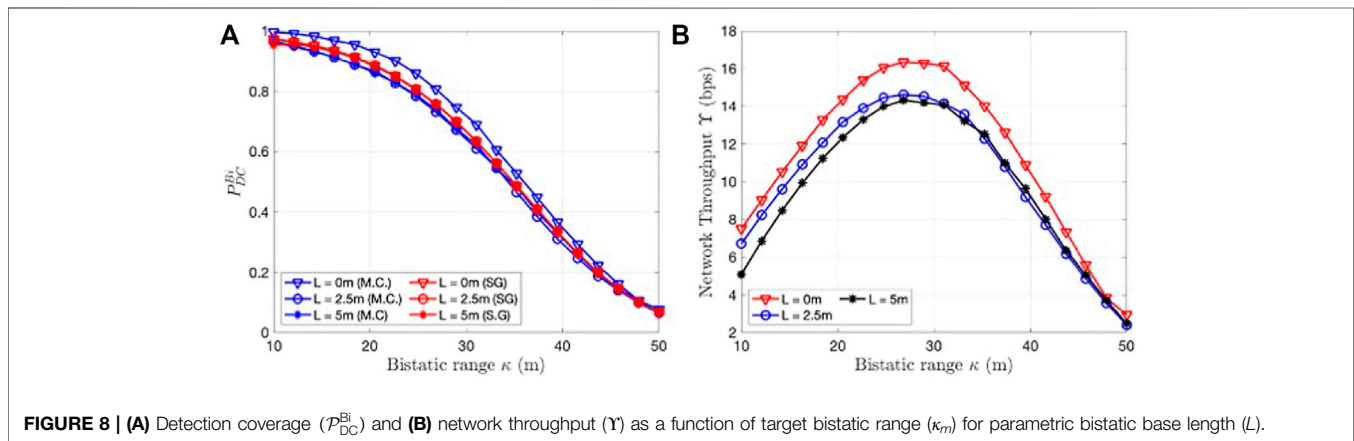


FIGURE 8 | (A) Detection coverage (P_{DC}^{Bi}) and **(B)** network throughput (Y) as a function of target bistatic range (κ_m) for parametric bistatic base length (L).

(the second term within the exponent). The SNR deteriorates as a function of the fourth power of the target distance while the SCR deteriorates linearly as a function of target radial distance. Hence, at greater distances we are limited by the clutter rather than the noise. We compare the monostatic and bistatic radar performances using the baseline length L as a parameter in **Figure 8**. Note that for all values of L and κ_m in the above study, the MU remains within the cosite region of the radar. The result show that the P_{DC}^{Bi} does not vary significantly for change from monostatic ($L = 0$) to bistatic ($L > 0$) conditions. In other words, the mean number of MU detected does not change significantly in both cases. The throughput, on the other hand, shown in **Figures 8B**, is higher for the monostatic case and appears to reduce slightly for increase in baseline length. This is because the circumference of the Cassini oval reduces slightly from the monostatic case to the bistatic case as per (24). Hence, fewer MU will be selected for a fixed bistatic range.

4.4 Pulse Repetition Interval

The maximum two-way unambiguous range of a radar, $R_{max} = (R_m^{tx} + R_m^{rx})_{max}$, is equal to cT_{PRI} . Through the intersection of the ellipse defined for a uniform R_{max} and the Cassini oval of constant κ_m , the two terms are related through

$$R_{max} = cT_{PRI} = L^2 + 2\kappa_m^2(1 + \cos\beta). \quad (34)$$

Note that in the above expression, the bistatic range changes for the parameter β . The maximum value that $\cos\beta$ can take is 1. Hence, for a given radar's T_{PRI}

$$\kappa_{max} = \frac{1}{2}(c^2T_{PRI}^2 - L^2)^{1/2}. \quad (35)$$

If we assume that at this range $\kappa_{max} \gg L$, then $P_{DC}^{Bi}(\kappa_{max})$ is given by

$$P_{DC}^{Bi}(\kappa_{max}) = \exp\left(\frac{-\gamma N_s (c^2T_{PRI}^2 - L^2)^2}{16\sigma_{mavg} P_{tx} G_0 B_0 H_0} + \frac{-\gamma P_r c \tau \sigma_{cavg} (c^2T_{PRI}^2 - L^2)^{1/2}}{4B_0 \epsilon (\sigma_{mavg} + \gamma \sigma_{cavg})}\right), \quad (36)$$

and the throughput is given by

$$Y(\kappa_{max}) = P_{DC}^{Bi}(\kappa_{max}) \frac{\pi}{2} (c^2T_{PRI}^2 - L^2)^{1/2} \rho_m c \tau (1 - \epsilon) D. \quad (37)$$

In the above throughput expression, it is evident that if the T_{PRI} is larger, the radar detection performance deteriorates. However, a larger number of MU are included in the region-of-interest due to which there are some gains in the throughput. We assume that if the R_{max} is high enough to ignore the effects of L , the radar operates under clutter limited conditions, and the throughput is a function of T_{PRI} , as given in

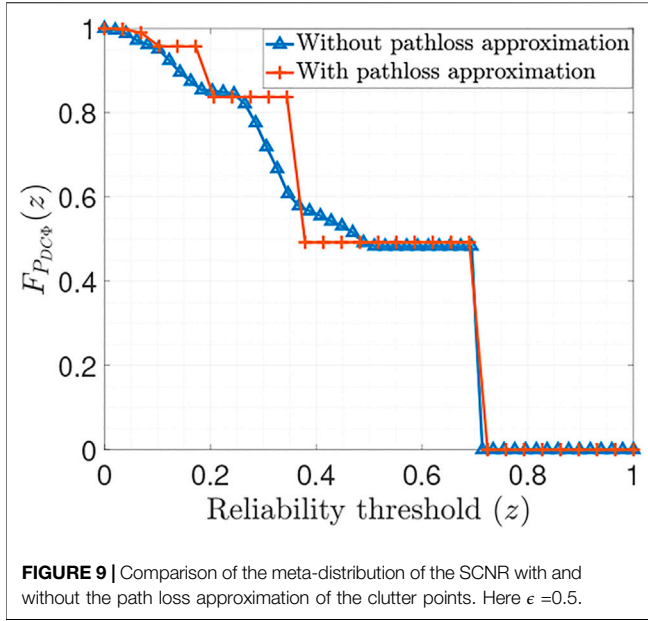


FIGURE 9 | Comparison of the meta-distribution of the SCNR with and without the path loss approximation of the clutter points. Here $\epsilon = 0.5$.

$$\Upsilon(T_{PRI}) = \exp\left(-\frac{\gamma\rho_c\sigma_{c_{avg}}c^2\tau T_{PRI}}{4B_0\epsilon(\sigma_{m_{avg}} + \gamma\sigma_{c_{avg}})}\right) \frac{\pi c^2\tau T_{PRI}\rho_m(1-\epsilon)D}{2} \quad (38)$$

Corollary 1.4. Accordingly, the optimum pulse repetition interval, \tilde{T}_{PRI} , can be estimated for maximum throughput as

$$\tilde{T}_{PRI} = \frac{4B_0\epsilon(\sigma_{m_{avg}} + \gamma\sigma_{c_{avg}})}{\gamma\rho_c\sigma_{c_{avg}}c^2\tau} \quad (39)$$

The above expression shows that higher ϵ (resulting in narrow beams) and shorter pulse duration (smaller τ) will allow for a longer pulse repetition interval and unambiguous range due to improvement in the link metrics.

4.5 Meta Distribution of Signal-to-Clutter-and-Noise Ratio in a Bistatic Radar

Although the \mathcal{P}_{DC}^{Bi} is a useful metric for tuning radar parameters, it only provides an average view of the network across all possible network realizations of the underlying point process. It is simply a *spatial* average of the detection performance of all radars across all clutter realizations in the region of interest. Hence, it does not reveal the performance of individual radars. This inhibits derivation of link-level reliability of the radar detection performance. In this regard, the meta-distribution, i.e., the distribution of the radar \mathcal{P}_{DC}^{Bi} conditioned on a realization of Φ provides a framework to study the same. For that, we introduce the random variable $\mathcal{P}_{DC\Phi}^{Bi}$, which denotes the bistatic detection coverage probability conditioned on the clutter realization, i.e., $\mathcal{P}_{DC\Phi}^{Bi} = \mathbb{P}(\text{SCNR}(\kappa_m) > \gamma | \Phi)$. The meta-distribution then is simply the distribution of the random variable $\mathcal{P}_{DC\Phi}^{Bi}$. Its complementary CDF,

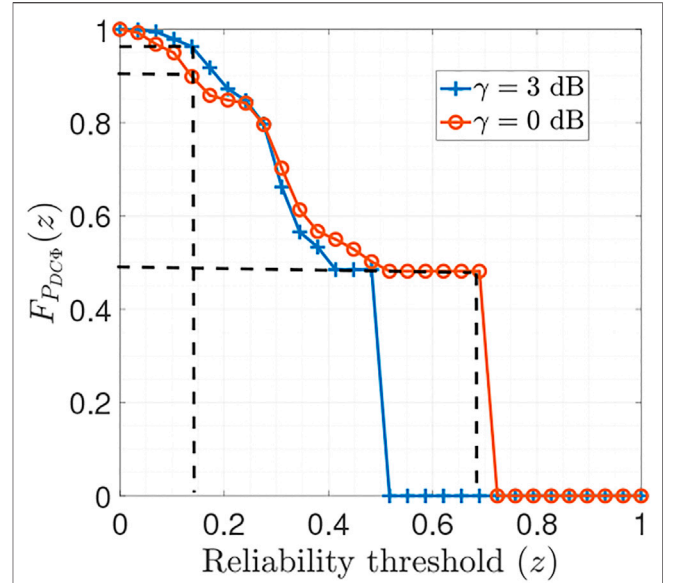


FIGURE 10 | Meta distribution of the SCNR for different SCNR thresholds.

i.e., $F_{\mathcal{P}_{DC\Phi}^{Bi}}(z) = \mathbb{P}(\mathcal{P}_{DC\Phi}^{Bi} \geq z)$, represents the probability with which at least z fraction of the bistatic radar links experience a successful radar detection when the SCNR threshold is set to γ . Mathematically,

$$\mathcal{P}_{DC\Phi}^{Bi} = \mathcal{P}(\text{SCNR}(\kappa_m) \geq \gamma | \Phi) = \mathcal{P}\left(\sigma_m \geq \sum_{c \in \Phi \cap \mathcal{A}_c(\kappa_m)} \frac{\gamma\sigma_c\kappa_m^4}{\kappa_c^4} + \frac{\gamma N_s\kappa_m^4}{P_{tx}G_0B_0\epsilon H_0} \middle| \Phi\right), \quad (40)$$

$$= \exp\left(-\frac{\gamma N_s\kappa_m^4}{\sigma_{m_{avg}}P_{tx}G_0B_0\epsilon H_0}\right) \times \left(\prod_{c \in \Phi \cap \mathcal{A}_c(\kappa_m)} \left(\frac{\gamma\sigma_{c_{avg}}(R_c^{tx})^{-2}(R_c^{rx})^{-2}\kappa_m^4}{\sigma_{m_{avg}} + \gamma\sigma_{c_{avg}}(R_c^{tx})^{-2}(R_c^{rx})^{-2}\kappa_m^4}\right)\right). \quad (41)$$

For a point clutter located at a distance, R_c^{tx} , from the transmitter at an angle θ_c^{tx} , we have $(R_c^{rx})^2 = (R_c^{tx})^2 + L^2 + 2R_c^{tx}L \cos(\theta_c^{tx})$. The direct evaluation of the exact distribution of $\mathcal{P}_{DC\Phi}^{Bi}$ is challenging. Thus, we take an indirect approach to evaluate it through the calculation of its moments. In particular, the b th moment of $\mathcal{P}_{DC\Phi}^{Bi}$ is given by:

$$\begin{aligned} M_b &= \mathbb{E}\left[T(b, \kappa_m) \left(\prod_{c \in \Phi \cap \mathcal{A}_c(\kappa_m)} \left(\frac{\gamma\sigma_{c_{avg}}(R_c^{tx})^{-2}(R_c^{rx})^{-2}\kappa_m^4}{\sigma_{m_{avg}} + \gamma\sigma_{c_{avg}}(R_c^{tx})^{-2}(R_c^{rx})^{-2}\kappa_m^4}\right)\right)^b\right] \\ &= T(b, m) \mathbb{E}\left[\left(\prod_{c \in \Phi \cap \mathcal{A}_c(\kappa_m)} \left(\frac{\gamma\sigma_{c_{avg}}(R_c^{tx})^{-2}(R_c^{rx})^{-2}\kappa_m^4}{\sigma_{m_{avg}} + \gamma\sigma_{c_{avg}}(R_c^{tx})^{-2}(R_c^{rx})^{-2}\kappa_m^4}\right)\right)^b\right] \\ &= \frac{1}{2\pi} T(b, m) \int_0^{2\pi} \exp\left(-\rho_c \int_{\theta_m^{tx} - \frac{\Delta\theta}{2}}^{\theta_m^{tx} + \frac{\Delta\theta}{2}} \int_{R_m^{tx} - \frac{\Delta R}{2}}^{R_m^{tx} + \frac{\Delta R}{2}} 1 - \left(\frac{\gamma\sigma_{c_{avg}}\gamma^{-2}\gamma_r^{-2}\kappa_m^4}{\sigma_{m_{avg}} + \gamma\sigma_{c_{avg}}\gamma^{-2}\gamma_r^{-2}\kappa_m^4}\right) y dy d\theta_c^{tx}\right) d\theta_m \\ &= \frac{1}{2\pi} T(b, m) \int_0^{2\pi} \exp\left(-\rho_c \sum_{k=1}^b \binom{b}{k} \int_{\theta_m^{tx} - \frac{\Delta\theta}{2}}^{\theta_m^{tx} + \frac{\Delta\theta}{2}} \int_{R_m^{tx} - \frac{\Delta R}{2}}^{R_m^{tx} + \frac{\Delta R}{2}} \left(\frac{\gamma\sigma_{c_{avg}}\gamma^{-2}\gamma_r^{-2}\kappa_m^4}{\sigma_{m_{avg}} + \gamma\sigma_{c_{avg}}\gamma^{-2}\gamma_r^{-2}\kappa_m^4}\right)^k y dy d\theta_c^{tx}\right) d\theta_m, \end{aligned} \quad (42)$$

where $\mathcal{T}(b, m) = \exp\left(-\frac{\gamma b N_s \kappa_m^4}{\sigma_{m_{avg}} P_{tx} G_0 B_0 \epsilon H_0}\right)$, $\gamma_r = (\gamma^2 + L^2 - 2\gamma L \cos(\theta_c^{rx}))^{\frac{1}{2}}$. Now, for a large bandwidth, the range-resolution cell is relatively small, and hence, with the path loss approximation $\sqrt{R_c^{tx} R_c^{rx}} = \kappa_m$ for all clutter points within the cell, we have:

$$\begin{aligned} M_b &= \exp\left(-\frac{\gamma b N_s \kappa_m^4}{\sigma_{m_{avg}} P_{tx} G_0 B_0 \epsilon H_0}\right) \mathbb{E}_n \left[\left(\frac{\sigma_{m_{avg}}}{\sigma_{m_{avg}} + \gamma \sigma_{c_{avg}}} \right)^{nb} \right] \\ &= \exp\left(-\frac{\gamma b N_s \kappa_m^4}{\sigma_{m_{avg}} P_{tx} G_0 B_0 \epsilon H_0}\right) \exp\left(\rho_c A_c(\kappa_m) \left(\left(\frac{\sigma_{m_{avg}}}{\sigma_{m_{avg}} + \gamma \sigma_{c_{avg}}} \right)^b - 1 \right)\right) \end{aligned} \quad (43)$$

We note here that with the path loss approximation, only the number of clutter points (and not their locations) inside the range resolution cell n impacts the moment. Then, the complementary CDF of the conditional $\mathcal{P}_{DC\Phi}^{Bi}$ can be evaluated using the Gil-Pelaez inversion theorem as:

$$F_{\mathcal{P}_{DC\Phi}^{Bi}}(z) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{\mathcal{I}(\exp(-ju \log(z))) M_{ju}}{u} du \quad (44)$$

where, $j = \sqrt{-1}$ and $M_{ju}(\cdot)$ is the ju -th moment of $\mathcal{P}_{DC\Phi}^{Bi}$.

In **Figure 9** we see the impact of the path loss approximation of the clutter points on the meta-distribution of the SCNR. In particular, we see that since with the path loss approximation, the meta-distribution depends only on the number of clutter points within the range resolution cell, the corresponding plot has a stepped behaviour, where each step corresponds to a certain number of clutter points. On the contrary, the plot without the path loss approximation takes into account the relative randomness in the locations of the clutter points within the range resolution cell. For a given κ_m , the path loss approximation may result in either an overestimation or an underestimation of the actual meta-distribution. However, such an analysis is out of scope of the current work and will be investigated in a future work. In **Figure 10** we plot the meta-distribution of the SCNR for different SCNR thresholds. This represents, qualitatively, a fine-grained analysis of the radar detection. For a given γ the meta-distribution evaluated at a given z represents the fraction of radar links that experience a successful radar detection at least $z\%$ of the time. For example, when the radar detection threshold is set at $\gamma = 0$ dB, we observe that about half ($F_{P_{DC\Phi}}(z) = 0.5$) of the targets are detected with a reliability of at least 70% (i.e., $z = 0.7$), while virtually no targets ($F_{P_{DC\Phi}}(z) = 0$) are detected with a reliability of 70% when the detection threshold is set at $\gamma = 3$ dB. On the lower reliability regime, interestingly, we observe that with $\gamma = 3$ dB, more than 95% of the targets ($F_{P_{DC\Phi}}(z) = 0.95$) are detected with a reliability of at least 15% (i.e., with $z = 0.15$) while the same for $\gamma = 0$ dB is lower (about 90%). This also indicates that for a lower SCNR threshold, not only the detection probability \mathcal{P}_{DC}^{Bi} is higher, but also guaranteeing higher reliability for individual links is more likely. Remarkably, we observe that regardless of the value of \mathcal{P}_{DC}^{Bi} , none of the targets can be guaranteed to be detected beyond 70% ($z = 0.7$) reliability, and

to achieve that, additional radar transceivers must be deployed.

5 CONCLUSION

We have provided an SG based analytical framework to provide system level planning insights into how radar based localization can enhance communication throughput of a JRC system. The key advantage of this framework is that it accounts for the significant variations in the radar, target and clutter conditions that may be encountered in actual deployments without requiring laborious system level simulations or measurement data collection. Specifically, we provide a theorem to optimize JRC system parameters such as the explore/exploit duty cycle, the transmitted power, bandwidth and pulse repetition interval for maximizing the network throughput. The results are presented for generalized bistatic radar scenarios from which the monostatic results are derived through limiting conditions. We also provide a study on the meta-distribution of the radar detection metric which provides the key insight that none of the mobile users can be reliably detected beyond 70% of the time with a single JRC configuration. Our results are validated with Monte Carlo simulations.

The analysis in this work is based on some assumptions: First, we have assumed a planar bistatic radar geometry where all the mobile users/radar targets fall in the cosite region (baseline length is below twice the bistatic range). These assumptions are satisfied in several JRC applications such as indoor localization using WiFi/WLAN devices and in radar enhanced vehicular communications. However, the assumption does not hold for GNSS based bistatic radar remote sensing where the transmitter is the satellite while the receiver is mounted close to the earth and a three-dimensional geometry would have to be considered. Hence, our future work will focus on the modification to the SG based analysis to analyze radar performance metrics under 3D, non-cosite conditions of the bistatic radar.

Second, we have considered short range line-of-sight links in our study which are applicable to mmWave JRC implementations. However, real world deployments encounter blockages that must be accounted for from a JRC system design perspective. Similarly, the radar will receive returns from sidelobes along with the main lobes which has not been considered in our work. Finally, in our throughput analysis, we have assumed that all the mobile users have uniform data rates that can be supported. In real world conditions, the requirements from individual users will differ and there may be system constraints on the maximum resource utilization. Therefore, the study of the performance bounds due to more realistic channel, radar and mobile user models will lead to more accurate estimation of the detection performance and network throughput and would form the basis of future studies.

Third, in this work, we have confined our discussion to a single bistatic radar framework. In the foreseeable future, we may encounter radar networks with a single transmitter and multiple receivers, or even multiple transmitters and receivers. In these conditions, there can be significant diversity in the radar and target geometry which can be effectively analysed through SG. Research into multistatic radar frameworks would form a natural extension to this work.

DATA AVAILABILITY STATEMENT

All the codes used to generate the figures in the document can be accessed at https://essrg.iitd.edu.in/?page_id=4355.

REFERENCES

- Al-Hourani, A., Evans, R. J., Kandeepan, S., Moran, B., and Eltom, H. (2017). Stochastic Geometry Methods for Modeling Automotive Radar Interference. *IEEE Trans. Intell. Transportation Syst.* 19, 333–344. doi:10.1109/tits.2016.2632309
- Ali, A., González-Prelcic, N., and Ghosh, A. (2020). Passive Radar at the Roadside Unit to Configure Millimeter Wave Vehicle-To-Infrastructure Links. *IEEE Trans. Veh. Technol.* 69, 14903–14917. doi:10.1109/tvt.2020.3027636
- Alloulah, M., and Huang, H. (2019). Future Millimeter-Wave Indoor Systems: A Blueprint for Joint Communication and Sensing. *Computer* 52, 16–24. doi:10.1109/mc.2019.2914018
- Andrews, J. G., Baccelli, F., and Ganti, R. K. (2011). A Tractable Approach to Coverage and Rate in Cellular Networks. *IEEE Trans. Commun.* 59, 3122–3134. doi:10.1109/tcomm.2011.100411.100541
- Bai, T., and Heath, R. W. (2014). Coverage and Rate Analysis for Millimeter-Wave Cellular Networks. *IEEE Trans. Wireless Commun.* 14, 1100–1114. doi:10.1109/TWC.2014.2364267
- Billingsley, J. B. (2002). *Low-angle Radar Land Clutter: Measurements and Empirical Models*. Norwich, NY: IET.
- Chen, X., Tharmarasa, R., Pelletier, M., and Kirubarajan, T. (2012). Integrated Radar Estimation and Target Tracking Using Poisson point Processes. *IEEE Trans. Aerosp. Electron. Syst.* 48, 1210–1235. doi:10.1109/taes.2012.6178058
- Chiu, S. N., Stoyan, D., Kendall, W. S., and Mecke, J. (2013). *Stochastic Geometry and its Applications*. West Sussex, United Kingdom: John Wiley & Sons.
- Dokhanchi, S. H., Mysore, B. S., Mishra, K. V., and Ottersten, B. (2019). A Mmwave Automotive Joint Radar-Communications System. *IEEE Trans. Aerosp. Electron. Syst.* 55, 1241–1260. doi:10.1109/taes.2019.2899797
- Duggal, G., Mishra, K. V., and Ram, S. S. (2019). “Micro-doppler and Micro-range Detection via Doppler-Resilient 802.11ad-Based Vehicle-To-Pedestrian Radar,” in *2019 IEEE Radar Conference* (Boston, MA: IEEE), 1–6. doi:10.1109/radar.2019.8835525
- Duggal, G., Vishwakarma, S., Mishra, K. V., and Ram, S. S. (2020). Doppler-Resilient 802.11ad-Based Ultrashort Range Automotive Joint Radar-Communications System. *IEEE Trans. Aerosp. Electron. Syst.* 56, 4035–4048. doi:10.1109/taes.2020.2990393
- Falcone, P., Colone, F., and Lombardo, P. (2012). Potentialities and Challenges of Wifi-Based Passive Radar. *IEEE Aerosp. Electron. Syst. Mag.* 27, 15–26. doi:10.1109/maes.2012.6380822
- Fang, Z., Wei, Z., Chen, X., Wu, H., and Feng, Z. (2020). Stochastic Geometry for Automotive Radar Interference with Rcs Characteristics. *IEEE Wireless Commun. Lett.* 9, 1817–1820. doi:10.1109/lwc.2020.3003064
- Ghatak, G., De Domenico, A., and Coupechoux, M. (2018). Coverage Analysis and Load Balancing in Hetnets with Millimeter Wave Multi-Rat Small Cells. *IEEE Trans. Wireless Commun.* 17, 3154–3169. doi:10.1109/twc.2018.2807426
- Ghatak, G., Koirala, R., De Domenico, A., Denis, B., Dardari, D., Uguen, B., et al. (2021). Beamwidth Optimization and Resource Partitioning Scheme for Localization Assisted Mm-Wave Communication. *IEEE Trans. Commun.* 69, 1358–1374. doi:10.1109/TCOMM.2020.3036864

AUTHOR CONTRIBUTIONS

The theoretical formulations and derivations and writing of the paper were carried out by SR in collaboration with GG. The Monte Carlo simulations for experimental validation were carried out by SS.

FUNDING

Project is funded through a grant from Ministry of Electronics and Information Technology, Government of India, No.13 (30/2020-CC&BT).

- Grossi, E., Lops, M., Tulino, A. M., and Venturino, L. (2021). Opportunistic Sensing Using Mmwave Communication Signals: A Subspace Approach. *IEEE Trans. Wireless Commun.* 20, 4420–4434. doi:10.1109/TWC.2021.3058775
- Grossi, E., Lops, M., and Venturino, L. (2017). Two-step Sequential Detection in Agile-Beam Radars: Performance and Tradeoffs. *IEEE Trans. Aerosp. Electron. Syst.* 53, 2199–2213. doi:10.1109/taes.2017.2688878
- Haenggi, M. (2012). *Stochastic Geometry for Wireless Networks*. Cambridge, United Kingdom: Cambridge University Press.
- Hassanien, A., Amin, M. G., Zhang, Y. D., and Ahmad, F. (2016). Signaling Strategies for Dual-Function Radar Communications: An Overview. *IEEE Aerosp. Electron. Syst. Mag.* 31, 36–45. doi:10.1109/maes.2016.150225
- Hu, J., Wu, Y., Chen, R., Shu, F., and Wang, J. (2019). Optimal Detection of Uav's Transmission with Beam Sweeping in covert Wireless Networks. *IEEE Trans. Vehicular Technol.* 69, 1080–1085.
- Jackson, M. C. (1986). The Geometry of Bistatic Radar Systems. *IEE Proc. F Commun. Radar Signal. Process.* UK 133, 604–612. doi:10.1049/ip-f-1.1986.0097
- Kay, S. M. (1998). *Fundamentals of Statistical Signal Processing Detection Theory*. Upper Saddle River, NJ: Prentice-Hall 146, 222.
- Kellet, D., Garmatyuk, D., Mudaliar, S., Condict, N., and Qualls, I. (2019). Random Sequence Encoded Waveforms for covert Asynchronous Communications and Radar. *IET Radar, Sonar & Navigation* 13, 1713–1720. doi:10.1049/iet-rsn.2018.5659
- Kumari, P., Choi, J., González-Prelcic, N., and Heath, R. W. (2017). Ieee 802.11 Ad-Based Radar: An Approach to Joint Vehicular Communication-Radar System. *IEEE Trans. Vehicular Technol.* 67, 3012–3027. doi:10.1109/TVT.2017.2774762
- Li, W., Piechocki, R. J., Woodbridge, K., Tang, C., and Chetty, K. (2020). Passive Wifi Radar for Human Sensing Using a Stand-Alone Access point. *IEEE Trans. Geosci. Remote Sensing* 59, 1986–1998. doi:10.1109/TGRS.2020.3006387
- Liu, F., Masouros, C., Petropulu, A. P., Griffiths, H., and Hanzo, L. (2020). Joint Radar and Communication Design: Applications, State-Of-The-Art, and the Road Ahead. *IEEE Trans. Commun.* 68, 3834–3862. doi:10.1109/tcomm.2020.2973976
- Ma, D., Shlezinger, N., Huang, T., Shavit, Y., Namer, M., Liu, Y., et al. (2021). Spatial Modulation for Joint Radar-Communications Systems: Design, Analysis, and Hardware Prototype. *IEEE Trans. Veh. Technol.* 70, 2283–2298. doi:10.1109/tvt.2021.3056408
- Miranda, S. L. C., Baker, C. J., Woodbridge, K., and Griffiths, H. D. (2007). Comparison of Scheduling Algorithms for Multifunction Radar. *IET Radar Sonar Navig.* 1, 414–424. doi:10.1049/iet-rsn:20070003
- Mishra, K. V., Bhavani Shankar, M. R., Koivunen, V., Ottersten, B., and Vorobyov, S. A. (2019). Toward Millimeter-Wave Joint Radar Communications: A Signal Processing Perspective. *IEEE Signal. Process. Mag.* 36, 100–114. doi:10.1109/msp.2019.2913173
- Moyer, L. R., Morgan, C. J., and Rugger, D. A. (1989). An Exact Expression for Resolution Cell Area in Special Case of Bistatic Radar Systems. *IEEE Trans. Aerosp. Electron. Syst.* 25, 584–587. doi:10.1109/7.32092
- Munari, A., Simic, L., and Petrova, M. (2018). Stochastic Geometry Interference Analysis of Radar Network Performance. *IEEE Commun. Lett.* 22, 2362–2365. doi:10.1109/lcomm.2018.2869742

- Nitsche, T., Cordeiro, C., Flores, A., Knightly, E., Perahia, E., and Widmer, J. (2014). IEEE 802.11ad: Directional 60 GHz Communication for Multi-Gigabit-Per-Second Wi-Fi [Invited Paper]. *IEEE Commun. Mag.* 52, 132–141. doi:10.1109/mcom.2014.6979964
- Park, J., and Heath, R. W. (2018). Analysis of Blockage Sensing by Radars in Random Cellular Networks. *IEEE Signal. Process. Lett.* 25, 1620–1624. doi:10.1109/lsp.2018.2869279
- Ram, S. S., and Ghatak, G. (2022). *Estimation of Bistatic Radar Detection Performance under Discrete Clutter Conditions Using Stochastic Geometry*. *arXiv e-prints*, arXiv: 2201.03221.
- Ram, S. S., Singh, G., and Ghatak, G. (2020). “Estimating Radar Detection Coverage Probability of Targets in a Cluttered Environment Using Stochastic Geometry,” in *2020 IEEE International Radar Conference (RADAR)* (Washington D.C.: IEEE), 665–670. doi:10.1109/radar42522.2020.9114637
- Ram, S. S., Singh, G., and Ghatak, G. (2021). Optimization of Radar Parameters for Maximum Detection Probability under Generalized Discrete Clutter Conditions Using Stochastic Geometry. *IEEE Open J. Signal. Process.* 2, 571–585. doi:10.1109/ojsp.2021.3121199
- Raynal, A. M., Bickel, D. L., Denton, M. M., Bow, W. J., and Doerry, A. W. (2011a). in *Radar Cross Section Statistics of Ground Vehicles at Ku-Band (SPIE Proceedings)*. United States: SPIE 8021.
- Raynal, A. M., Burns, B. L., Verge, T., Bickel, D. L., Dunkel, R., and Doerry, A. W. (2011b). Radar Cross Section Statistics of Dismounts at Ku-Band. *Radar Sensor Techn.* XV, 8021. doi:10.1117/12.882873
- Ren, P., Munari, A., and Petrova, M. (2018). Performance Tradeoffs of Joint Radar-Communication Networks. *IEEE Wireless Commun. Lett.* 8, 165–168. doi:10.1109/LWC.2018.2865360
- Ruoskanen, J., Eskelinen, P., and Heikkilä, H. (2003). Millimeter Wave Radar with Clutter Measurements. *IEEE Aerosp. Electron. Syst. Mag.* 18, 19–23. doi:10.1109/maes.2003.1244771
- Sekine, M., Mao, Y., and Mao, Y. (1990). *Weibull Radar Clutter*. London, United Kingdom: IEE Radar, Sonar, Navigation and Avionics Series 3, by Peter Peregrinus Ltd., 3.
- Skolnik, M. I. (1961). An Analysis of Bistatic Radar. *IRE Trans. Aeronaut. Navig. Electron.* ANE-8, 19–27. doi:10.1109/tane3.1961.4201772
- Skolnik, M. I. (1980). *Introduction to Radar Systems*. New York: McGraw Hill Book Co., 590.
- Storrer, L., Yildirim, H. C., Crauwels, M., Copa, E. I. P., Pollin, S., Louveaux, J., et al. (2021). Indoor Tracking of Multiple Individuals with an 802.11ax Wi-Fi-Based Multi-Antenna Passive Radar. *IEEE Sensors J.* 21, 20462–20474. doi:10.1109/jsen.2021.3095675
- Tan, B., Woodbridge, K., and Chetty, K. (2016). A Wireless Passive Radar System for Real-Time through-wall Movement Detection. *IEEE Trans. Aerosp. Electron. Syst.* 52, 2596–2603. doi:10.1109/taes.2016.140207
- Thornburg, A., Bai, T., and Heath, R. W. (2016). Performance Analysis of Outdoor Mmwave Ad Hoc Networks. *IEEE Trans. Signal. Process.* 64, 4065–4079. doi:10.1109/tsp.2016.2551690
- Willis, N. J. (2005). *Bistatic Radar*, 2. Rayleigh, NC: SciTech Publishing.
- Yildirim, H. C., Determe, J. F., Storrer, L., Rottenberg, F., De Doncker, P., Louveaux, J., et al. (2021). Super Resolution Passive Radars Based on 802.11ax Wi-Fi Signals for Human Movement Detection. *IET Radar, Sonar & Navigation* 15, 323–339. doi:10.1049/rsn2.12038
- Zavorotny, V. U., Gleason, S., Cardellach, E., and Camps, A. (2014). Tutorial on Remote Sensing Using Gns Bistatic Radar of Opportunity. *IEEE Geosci. Remote Sens. Mag.* 2, 8–45. doi:10.1109/mgrs.2014.2374220
- Zhou, P., Cheng, K., Han, X., Fang, X., Fang, Y., He, R., et al. (2018). IEEE 802.11ay-Based mmWave WLANs: Design Challenges and Solutions. *IEEE Commun. Surv. Tutorials* 20, 1654–1681. doi:10.1109/comst.2018.2816920

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Ram, Singhal and Ghatak. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Small Object Detection and Tracking in Satellite Videos With Motion Informed-CNN and GM-PHD Filter

Camilo Aguilar¹, Mathias Ortner² and Josiane Zerubia^{1*}

¹Inria, Université Côte d'Azur, Sophia-Antipolis, France, ²Airbus DS, Toulouse, France

Small object tracking in low-resolution remote sensing images presents numerous challenges. Targets are relatively small compared to the field of view, do not present distinct features, and are often lost in cluttered environments. In this paper, we propose a track-by-detection approach to detect and track small moving targets by using a convolutional neural network and a Bayesian tracker. Our object detection consists of a two-step process based on motion and a patch-based convolutional neural network (CNN). The first stage performs a lightweight motion detection operator to obtain rough target locations. The second stage uses this information combined with a CNN to refine the detection results. In addition, we adopt an online track-by-detection approach by using the Probability Hypothesis Density (PHD) filter to convert detections into tracks. The PHD filter offers a robust multi-object Bayesian data-association framework that performs well in cluttered environments, keeps track of missed detections, and presents remarkable computational advantages over different Bayesian filters. We test our method across various cases of a challenging dataset: a low-resolution satellite video comprising numerous small moving objects. We demonstrate the proposed method outperforms competing approaches across different scenarios with both object detection and object tracking metrics.

Keywords: object detection, object tracking, PHD filter, CNNs, remote sensing

OPEN ACCESS

Edited by:

Maria Sabrina Greco,
University of Pisa, Italy

Reviewed by:

Allan De Freitas,
University of Pretoria, South Africa
Deepayan Bhowmik,
University of Stirling, United Kingdom

*Correspondence:

Josiane Zerubia
josiane.zerubia@inria.fr

Specialty section:

This article was submitted to
Image Processing,
a section of the journal
Frontiers in Signal Processing

Received: 01 December 2021

Accepted: 04 March 2022

Published: 29 April 2022

Citation:

Aguilar C, Ortner M and Zerubia J
(2022) Small Object Detection and
Tracking in Satellite Videos With Motion
Informed-CNN and GM-PHD Filter.
Front. Sig. Proc. 2:827160.
doi: 10.3389/frsip.2022.827160

INTRODUCTION

In recent years, object detection and tracking in remote sensing videos have become a widely attractive area of research. Novel satellite and Wide Area Motion Imagery (WAMI) technologies have created an unprecedented demand for fast and automatic information retrieval. For example, Airbus' Zephyr high altitude drones can cover up to 20, ×, 30 km² of continuous video surveillance, or the Chinese Jilin-1 satellite captures ground images spanning several kilometers with a 1-m spatial resolution imaged at 20 Hz.

The generated images contain essential information for civilian and military domains when ground sensors are not locally available. Sample civilian applications include urban planning (Wijnands et al., 2021), automatic traffic monitoring (Kaack et al., 2019), driving behavioral research (Chen et al., 2021), or commerce management with ship monitoring (Cao et al., 2019). Similarly, object detection and tracking contribute to military applications such as border protection or abnormal activity monitoring. For example, the work proposed by Kirubarajan et al. (2000) presents an approach to detect and tracks convoys in different scenarios such as road networks or open fields.

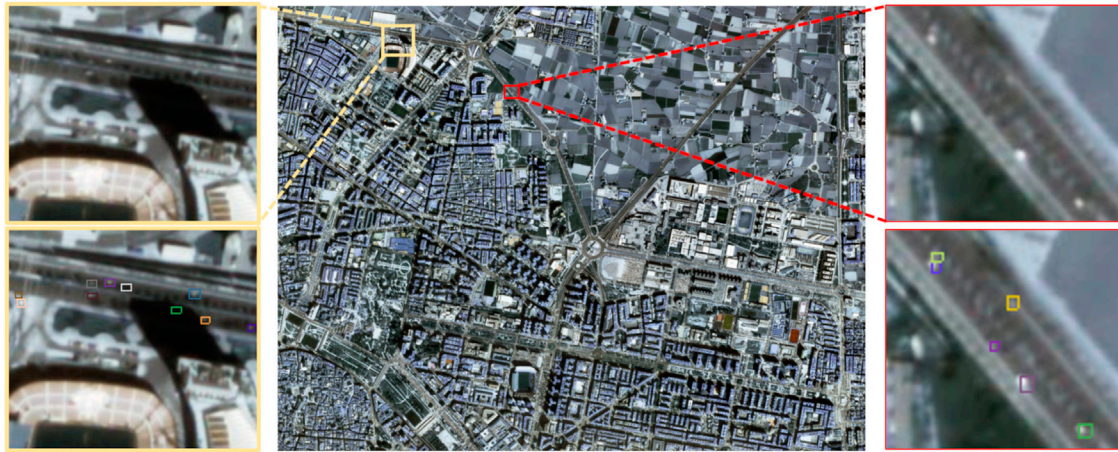


FIGURE 1 | Jilin-1 satellite image with provided annotations. Each colored box represents a target instance.

While object tracking has dramatically improved during the last years, a significant amount of approaches solve problems that contain large training datasets and feature-rich targets, such as pedestrian tracking in surveillance cameras or city landscapes. Nevertheless, novel methods need to tackle application-related challenges such as small object tracking in remote sensing images and have to overcome challenges such as datasets with scarce and incomplete annotations.

Particularly, targets in satellite images and high altitude drones present notable challenges to common detectors and trackers. First, objects of interest are very small compared to the field of view. For instance, **Figure 1** shows a ground image with a resolution of 1 m/pixel where vehicles span on average 5×6 pixels and resemble white moving blobs. In fact, numerous small objects appear at subpixel levels such as motorcycles and are not detectable for common appearance-based object detectors. Additionally, images show diverse noise sources such as illumination changes, clouds, shadows and environmental phenomena such as wind or rain. These noise sources generate numerous false positives when using motion as the main feature for object detection. Moreover, satellites and drones orbit introduce parallax effect noise for object detectors and motion prediction noise for object trackers.

In this paper, we present improvements and further results of our work presented by Aguilar et al. (2021) where we detect small objects using motion and appearance information. We use three consecutive frames to estimate moving object locations and we refine the detections using a patch-based Faster RCNN (Ren et al. (2015)). Specifically, in this paper we improve the patch-based detection by adding the motion response into the Faster RCNN input. The combination of motion and appearance information on extracted patches improves significantly Faster RCNN's object detection.

Once we obtain object measurements, we feed the extracted data to the probability hypothesis density (PHD) filter, proposed by Mahler (2003). This filter models multi-object states under a Markovian framework, where the state of each tracked object is

conditionally independent of all but the previous step. This assumption simplifies the filter and allows it to be computationally efficient in comparison to other related filters at the cost of tracking single state instances instead of full target trajectories. In this paper, we propose an enhanced version of the PHD filter to propagate labels in time without compromising the filter's performance and also to discriminate surviving and appearing objects in each frame.

This paper is divided into five sections. We discuss popular object detection and tracking approaches used in satellite images in **Section 2**. We discuss the proposed method in **Section 3** where we present the object detection and object tracking approaches. We show results for a challenging dataset in **Section 4** and we discuss the conclusion and future work in **Section 5**.

RELATED WORK

While object detection and tracking are related, for sake of simplicity, we divide our literature review into two categories composed of object detection and tracking applied to satellite images.

Object Detection

Static Image Object Detection

Static image object detection methods rely on spatial information to extract features and obtain object segmentation masks or bounding boxes. Popular approaches include Faster-RCNN, proposed by Ren et al. (2015), YOLO, proposed by Redmon et al. (2016), Retina-Net, proposed by Lin et al. (2017). Although these works obtain remarkable results across several benchmarks, their performance decreases significantly when tested with small objects or weakly labeled datasets such as in remote sensing images. In fact, Acatay et al. (2018) presented a comprehensive review and the drawbacks from using the base Faster-RCNN, YOLO, and Single Shot Detectors (SSD) on aerial images. Several researchers approached satellite object detection with modified

appearance-based object detector approach for remote sensing images. For example, Ren et al. (2018) proposed a modified Faster-RCNN to detect small objects in satellite images by modifying the anchor boxes, adding skipped connections, and including contextual information. However, this method focuses on capturing relatively large objects such as planes and large ships. Similarly, Qian et al. (2020) proposed a modified version of Faster-RCNN with a new architecture, new metric, and loss to optimize the training of small objects bounding boxes that do not overlap.

Motion-Only Object Detection

Motion-based detections consist principally in background subtraction and frame differencing. A popular approach is to model backgrounds with Gaussian distributions and parameters derived from observations. This model has been extensively expanded such as with the method proposed by Stauffer and Grimson (2000) to use Gaussian mixture models (GMM) instead of a single Gaussian distribution, or the work proposed by Han and Davis (2012) which uses kernel density estimators (KDE) to estimate background distributions and support vector machines (SVM) to discriminate objects. Yang et al. (2016) proposed ViBe, an approach that updates the background estimation persistently and locally by using random selection. However, background subtraction methods generate noisy results when dealing with long sequences of images with a moving imaging system such as a satellite or drone.

Similarly, frame differencing has shown robustness across several methods. For example, Teutsch and Grinberg (2016) proposed to use frame differencing together with numerous post-processing filters to perform object detection in WAMI images. Also, Ao et al. (2020) proposed to use frame differencing together with noise estimation and shape-based filters to extract objects. These approaches obtain reasonable results but they rely on complex hand-crafted post-processing steps that can be hardly adapted to different noise sources.

Motion models are often robust and computationally lightweight; however, their performance relies heavily on frame registration. Small errors in frame registration or illumination changes often lead to large errors in motion-based object detection.

Spatio-Temporal Convolutional Neural Networks

State-of-the-art methods aim to combine approaches from both appearance and motion to improve object detection. Generally, these methods use CNNs that take into account both motion and appearance information to extract object locations. For instance, LaLonde et al. (2018) proposed ClusterNet and FoveaNet, a two-stage approach for exploiting spatial and temporal data in small object detection. They use five consecutive frames as input to an under-sampling network to create clusters of object locations (ClusterNet), and then they use a region specialized network (FoveaNet) to refine the outputs of the first network. Also, Canepa et al. (2021) proposed T-Rex Net, a network that uses frame differencing as inputs to the network to improve small object detection performance. Sommer et al. (2021) proposed

an appearance-based and motion-based object detector by combining two networks, one to estimate moving objects locations, and one to extract image features. These methods showed promising results for ultra high resolution datasets such as the WPAFB 2009 (AFRL (2009)) dataset which contains a resolution of up to 0.25 cms/pixel; however, these approaches cannot be directly applied to lower resolution data such as at 1m/pixel as the target features are lost and performing undersampling could miss the small targets.

Object Tracking Feature Tracking

Common tracking approaches for satellite images include the use of correlation filters and expansions to this approach. Correlation filters find similarities between frames to responses to learned filters and match the coordinates and responses. For example, Du et al. (2017) employed a correlation filter combined with three frame difference to track objects in satellite images, and Xuan et al. (2020) used correlation filters together with linear equations to track objects even under occlusions. While these methods are robust for object tracking, they rely on initialization and are normally adapted to track single objects.

Joint Tracking and Detection

Numerous state-of-the-art tracking methods are deep learning-based and learn to jointly detect and track objects. For instance Bergmann et al. (2019) proposed Tracktor++ to use a CNN to perform both object detection and tracking. Similarly, Feichtenhofer et al. (2017) proposed Track to Detect and Detect to Track to regress both bounding boxes for the object dimensions and for the object temporal displacement. Among robust CNN tracking approaches are attention-based methods such as Patchwork, proposed by Chai (2019), which consists in using an attention mechanism to predict the location of an object in future frames. Jiao et al. (2021) created a survey of novel generation deep learning-based techniques used for object tracking, where methods mostly depend on correlating learned features in time.

Track by Detection

Tracking by detection approaches include SORT, proposed by Bewley et al. (2016) and its extension DeepSORT, proposed by Wojke et al. (2017). SORT consists of an online multiple object tracker (MOT) that uses multiple Kalman filters for tracking and the Hungarian algorithm (Kuhn and Yaw (1955)) for data association, and DeepSORT is an extension that uses object features similarity to modify the data association step. These approaches obtain state-of-the-art results in remarkable computational times; however, due to their pragmatic approach, they do not process a unified multi-object data uncertainty model that can model ambiguous target paths.

Reid (1979) proposed a Bayesian framework named multiple hypothesis tracking (MHT) and Fortmann et al. (1980) proposed the joint probabilistic data association (JPDA). These approaches consider unified probabilistic models and propagate the data

association combinatoric metrics on time. However, these filters are often slow due to the complicated data association process and the exponential increase of complexity with time.

Finally, the random finite set (RFS) framework and random finite set statistics proposed by Mahler (2007) propose an attractive track-by-detection paradigm without compromising the computational time. Among popular trackers are the PHD filter, proposed by Mahler (2003), the cardinally PHD filter, presented by Vo et al. (2006), and novel methods such as the Labelled Multi-Bernoulli Filter, developed by Vo and Vo (2013) and its computationally efficient version Vo et al. (2017). In our case, we propose an extended version of PHD filter due to its robust results and significant computational advantages.

PROPOSED APPROACH

In this paper, we extend the work proposed by Aguilar et al. (2021) which employs a 3-frame difference algorithm to approximate target locations and a patch-based CNN to refine detections. We extend this work by 1) concatenating the frame difference response to the input for the neural network, 2) by performing a tile-based patch selection rather than coordinate-based patch selection. Finally, we use an extended version of the PHD filter, a Bayesian multi-object tracker, to convert frame-wise object detections into track hypothesis.

Motion Aware CNN for Object Detection Motion Detector

We estimate object motion by finding differences between consecutive frames and adding their responses to create a likelihood $3FD_k(i, j)$, where $(i, j) \in \mathbb{R}^2$ are the pixel coordinates and $k \in \mathbb{N}$ is the time index. This process is summarized in the equations:

$$\Delta I_k(i, j) = I_k(i, j) - I_{k-1}(i, j) \quad (1)$$

$$3FD_k(i, j) = |\Delta I_k(i, j)| + |\Delta I_{k+1}(i, j)| \quad (2)$$

Sequentially, we binarize the $3FD_k(i, j)$ response with a frame-adaptive threshold to obtain rough object location estimates by applying the formulas:

$$G(i, j) = \begin{cases} 1 & 3FD_k(i, j) > T_k \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$T_k = c * \max(3FD_k(i, j)) \quad (4)$$

Where $c \in (0, 1)$ is a percentage-based threshold hyperparameter and is used to remove noisy 3-frame difference responses. We chose c by performing grid search and choosing values of c that would favor higher detection rates, in particular we set $c = 15\%$ for all the experiments shown in Section 4. The 3-frame difference approach yields good object location estimates but it fails to perform shape regularization, detect low contrast objects, and detect slow-moving targets. Therefore, we complement the frame difference response with Faster RCNN (Ren et al. (2015)). This addition helped to filter false positives, discriminate nearby objects, and increase the detection rate.

We use the frame difference for two objectives: to reduce the target search space and to feed this information to the neural network. We begin by tiling the image starting at the origin and using the response $G(i, j)$ to find patches with moving objects. The patch-based approach rather than full image-based approach presents significant advantages: it contributes to focusing on relevant areas rather than the whole image space, and it contributes to training a network with scarce data because one image can yield several training patches. We extract patches that contain object hypothesis (given by the frame difference response) and refine the detections using Faster RCNN.

We modify the inputs to the traditional Faster RCNN by including three consecutive frames (shown in Figure 2B) and by concatenating these images to the frame difference response (shown in Figure 2C). This step is different from our previous approach Aguilar et al. (2021) where we used only one patch as input for the CNN. Using three frames together with the frame-difference response provides an additional cue for the network to detect moving objects (denoted by cyan and yellow colors in the concatenated inputs in Figure 2D). Figure 2E shows that our approach detects very small moving objects such as motorcycles that would have been missed by using only one frame as input. The addition of motion information improves detection rates for small moving objects and also reduces false positives of vehicle-looking static objects. Section 4.3 shows further details in the effect of using three frames and frame difference as opposed to one frame.

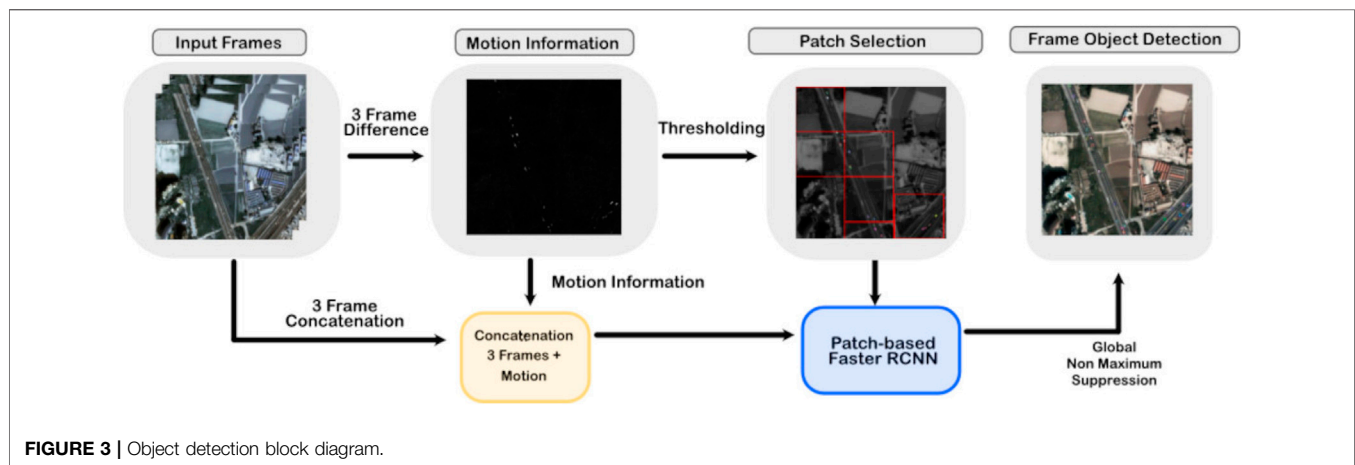
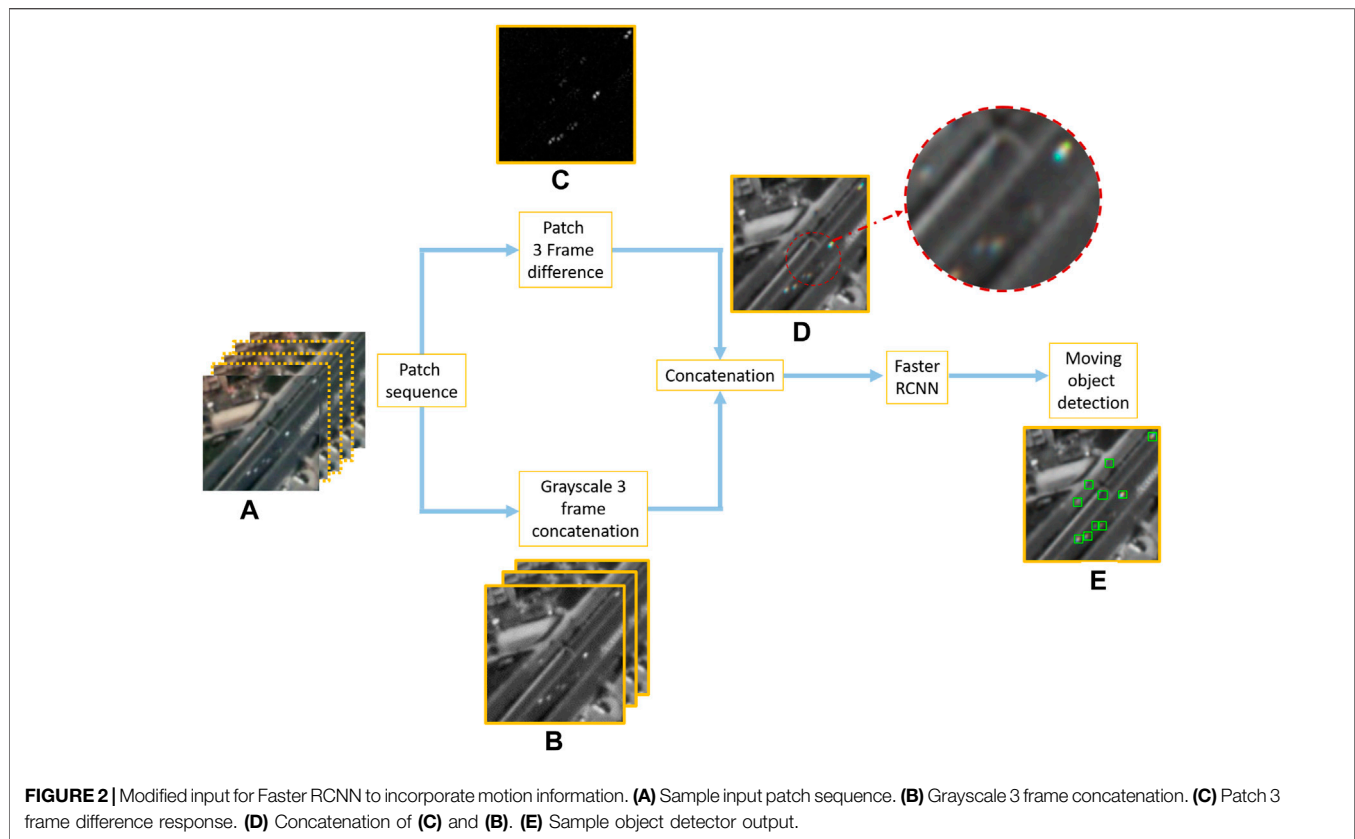
Finally, we merge the patch results by performing global non-maximum suppression and applying the respective offset to the patch-based detections. The whole object detection process is summarized in Figure 3.

Object Tracking With the GM-PHD Filter Motion and Measurement Modeling

We define the state vector for the j th target at time k as $\mathbf{x}_k^j = [p_x, p_y, v_x, v_y, w, h]^T$ where $p_x, p_y \in \mathbb{R}$ denote the target x and y position, $v_x, v_y \in \mathbb{R}$ denote the target velocity components, and w, h denote the target width and height respectively. We assume the target motion is linear and adopt the constant velocity (CV) model with Gaussian noise. Hence we assume the targets evolve according to the equation: $f_{k|k-1}(\mathbf{x}_k^j | \mathbf{x}_{k-1}^j) = N(\mathbf{x}_k^j; F_k \mathbf{x}_{k-1}^j, Q_{k-1})$ where Q_k is the motion covariance and F_k is the transition matrix defined as:

$$F_k = \begin{bmatrix} 1 & 0 & \tau & 0 & 0 & 0 \\ 0 & 1 & 0 & \tau & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (5)$$

Where τ is a hyperparameter related to the sampling frequency. Similarly, we define the i th measurement at time k as $\mathbf{z}_k^i = [p_x, p_y, w, h]^T$, where $p_x, p_y, w, h \in \mathbb{R}$ denote the x, y coordinates, width and height respectively. We assume the noisy and Gaussian measurements in the form of



$g_k(\mathbf{z}_k^i | \mathbf{x}_k) = \mathcal{N}(\mathbf{z}_k^i; H_k \mathbf{x}_k, R_k)$, where R_k is the measurement noise covariance and H_k denotes the measurement matrix defined as:

$$H_k = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (6)$$

PHD Filter

We aim to estimate the multi-target states from a sequence of possibly noisy or cluttered measurements. We approach this task

by using the random finite set (RFS) statistics defined by Mahler (2007). This setup provides a Bayesian formulation for modeling objects and observations as set-valued random variables. Specifically, the collection of targets state at time k is defined by $\mathbf{X}_k = \{\mathbf{x}_k^1, \mathbf{x}_k^2, \dots, \mathbf{x}_k^{N_k}\}$, where \mathbf{x}_k^j denotes the j th target state vector at time k , and N_k denotes the cardinality of \mathbf{X}_k . Similarly, the measurements at frame k are defined by the RFS $\mathbf{Z}_k = \{\mathbf{z}_k^1, \mathbf{z}_k^2, \dots, \mathbf{z}_k^{M_k}\}$, where M_k denotes the cardinality for the measurement RFS at time k . Our objective is to model the multi-target state posterior of \mathbf{X}_k given all the previous measurements $\mathbf{Z}_{1:2}, \dots, \mathbf{Z}_k$, namely we aim to find $p_{k|1:k}(\mathbf{X}_k | \mathbf{Z}_{1:k})$.

The PHD filter provides an approximation to the optimal multi-target filter by modeling the posterior $p_{k|1:k}(\mathbf{X}_k|\mathbf{Z}_{1:k})$ as a Poisson random finite set and by recursively propagating its first-order statistical moment, called probability hypothesis density (PHD) function. The PHD filter achieves this task by iteratively performing a two step process: the prediction step and the update step.

The prediction step consists on estimating the PHD function $D_{k|1:k-1}(\mathbf{X}_k|\mathbf{Z}_{1:k-1})$ at time k given only previous measurements, abbreviated as $D_{k|k-1}(x)$. The update step consists on estimating the posterior PHD $D_{k|1:k}(\mathbf{X}_k|\mathbf{Z}_{1:k})$ using the predicted information and the new measurement obtained at time k and is abbreviated to $D_{k|k}(x)$.

The GM-PHD Filter

The Gaussian Mixture PHD Filter (GM-PHD), proposed by Vo and Ma (2006), is a close form solution to the PHD recursion and its convergence properties are analyzed by Clark and Vo (2007). The GM-PHD relies on the assumptions of linear Gaussian motion and measurement models explained in Section 3.2.1. Additionally, the GM-PHD assumes the form of the posterior at the previous time frame, $D_{k-1|k-1}(x)$, has the form of a Gaussian mixture given by:

$$D_{k-1|k-1}(x) = \sum_{j=1}^{J_{k-1|k-1}} \omega_{k-1|k-1}^j \mathcal{N}(x; \mathbf{m}_{k-1|k-1}^j, \mathbf{P}_{k-1|k-1}^j) \quad (7)$$

Where $J_{k-1|k-1}$ is the number of Gaussian components and $\omega_{k-1|k-1}^j$, $\mathbf{m}_{k-1|k-1}^j$, $\mathbf{P}_{k-1|k-1}^j$ are the weight, mean, and covariance for each GM component in the posterior distribution at time $k-1$.

The GM-PHD filter estimates the predicted $D_{k|k-1}(x)$ and updated $D_{k|k}(x)$ PHDs with Gaussian mixtures. The closed form solution for the GM-PHD prediction step is given by the equation:

$$D_{k|k-1}(x) = \lambda(x) + p_s \sum_{j=1}^{J_{k|k-1}} \omega_{k|k-1}^j \mathcal{N}(x; F_k \mathbf{m}_{k-1|k-1}^j, Q + F_k \mathbf{P}_{k-1|k-1}^j F_k^T) \quad (8)$$

Where F_k and Q are respectively the transition and motion covariance matrices defined in Section 3.2.1, p_s is the survival probability, and $\lambda(x)$ is the birth RFS intensity which will be described in Section 3.2.4. Finally, we update the GM-PHD posterior following the equation:

$$D_{k|k}(x) = (1 - p_D) D_{k|k-1}(x) + \sum_{z \in Z_k} \omega_{k|k}^j(z) \mathcal{N}(x; \mathbf{m}_{k|k}^j(z), \mathbf{P}_{k|k}^j(z)) \quad (9)$$

Where $D_{k|k-1}(x)$ denotes the predicted GM components and p_D is the probability of detection. The terms $\mathbf{m}_{k|k}^j(z)$ and $\mathbf{P}_{k|k}^j(z)$ represent the updated component mean and covariance and are defined as:

$$\mathbf{m}_{k|k}^j(z) = \mathbf{m}_{k|k-1}^j + \mathbf{K}_k^j [z - H_k \mathbf{m}_{k|k-1}^j] \quad (10)$$

$$\mathbf{P}_{k|k}^j = [\mathbf{I} - \mathbf{K}_k^j H_k] \mathbf{P}_{k|k-1}^j \quad (11)$$

$$\mathbf{K}_k^j = \mathbf{P}_{k|k-1}^j H_k^T [H_k \mathbf{P}_{k|k-1}^j H_k^T + \mathbf{R}]^{-1} \quad (12)$$

The updated component weight $\omega_{k|k}^j(z)$ is defined as:

$$\omega_{k|k}^j(z) = \frac{p_D \omega_{k|k-1}^j l_k^j(z)}{\kappa_k(z) + p_D \sum_{i=1}^{J_{k|k-1}} \omega_{k|k-1}^i l_k^i(z)} \quad (13)$$

Where $\kappa_k(z)$ denotes the clutter process intensity (modeled with a Poisson Random Finite Set) and $l_k^j(z)$ denotes the target-measurement association likelihood defined as:

$$l_k^j(z) = \mathcal{N}(z; H_k \mathbf{m}_{k|k-1}^j, \mathbf{S}_k^j) \quad (14)$$

$$\mathbf{S}_k^j = \mathbf{R}_k + [H_k \mathbf{P}_{k|k-1}^j H_k^T] \quad (15)$$

We estimate the filter's inference cardinality by adding all the weights in the posterior PHD and we apply merging and pruning for components with very small weights in order to preserve the computational advantages of the PHD filter.

PHD Filter Enhancements

We use a measurement-driven approach to estimate the birth $\lambda(x)$ intensity. Specifically, we use an adapted measurement classification similar to Fu et al. (2018) to discriminate measurements into surviving measurements, \mathbf{Z}_k^s and birth measurements \mathbf{Z}_k^b . During each iteration, we use the Hungarian algorithm to find the optimal matching between the new measurement set, \mathbf{Z}_k , and the set of spatial components of the predicted GM-PHD: $\{H \mathbf{m}_{k|k-1}^j\}_{j=1,2,\dots,J_{k|k-1}}$. If the distance between a measurement and a predicted component mean is less than a threshold, we classify the target as surviving measurement, otherwise, all the unassigned measurements are classified as a birth-proposal.

We implement the label preserving structure proposed by Panta et al. (2009) as the original GM-PHD filter does not account for target labels or past trajectories. This extension initializes a label for every Gaussian mixture component and propagates the label in time without affecting the filter performance. Each birth step initializes new labels for each birth component and the labels are tracked during the prediction and the data association step. These advantages contribute to keeping track of possible target trajectories without compromising the filter computational load.

RESULTS

Evaluation Metrics

We evaluate our methods by using object detection and object tracking metrics. We use ground truth annotations in the form of $\mathbf{o}_k = \{o_1, o_2, \dots, o_N\}$, where k is the frame number and $o_i = (p_x, p_y, l)$ is a single annotated object at coordinates (p_x, p_y) with associated label l . We let an estimated target be $\hat{o}_i = (\hat{p}_x, \hat{p}_y, \hat{l})$, where \hat{p}_x, \hat{p}_y are the location components from the GM-PHD filter inferred object state, and \hat{l} is the inferred associated label. At

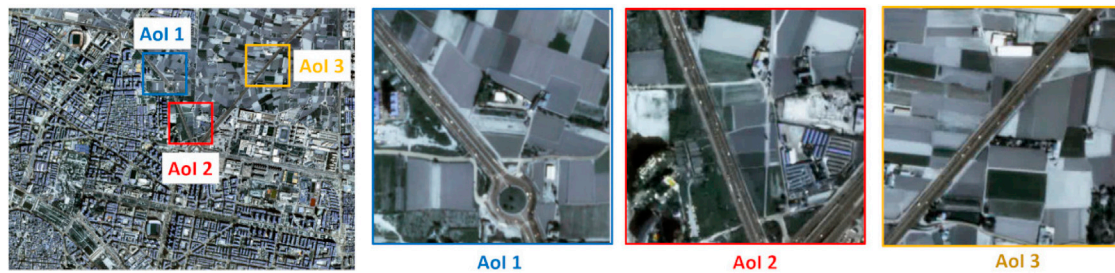


FIGURE 4 | Areas of interest (Aols) for method evaluation.

TABLE 1 | Ablation studies.

	Precision	Recall	F1
Faster RCNN	56.73	72.76	61.69
Faster RCNN + Motion Information	69.46	73.33	70.05
Patch-Based Faster RCNN	69.06	70.96	69.22
Patch-Based Faster RCNN + Motion Information	78.13	70.40	76.14

TABLE 2 | Average F1 scores for different patch sizes.

Patch Size	32 × 32	64 × 64	128 × 128	256 × 256	512 × 512 (full image)
F1 score	51.66	70.66	76.14	72.66	70.05

every frame, we match the set of detected targets with the set of ground truth objects, we label an estimated target \hat{o}_i as true positive (*TP*) if is within five pixels away from an unmatched ground truth object, otherwise, we label the object as a false positive (*FP*). Similarly, we label any ground truth target that has not been matched to an estimated target as a false negative (*FN*). Finally, we call a track an identity switch (*IDS*) if its object track hypothesis is associated with more than one ground truth label l .

Object Detection Metrics

For object detection, we report the *F1* score which is a widely accepted evaluation metric to evaluate the quality of the detector. The *F1* score is defined as:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (16)$$

Where precision denotes the ratio of relevant hypothesis proposed by the object detector and is defined as:

$$Precision = \frac{TP}{TP + FP} \quad (17)$$

Recall denotes the percent of correctly detected objects in comparison to the total number of available objects and is defined as:

$$Recall = \frac{TP}{TP + FN} \quad (18)$$

We report these metrics as percentages, where the best score is of 100 and the worst score is 0. Additionally, we present a precision-recall curve to show the robustness of the proposed approach over the possible parameter ranges and to show its improved performance over possible competing approaches. We use these tests to choose the parameters for running the *F1* score for each listed method.

Object Tracking Metrics

We also report tracking metric ClearMOT, proposed by Bernardin and Stiefelhausen (2008), as it has become a popular and robust metric for tracking algorithms. We report the multiple object tracking accuracy (MOTA) which evaluates the quality of the recovered tracks. It considers FPs, FNs, and identity switches (IDSs). The MOTA score is defined as:

$$MOTA = 1 - \frac{\sum_{k=1}^N (FN_k + FP_k + IDS_k)}{\sum_{k=1}^N GT_k} \quad (19)$$

Where N refers to the number of frames, and FN_k , FP_k , IDS_k , GT_k refers to the false negatives, false positives, identity switches and number of ground truth objects at frame k respectively. The MOTA score has a range in $(-\infty, 1)$, where negative values report poor performances, and one is the best possible score. In this work, we report the scores as a percentages to keep consistency with literature. We also report the multiple object tracking precision (MOTP), which considers the average distance error between the detected objects and the ground truth objects. The MOTP is defined as:

$$MOTP = \frac{\sum_{k=1}^N \sum_{i=1}^{c_k} d_{i,k}}{\sum_k c_k} \quad (20)$$

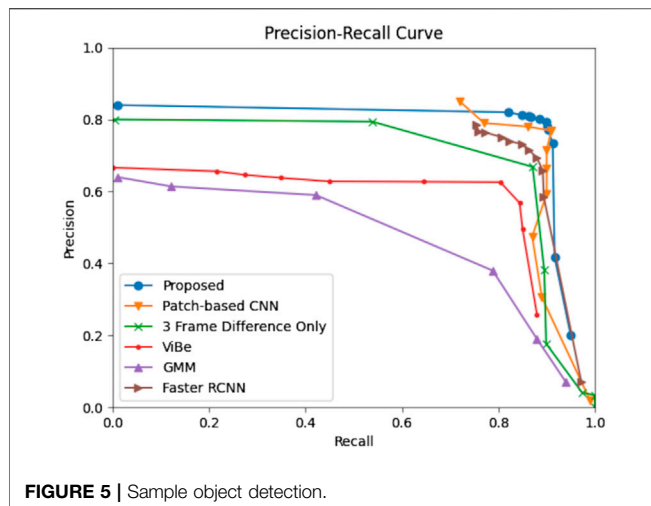


FIGURE 5 | Sample object detection.

TABLE 3 | Object detection metrics.

AoI	Detector	Precision	Recall	F1
1	3 Frame-based, Ao et al. (2020)	85.8	79.3	82.42
	ViBe, Yang et al. (2016)	80.9	63.8	71.33
	GMM, Wren et al. (1997)	78.9	38.3	51.57
	Faster-RCNN, Ren et al. (2015)	80.6	75.1	77.76
	Patch-based-CNN, Aguilar et al. (2021)	91.5	76.9	83.57
	Proposed Object Detection	90.2	80.9	85.32
2	3 Frame-based, Ao et al. (2020)	70.0	73.1	71.52
	ViBe, Yang et al. (2016)	41.1	65.1	50.38
	GMM, Wren et al. (1997)	61.0	65.1	62.95
	Faster-RCNN, Ren et al. (2015)	27.2	66.9	38.65
	Patch-based-CNN, Aguilar et al. (2021)	50.2	70.8	58.76
	Proposed Object Detection	71.3	74.5	72.84
3	3 Frame-based, Ao et al. (2020)	62.3	48.7	54.68
	ViBe, Yang et al. (2016)	74.4	56.9	64.47
	GMM, Wren et al. (1997)	35.9	54.9	43.43
	Faster-RCNN, Ren et al. (2015)	62.4	76.3	68.68
	Patch-based-CNN, Aguilar et al. (2021)	65.5	65.2	65.33
	Proposed Object Detection	72.9	67.8	70.26

Where c_k refers to the number of correctly detected objects at frame k and $d_{i,k}$ denotes the distance between a ground truth object and the detected hypothesis. The MOTP score is in the range $[0, \infty)$ where 0 denotes the perfect score and large values denote worse performances.

Finally, we report track quality measures in a similar format to Dendorfer et al. (2021). We call a trajectory mostly tracked (MT) if we can persistently track at least 80% of its path. Similarly, we call a trajectory mostly lost (ML) if we can track 20% or less of its ground truth trajectory. We report these scores as percentages where larger percentages of MT scores denote better performances but larger percentages of ML scores denote worse performances.

Experiment Set up

For evaluation purposes, we use the CGSTL dataset, available at <https://mall.charmingglobe.com>. This dataset contains a

video of the city of Valencia, Spain, recorded on 7 March 2017, by the Jilin-1 satellite. Its spatial resolution is 1 m/pixel and the video spans 12 kms², with a size of 3,071 × 4,096 pixels. The video contains 580 frames and represents 29 s of video imaged at 20 frames per second. The labels were provided by Ao et al. (2020) and contain the (x, y) object center coordinates, the width, and height of the object bounding boxes. The provided ground truth contains strong labeling for only moving targets in three areas of interest (AoI) of size 500, ×, 500 pixels (shown in Figure 4). The approximate coordinate location for each area are AoI 1 [520, 1616], AoI 2 [1074, 1895] and AoI 3 [450, 2810] with respect to the first frame. Additionally, we performed image stabilization (ORB(Rublee et al. (2011))) to compensate for the satellite motion during the recorded video. Finally, only one every ten frames is labeled (58 total labeled frames), hence, we used the stabilization procedure and linear interpolation between frames to fill the label subsampling. The stabilization procedure has a significant impact on object detection, object tracking, and score evaluation across all 580 frames as these methods depend on linear object motion and static background. It is worth mentioning we improve the stabilization procedure over our previous work (Aguilar et al. (2021)) by using the Python OpenCV implementation of ORB(Rublee et al. (2011)); hence our ‘true positive’ distance criteria is set to five pixels rather than 20 pixels as in our previous work.

All of the AoIs contain highways and moving vehicles at high speed. AoI one contains a roundabout, where objects reduce their velocity and travel in clusters. AoI two contains a highway next to farming structures that create numerous false positives for both motion and appearance-based object detectors. AoI three contains a highway with objects moving at high speeds. It is worth mentioning all AoIs contain numerous motorcycles and very small objects that are often missed in the ground truth annotations due to the difficulty of labeling such objects at such low image resolution. For each AoI, we trained the network using the other two AoIs as training data due to the ground truth data scarcity. We trained the networks using extracted patches of size 128 × 128 centered at ground truth objects and we augmented data by using patch vertical and horizontal flips, and random translations. We used the Pytorch implementation for Faster-RCNN using a pre-trained ResNet50 proposed by He et al. (2016) as backbone for feature extraction. The networks were trained using an NVIDIA QUADRO using stochastic gradient descend as optimizer with a learning rate of $lr = 0.005$ and a weight decay of 0.0005.

Ablation Studies

We perform ablation studies to investigate the impact of using patch-based inference and the impact of including motion information on object detection quality. We report the F1 scores for our method using path-selection only, motion-information only, and patch-selection and motion-information combined. We evaluate these scores across all

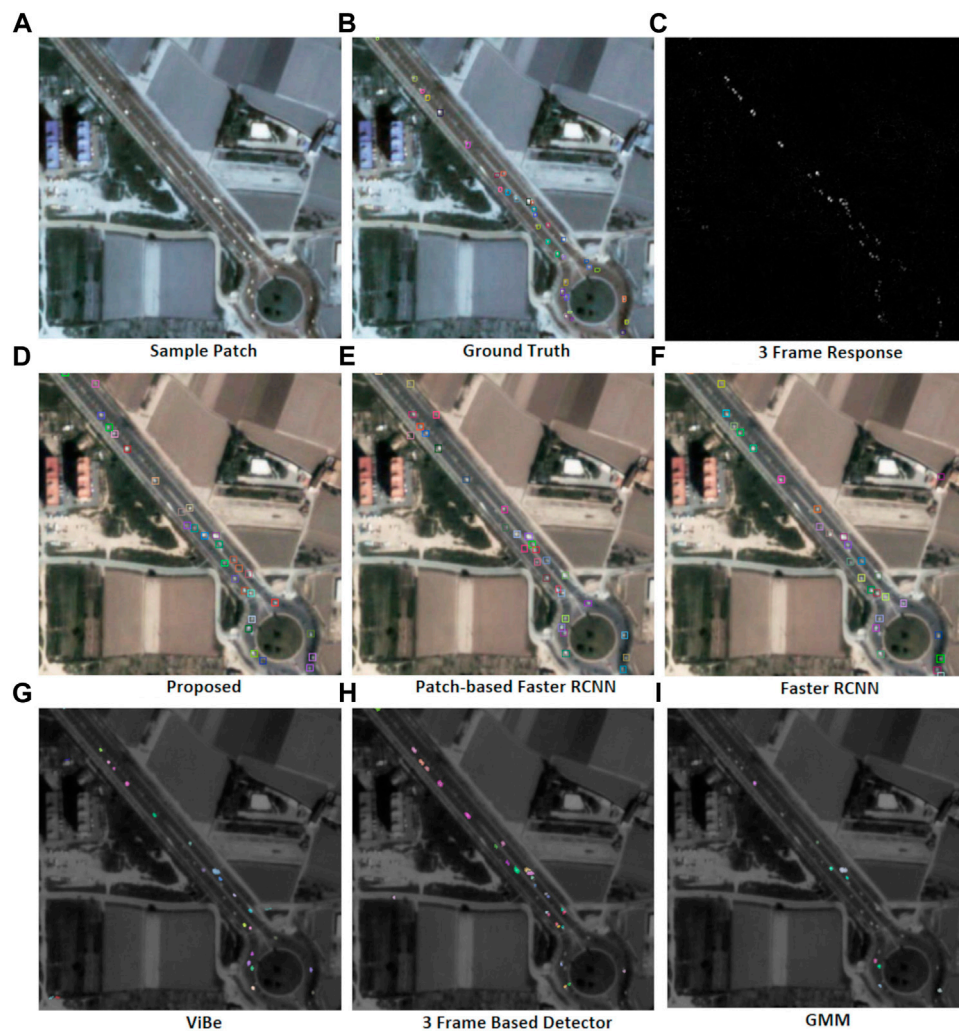


FIGURE 6 | Sample object detection in sub-region of AoI 1. First row: **(A)** sample patch . **(B)** Ground truth bounding boxes. **(C)** 3-frame difference response. Second row: **(D)** Proposed method. **(E)** Output of patched-based Faster RCNN. **(F)** Output of Faster RCNN. Third row: **(G)** Output of ViBe. **(H)** Output of 3 frame based detector. **(I)** Output of GMM detector.

AoIs and report the average precision, recall, and the *F1* scores for each combination.

Patch-Based Inference

We test the effect of using a patch-based method by comparing a full-image and patch-based inference with Faster RCNN. **Table 1** shows that a full-image Faster RCNN obtains a *F1* metric of 61.69 but using a patch-based Faster RCNN increased the *F1* score to 69.22. The patch-based approach outperforms Faster RCNN in the precision score because it reduces the search space to areas with moving objects and decreases the ratio of *FPs*. This result is expected as satellite images contain numerous blob-looking objects that yield false positives and Faster RCNN alone would detect the objects as vehicles. These results are developed further and shown numerically and visually in **Section 4.4**. Additionally, we test the effect of varying the patch size by evaluating average object detection metrics

using patch sizes of 32, 64, 128, 256, and 512 (full image). The size effects for the patch selection are depicted in **Table 2**, where the highest *F1* score is obtained for the patch size of 128×128 pixels. During our experiments, we concluded that the patch size of 128×128 focuses the CNN to smaller regions while preserving contextual information. In fact, a patch size of 64×64 yielded numerous false positives from static objects with white-blob appearance. On the contrary, large patch sizes such as 256×256 and 512×512 obtained large numbers of misdetections due to the small object size in comparison with the field of view.

Motion-Based Inference

We investigate the effect of including motion information by testing the full-image Faster RCNN combined with motion information. We achieve this task by feeding three consecutive frames concatenated with the three frame difference algorithm to Faster

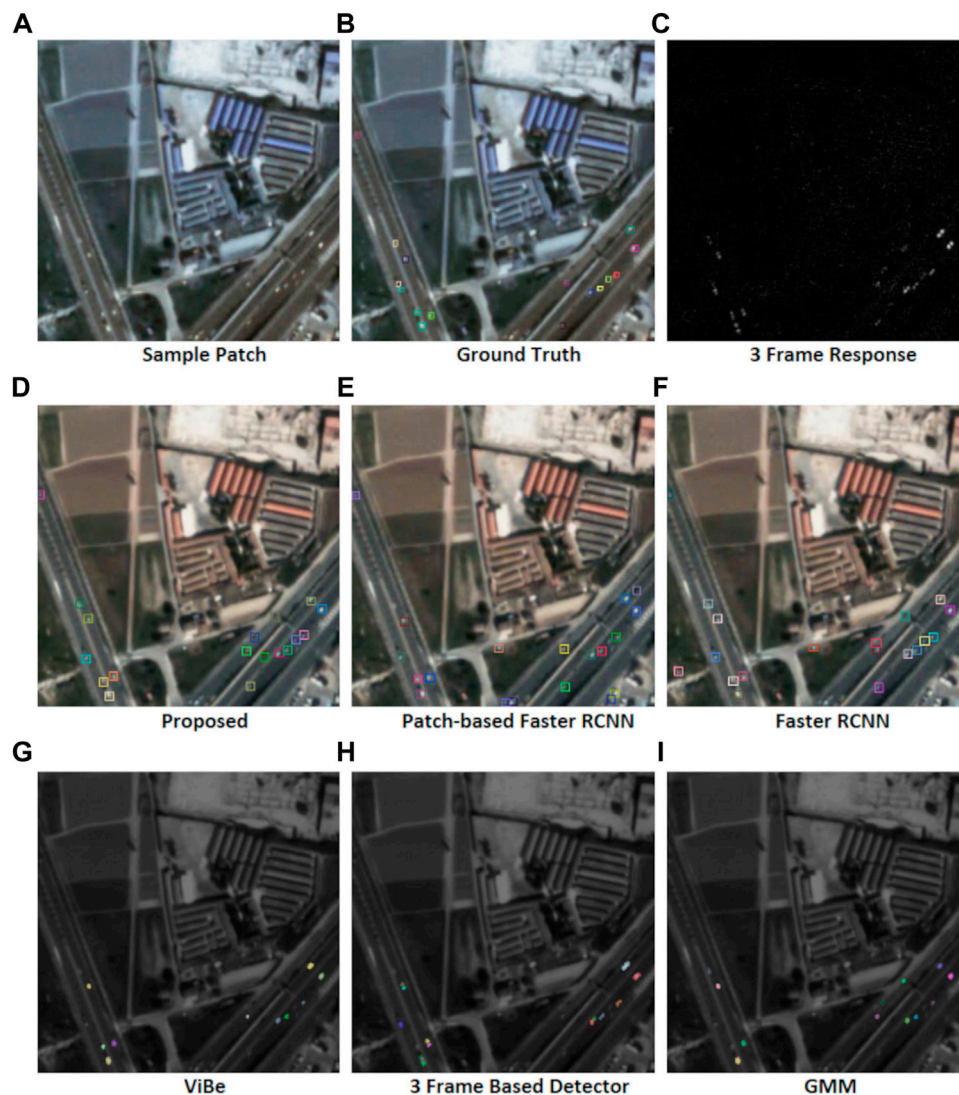


FIGURE 7 | Sample object detection in sub-region of Aol 2. First row: **(A)** sample patch . **(B)** Ground truth bounding boxes. **(C)** 3-frame difference response. Second row: **(D)** Proposed method. **(E)** Output of patched-based Faster RCNN. **(F)** Output of Faster RCNN. Third row: **(G)** Output of ViBe. **(H)** Output of 3 frame based detector. **(I)** Output of GMM detector.

RCNN. **Table 1** shows that including motion information for the full-image Faster RCNN improves the *F1* score from 61.69 to 70.05. This improvement occurs due to the increase in the precision score, from 56.73 to 69.46. Our results show that including motion information also helps Faster RCNN to filter non-moving objects in a similar fashion to using a patch-based approach.

Motion and Patch-Based Inference

Finally, we test the effects of including motion information and a patch-based approach to the original Faster RCNN. **Table 1** shows that adding both motion information and patch-based inference increased the *F1* score of the original Faster RCNN by 6 and 7% respectively. The combined effect of using a patch inference and including motion information reduced the false-positive ratios

further, thus, increasing the precision score from 69.46 to 69.06 to 78.13. It is worth noting that neither the addition of motion or a patch-based approach contributed to increasing the recall score. In fact, full-image Faster RCNN obtains higher recall values than the proposed approach at the cost of increasing the number of false detections. These results suggest further development explained in **Section 5**.

Object Detection Evaluation

We evaluate the proposed object detector using the *F1* metric mentioned in **Section 4.1.1** and we compare its performance with five competing object detectors: custom 3-frame difference proposed by Ao et al. (2020), background subtraction using Gaussian mixture models proposed by Wren et al. (1997),

TABLE 4 | Tracking Metrics for AoI 1. *Denotes ground truth measurements used for calibration and filter-only testing.

AoI	Tracker	Detector	F1	MOTA	MOTP	MT	ML
1	SORT	Ground Truth Detections (Calibration)*	99.4*	99.1*	0.91*	63*	0*
		3 Frame-based, Ao et al. (2020)	50.4	27.2	2.75	7	23
		ViBe, Yang et al. (2016)	65.4	40.5	2.50	27	19
		GMM, Wren et al. (1997)	49.3	30.7	2.57	13	33
		Faster-RCNN, Ren et al. (2015)	70.3	48.2	2.94	23	14
		Patch-based-CNN, Aguilar et al. (2021)	44.9	19.6	2.93	1	31
		Proposed Object Detection	78.8	63.0	2.34	34	11
	GLMB	Ground Truth Detections (Calibration)*	94.95*	85.1*	1.80*	63*	1*
		3 Frame-based, Ao et al. (2020)	71.14	36.3	2.02	35	6
		ViBe, Yang et al. (2016)	67.02	37.4	1.53	32	13
		GMM, Wren et al. (1997)	49.90	22.1	1.61	15	30
		Faster-RCNN, Ren et al. (2015)	66.76	30.8	2.03	29	13
		Patch-based-CNN, Aguilar et al. (2021)	73.86	46.9	1.93	33	11
		Proposed Object Detection	83.8	66.6	1.19	35	12
	GM-PHD	Ground Truth Detections (Calibration)*	94.5*	89.7*	0.19*	58*	3*
		3 Frame-based, Ao et al. (2020)	69.9	47.2	2.17	24	11
		ViBe, Yang et al. (2016)	63.0	35.6	1.92	21	23
		GMM, Wren et al. (1997)	48.5	27.0	1.98	14	33
		Faster-RCNN, Ren et al. (2015)	71.7	47.7	2.36	31	22
		Patch-based-CNN, Aguilar et al. (2021)	76.1	56.7	2.40	31	17
		Proposed Object Detection	81.9	64.3	1.49	46	8

TABLE 5 | Tracking Metrics for AoI 2. *Denotes ground truth measurements used for calibration and filter-only testing.

AoI	Tracker	Detector	F1	MOTA	MOTP	MT	ML
2	SORT	Ground Truth Detections (Calibration)*	99.6*	99.5*	0.857*	61*	0*
		3 Frame-based, Ao et al. (2020)	54.81	26.6	2.14	17	26
		ViBe, Yang et al. (2016)	50.50	-18.2	2.32	38	17
		GMM, Wren et al. (1997)	74.54	54.1	2.26	35	13
		Faster-RCNN, Ren et al. (2015)	53.28	-22.1	2.38	32	18
		Patch-based-CNN, Aguilar et al. (2021)	42.33	15.1	2.72	6	22
		Proposed Object Detection	82.78	66.4	2.08	47	12
	GLMB	Ground Truth Detections (Calibration)*	97.94*	93.3*	1.543*	35*	0*
		3 Frame-based, Ao et al. (2020)	70.44	31.1	1.99	40	9
		ViBe, Yang et al. (2016)	49.69	-34.5	1.49	39	13
		GMM, Wren et al. (1997)	65.75	25.7	1.39	37	16
		Faster-RCNN, Ren et al. (2015)	72.50	31.8	1.41	50	4
		Patch-based-CNN, Aguilar et al. (2021)	57.60	33.3	1.88	39	5
		Proposed Object Detection	83.75	65.1	1.21	52	2
	GM-PHD	Ground Truth Detections (Calibration)*	98.7*	97.7*	0.18*	36*	0*
		3 Frame-based, Ao et al. (2020)	69.62	44.5	1.93	30	12
		ViBe, Yang et al. (2016)	44.86	-31.6	1.77	23	23
		GMM, Wren et al. (1997)	71.14	44.8	1.67	33	15
		Faster-RCNN, Ren et al. (2015)	50.00	-51.9	1.87	40	9
		Patch-based-CNN, Aguilar et al. (2021)	61.18	40.7	2.40	37	9
		Proposed Object Detection	82.61	64.1	1.58	47	3

ViBe, proposed by Yang et al. (2016), Faster RCNN, proposed by Ren et al. (2015) and the Patch-based object detector presented by Aguilar et al. (2021). We calibrate each method parameters by running a precision-recall curve on AoI 1, shown in **Figure 5**. We also show visual and numerical results for each AoI by reporting the precision, recall, and F1 scores for each competing method in **Table 3** and by showing sample object detection results in **Figure 6** and in **Figure 7**.

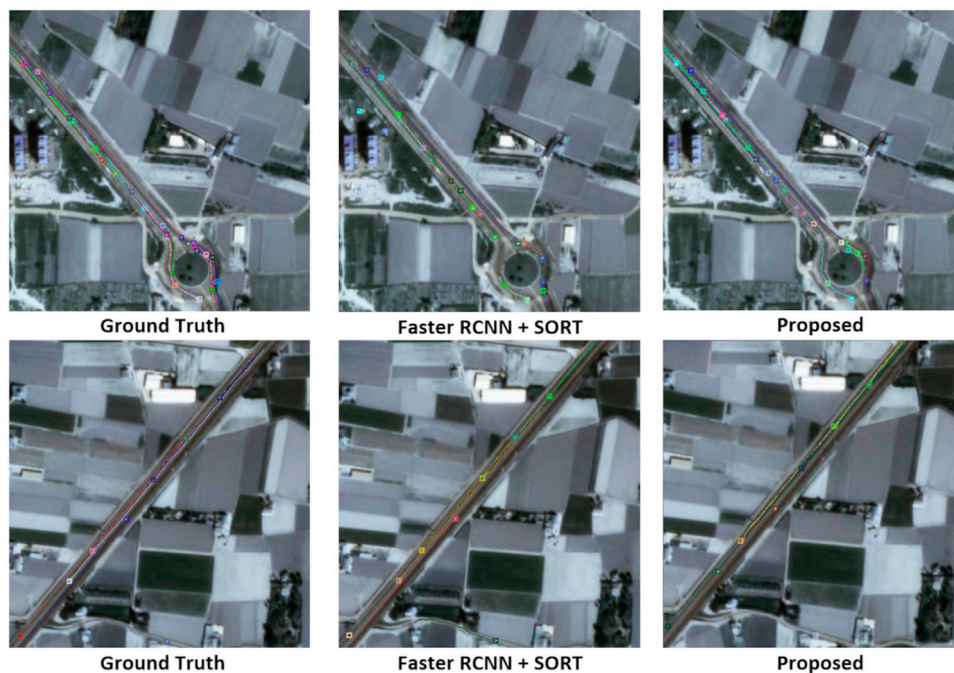
We varied the threshold and confidence parameters for 11 points in the range (0, 1) for the methods: 3-frame difference,

GMM, Faster RCNN, Patch-based RCNN, and the proposed approach. For ViBe, we changed the neighbor radius parameter: R for 11 points in the range (5, 50). **Figure 5** shows that our method is robust to parameter variations: it obtains better F1 scores across a diverse parameter range as the combination of appearance and time information increases true positives and decreases false negatives.

Figure 6 shows sample results for AoI 1. This area contains clusters of small moving objects at a roundabout and also presents

TABLE 6 | Tracking Metrics for AoI 2. *Denotes ground truth measurements used for calibration and filter-only testing.

AoI	Tracker	Detector	F1	MOTA	MOTP	MT	ML
3	SORT	Ground Truth Detections (Calibration)*	99.2*	98.4*	0.78*	46*	1*
		3 Frame-based, Ao et al. (2020)	58.20	39.2	2.46	14	25
		ViBe, Yang et al. (2016)	63.81	37.5	2.52	23	20
		GMM, Wren et al. (1997)	68.77	47.9	2.54	26	16
		Faster-RCNN, Ren et al. (2015)	70.3	48.2	2.94	23	14
		Patch-based-CNN, Aguilar et al. (2021)	37.44	14.9	3.39	0	37
	GLMB	Proposed Object Detection	73.49	53.7	1.72	23	13
		Ground Truth Detections (Calibration)*	99.60*	98.9*	1.43*	39*	0*
		3 Frame-based, Ao et al. (2020)	54.40	9.70	1.85	22	23
		ViBe, Yang et al. (2016)	63.13	33.8	1.70	25	19
		GMM, Wren et al. (1997)	41.74	-62.6	1.78	21	18
		Faster-RCNN, Ren et al. (2015)	71.59	48.7	1.19	34	7
	GM-PHD	Patch-based-CNN, Aguilar et al. (2021)	68.13	31.2	1.99	29	8
		Proposed Object Detection	78.18	56.0	1.16	34	5
		Ground Truth Detections (Calibration)*	99.8*	99.8*	0.12*	46*	1*
		3 Frame-based, Ao et al. (2020)	61.80	39.7	2.11	22	24
		ViBe, Yang et al. (2016)	60.42	32.3	2.13	19	21
		GMM, Wren et al. (1997)	59.74	23.3	2.15	25	16
		Faster-RCNN, Ren et al. (2015)	71.7	47.7	2.36	31	22
		Patch-based-CNN, Aguilar et al. (2021)	69.44	42.7	2.79	23	12
		Proposed Object Detection	77.53	57.1	1.26	32	7

**FIGURE 8 |** Sample Object Tracking. The square denotes the object current location and the line the object past locations. First column: ground truth marks. Second Column: Faster RCNN (Ren et al., 2015) and SORT (Bewley et al., 2016). Third column: proposed tracking algorithm.

numerous small vehicles such as motorcycles or bicycles. **Figure 6** shows that ViBe and GMM struggle to detect small and low contrast targets; hence, their recall values in **Table 3** are the lowest for AoI 1. Similarly, the 3-frame difference approach merges and splits nearby targets. On the other side, **Figure 6** shows that the supervised approaches detect a large number of relevant objects; thus their

recall score for all these methods is greater than 75%. However, both Faster RCNN and patch-based RCNN suffer from false positives such as detecting objects in farms or buildings. These artifacts reduce the overall F1 score for the detectors.

Figure 7 shows AoI two which contains two high-speed highways next to buildings with rich textures that generate

TABLE 7 | Computing times for modified GM-PHD and GLMB filters.

AoI	Tracks	Tracker	Computing Time(s)
1	64	GLMB	227.02
		GM-PHD	45.39
2	47	GLMB	129.06
		GM-PHD	27.56
3	22	GLMB	82.26
		GM-PHD	19.79

false positives. For example, **Figure 7** shows clusters of moving objects. **Figure 7** shows that both Faster RCNN and the patch-based RCNN detect false positives in the static background while our approach can discriminate only moving objects. **Table 3** shows that the proposed approach obtains better *F1* scores than all the competing methods, thanks to the better combination of precision-recall. It detects more relevant objects while reducing the overall ratio of false positives.

Object Tracking Evaluation

We compare object tracking using the MOTA, MOTP, MT and ML and *F1* scores shown in **Tables 4, 5, 6**. We compare the proposed GM-PHD tracker with the SORT tracker, developed by Bewley et al. (2016) and with the Generalized Labeled Multi-Bernoulli filter (GLMB), developed by Vo et al. (2017). We test the tracking outputs applied to each object detector shown in **Table 3** combined with all 3 filters.

The rows marked with an asterisk* in **Tables 4, 5, 6** show tracking metrics using ground truth object detections as filter inputs. These measurements simulate ideal object detectors and contribute to calibrating the filters' parameters. **Tables 5, 6** show robust performance for all three trackers across AoI two and AoI 3 (high-speed highways): all three filters obtain MOTA scores close to 99%. However, **Table 4** shows a case where SORT outperforms the GM-PHD and the GLMB filter when tracking with ground truth labels. SORT obtains a MOTA score of 99.4% while the GLMB filter 85.1% and GM-PHD filter obtains 89.7%. The GM-PHD and GLMB filter decrease their performance mostly due to the increased uncertainty and label switches for nearby slow-moving targets inside the roundabout of AoI 1.

The second to seventh row of **Tables 4, 5, 6** show metrics for tracking results applied to each object detector output. These detectors present considerable challenges for trackers due to clutter measurements and numerous misdetections. **Tables 4, 5, 6** show that both the GLMB and GM-PHD filter outperform the SORT filter for object detectors with high detection rate. For instance, the GM-PHD filter obtains higher MOTA scores for 3-frame difference, Faster-RCNN, patch-based Faster-RCNN, and the proposed method. These results are reflected in **Figure 8** where the GM-PHD recovers most of the objects moving in the roundabout. On the other hand, SORT outperforms the GM-PHD and GLMB filters for object detection with low detection rate such as ViBe and GMM,

where SORT obtains higher MOTA scores than the GM-PHD filter but lower MOTA scores compared to the proposed object detection and GM-PHD filter.

During our experiments, we determined that SORT performs better in tracking cases with linear constant motions, such as in AoI one and AoI 2. In fact, SORT obtained better results than the GM-PHD and GLMB filter for AoI two when applied in our proposed method. However, SORT presented difficulties adapting to high-speed tracks as in AoI 3. **Figure 8** shows the incomplete track trajectories of applying SORT to the outputs of Faster RCNN.

Finally, our modified GM-PHD filter presents similar tracking performances to the GLMB filter. The GLMB tracker slightly outperforms the modified GM-PHD filter in most tracking scores in all three AoIs. This is an expected result as the GLMB tracker shares the RFS framework with GM-PHD but has been extended to jointly estimate object states and tracks. Nevertheless, the GLMB filter retrieves tracks at the cost of a high computational burden. In fact, the efficient implementation of the GLMB filter (Vo et al. (2017)) relies on a pre-processing PHD filter lookup step and a Gibbs sampler step to perform joint prediction and update. Vo et al. (2017) explain that the efficient GLMB filter has a complexity of $\mathcal{O}(P^2M)$, where P denotes the number of hypothesis and M the number of measurements. On the other hand, our proposed GM-PHD filter has a linear complexity of $\mathcal{O}(PM)$. Additionally, we present sample computational times using the default GM-PHD ($\mathcal{O}(PM)$) filter and default GLMB ($\mathcal{O}(P^2M)$) filter implemented in Matlab by Vo et al. (2017). **Table 7** shows that the default GLMB filter is on average 4.77 times slower than the default GM-PHD filter. While our implementation of the GM-PHD filter obtains slightly lower tracking scores, it presents a considerable advantage in terms of computational demands. This advantage is particularly important for on-board applications where robust online tracking algorithms are preferred.

CONCLUSION AND FUTURE WORK

In this paper, we presented an improved track-by-detection approach where we use motion information together with neural networks to detect small moving objects on satellite images. Additionally, we perform tracking by using a modified version of the GM-PHD filter. Our version of the GM-PHD uses a measurement-driven birth intensity approximation and a label propagation in time. We present results for three AoIs in a challenging dataset where our approaches do not only outperform competing detection and tracking algorithms, but also detect objects not labeled by the ground truth annotations.

While our method performs detection and tracking, the method still requires several improvements. For example, our approach still misses several objects at sub-pixel level that appear and disappear. This drawback could be improved by including

the tracking information into the object detection in order to perform a unified track-and-detection approach.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

AUTHOR CONTRIBUTIONS

CA contributed with the experimental design, production of results, and writing of this manuscript. MO and JZ contributed with developing the main idea, data analysis, and supervision

of this study. All authors contributed to the manuscript and approved the submitted version.

FUNDING

This research work has been funded by BPI France under the LiChIE contract. The open access publication fees are provided by Inria.

ACKNOWLEDGMENTS

The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support. Additionally, the authors would like to thank BPI France for the financial support under the LiChiE contract.

REFERENCES

- Acatay, O., Sommer, L., Schumann, A., and Beyerer, J. (2018). "Comprehensive Evaluation of Deep Learning Based Detection Methods for Vehicle Detection in Aerial Imagery," in 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 27-30 Nov. 2018 (IEEE), 1-6. doi:10.1109/AVSS.2018.8639127
- [Dataset] AFRL (2009). Wright-Patterson Air Force Base (WPAFB) Dataset. Available at: <https://www.sdms.af.mil/index.php?collection=public-data&page=public-data-list>.
- Aguilar, C., Ortner, M., and Zerubia, J. (2021). "Small Moving Target MOT Tracking with GM-PHD Filter and Attention-Based CNN," in 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP), Gold Coast, Australia, 25-28 Oct. 2021 (IEEE), 1-6. doi:10.1109/MLSP52302.2021.9596204
- Ao, W., Fu, Y., Hou, X., and Xu, F. (2020). Needles in a Haystack: Tracking City-Scale Moving Vehicles from Continuously Moving Satellite. *IEEE Trans. Image Process.* 29, 1944-1957. doi:10.1109/TIP.2019.2944097
- Bergmann, P., Meinhardt, T., and Leal-Taixé, L. (2019). "Tracking without bells and Whistles," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 27 Oct.-2 Nov. 2019, 941-951. doi:10.1109/iccv.2019.00103
- Bernardin, K., and Stiefelhagen, R. (2008). Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP J. Image Video Process.* 2008, 1-10. doi:10.1155/2008/246309
- Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. (2016). "Simple Online and Realtime Tracking," in 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25-28 Sept. 2016 (IEEE), 3645-3649. doi:10.1109/ICIP.2016.7533003
- Bohyung Han, B., and Davis, L. S. (2012). Density-based Multifeature Background Subtraction with Support Vector Machine. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 1017-1023. doi:10.1109/TPAMI.2011.243
- Canepa, A., Ragusa, E., Zunino, R., and Gastaldo, P. (2021). T-RexNet-A Hardware-Aware Neural Network for Real-Time Detection of Small Moving Objects. *Sensors* 21, 1252. doi:10.3390/s21041252
- Cao, C., Mao, X., Zhang, J., Meng, J., Zhang, X., and Liu, G. (2019/2018). "Ship Detection Using X-Bragg Scattering Model Based on Compact Polarimetric SAR," in *The Proceedings of the International Conference on Sensing and Imaging*. Editors E. T. Quinto, N. Ida, M. Jiang, and A. K. Louis (Cham: Springer International Publishing), 87-96. doi:10.1007/978-3-030-30825-4_8
- Chai, Y. (2019/2019). *IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, 3414-3423. doi:10.1109/ICCV.2019.00351
- Patchwork: A Patch-wise Attention Network for Efficient Object Detection and Segmentation in Video Streams
- Chen, Y., Qin, R., Zhang, G., and Albanwan, H. (2021). Spatial Temporal Analysis of Traffic Patterns during the Covid-19 Epidemic by Vehicle Detection Using Planet Remote-Sensing Satellite Images. *Remote Sensing* 13, 208. doi:10.3390/rs13020208
- Clark, D., and Vo, B.-N. (2007). Convergence Analysis of the Gaussian Mixture PHD Filter. *IEEE Trans. Signal. Process.* 55, 1204-1212. doi:10.1109/TSP.2006.888886
- Dendorfer, P., Ošep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., et al. (2021). MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking. *Int. J. Comput. Vis.* 129, 845-881. doi:10.1007/s11263-020-01393-0
- Du, B., Sun, Y., Cai, S., Wu, C., and Du, Q. (2018). Object Tracking in Satellite Videos by Fusing the Kernel Correlation Filter and the Three-Frame-Difference Algorithm. *IEEE Geosci. Remote Sensing Lett.* 15, 168-172. doi:10.1109/LGRS.2017.2776899
- Feichtenhofer, C., Pinz, A., and Zisserman, A. (2017). "Detect to Track and Track to Detect," in International Conference on Computer Vision (ICCV), Venice, Italy, 22-29 Oct. 2017, 1-11. doi:10.1109/iccv.2017.330
- Fortmann, T., Bar-Shalom, Y., and Scheffe, M. (1980). "Multi-target Tracking Using Joint Probabilistic Data Association," in 1980 19th IEEE Conference on Decision and Control including the Symposium on Adaptive Processes, Albuquerque, NM, USA, 10-12 Dec. 1980, 807-812. doi:10.1109/CDC.1980.271915
- Fu, Z., Angelini, F., Naqvi, S. M., and Chambers, J. A. (2018). "GM-PHD Filter Based Online Multiple Human Tracking Using Deep Discriminative Correlation Matching," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15-20 April 2018, 4299-4303. doi:10.1109/ICASSP.2018.8461946
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27-30 June 2016, 770-778. doi:10.1109/CVPR.2016.90
- Jiao, L., Zhang, R., Liu, F., Yang, S., Hou, B., Li, L., et al. (2021). New Generation Deep Learning for Video Object Detection: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* 2021, 1-21. doi:10.1109/tnnls.2021.3053249
- Kaack, L. H., Chen, G. H., and Morgan, M. G. (2019). Truck Traffic Monitoring with Satellite Images. *COMPASS '19*, 155-164. doi:10.1145/3314344.3332480
- Kirubarajan, T., Bar-Shalom, Y., Pattipati, K. R., and Kadar, I. (2000). Ground Target Tracking with Variable Structure IMM Estimator. *IEEE Trans. Aerosp. Electron. Syst.* 36, 26-46. doi:10.1109/7.826310
- Kuhn, H. W., and Yaw, B. (1955). The Hungarian Method for the Assignment Problem. *Naval Res. Logistics* 2, 83-97. doi:10.1002/nav.3800020109
- LaLonde, R., Zhang, D., and Shah, M. (2018). "Clusternet: Detecting Small Objects in Large Scenes by Exploiting Spatio-Temporal Information," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

- (CVPR), Salt Lake City, UT, USA, 18–23 June 2018 (IEEE), 4003–4012. doi:10.1109/CVPR.2018.00421
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). “Focal Loss for Dense Object Detection,” in 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 Oct. 2017 (IEEE), 2999–3007. doi:10.1109/ICCV.2017.324
- Mahler, R. P. S. (2003). Multitarget Bayes Filtering via First-Order Multitarget Moments. *IEEE Trans. Aerosp. Electron. Syst.* 39, 1152–1178. doi:10.1109/TAES.2003.1261119
- Mahler, R. P. S. (2007). *Statistical Multisource-Multitarget Information Fusion*. USA: Artech House, Inc.
- Panta, K., Clark, D. E., and Vo, B.-N. (2009). Data Association and Track Management for the Gaussian Mixture Probability Hypothesis Density Filter. *IEEE Trans. Aerosp. Electron. Syst.* 45, 1003–1016. doi:10.1109/TAES.2009.5259179
- Qian, X., Lin, S., Cheng, G., Yao, X., Ren, H., and Wang, W. (2020). Object Detection in Remote Sensing Images Based on Improved Bounding Box Regression and Multi-Level Features Fusion. *Remote Sensing* 12, 143. doi:10.3390/rs12010143
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). “You Only Look once: Unified, Real-Time Object Detection,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016 (IEEE), 779–788. doi:10.1109/CVPR.2016.91
- Reid, D. (1979). An Algorithm for Tracking Multiple Targets. *IEEE Trans. Automat. Contr.* 24, 843–854. doi:10.1109/TAC.1979.1102177
- Ren, S., He, K., Girshick, R. B., and Sun, J. (2015). “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” in *Neural Information Processing Systems*. Editors C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Cambridge, Massachusetts: MIT Press), 91–99.
- Ren, Y., Zhu, C., and Xiao, S. (2018). Small Object Detection in Optical Remote Sensing Images via Modified Faster R-CNN. *Appl. Sci.* 8, 813. doi:10.3390/app8050813
- Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). “Orb: An Efficient Alternative to SIFT or SURF,” in 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 Nov. 2011, 2564–2571. doi:10.1109/ICCV.2011.6126544
- Sommer, L., Kruger, W., and Teutsch, M. (2021). “Appearance and Motion Based Persistent Multiple Object Tracking in Wide Area Motion Imagery,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, BC, Canada, 11–17 Oct. 2021, 3878–3888. doi:10.1109/iccvw54120.2021.00434
- Stauffer, C., and Grimson, W. E. L. (2000). Learning Patterns of Activity Using Real-Time Tracking. *IEEE Trans. Pattern Anal. Machine Intell.* 22, 747–757. doi:10.1109/34.868677
- Teutsch, M., and Grinberg, M. (2016). “Robust Detection of Moving Vehicles in Wide Area Motion Imagery,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016 (IEEE), 1434–1442. doi:10.1109/CVPRW.2016.180
- Vo, B.-N., and Ma, W.-K. (2006). The Gaussian Mixture Probability Hypothesis Density Filter. *IEEE Trans. Signal. Process.* 54, 4091–4104. doi:10.1109/TSP.2006.881190
- Vo, B.-N., Vo, B.-T., and Hoang, H. G. (2017). An Efficient Implementation of the Generalized Labeled Multi-Bernoulli Filter. *IEEE Trans. Signal. Process.* 65, 1975–1987. doi:10.1109/TSP.2016.2641392
- Vo, B.-T., Vo, B.-N., and Cantoni, A. (2006). “The Cardinalized Probability Hypothesis Density Filter for Linear Gaussian Multi-Target Models,” in 2006 40th Annual Conference on Information Sciences and Systems, Princeton, NJ, USA, 22–24 March 2006, 681–686. doi:10.1109/CISS.2006.286554
- Vo, B.-T., and Vo, B.-N. (2013). Labeled Random Finite Sets and Multi-Object Conjugate Priors. *IEEE Trans. Signal. Process.* 61, 3460–3475. doi:10.1109/TSP.2013.2259822
- Wijnands, J. S., Zhao, H., Nice, K. A., Thompson, J., Scully, K., Guo, J., et al. (2021). Identifying Safe Intersection Design through Unsupervised Feature Extraction from Satellite Imagery. *Computer-Aided Civil Infrastructure Eng.* 36, 346–361. doi:10.1111/mice.12623
- Wojke, N., Bewley, A., and Paulus, D. (2017). “Simple Online and Realtime Tracking with a Deep Association Metric,” in 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 Sept. 2017 (IEEE), 3645–3649. doi:10.1109/ICIP.2017.8296962
- Wren, C. R., Azarbayejani, A., Darrell, T., and Pentland, A. P. (1997). Pfunder: Real-Time Tracking of the Human Body. *IEEE Trans. Pattern Anal. Machine Intell.* 19, 780–785. doi:10.1109/34.598236
- Xuan, S., Li, S., Han, M., Wan, X., and Xia, G.-S. (2020). Object Tracking in Satellite Videos by Improved Correlation Filters with Motion Estimations. *IEEE Trans. Geosci. Remote Sensing* 58, 1074–1086. doi:10.1109/TGRS.2019.2943366
- Yang, Y., Han, D., Ding, J., and Yang, Y. (2016). “An Improved ViBe for Video Moving Object Detection Based on Evidential Reasoning,” in 2016 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Baden-Baden, Germany, 19–21 Sept. 2016 (IEEE), 1709–1724. doi:10.1109/MFI.2016.7849462

Conflict of Interest: Author MO was employed by Airbus Defense and Space, France.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Aguilar, Ortner and Zerubia. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



How Scalable Are Clade-Specific Marker K-Mer Based Hash Methods for Metagenomic Taxonomic Classification?

Melissa Gray, Zhengqiao Zhao and Gail L. Rosen*

Ecological and Evolutionary Signal-Processing and Informatics Laboratory, Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA, United States

OPEN ACCESS

Edited by:

Hagit Messer,
Tel Aviv University, Israel

Reviewed by:

David Koslicki,
The Pennsylvania State University,
United States
Cinzia Pizzi,
University of Padua, Italy

*Correspondence:

Gail L. Rosen
glr26@drexel.edu

Specialty section:

This article was submitted to
Statistical Signal Processing,
a section of the journal
Frontiers in Signal Processing

Received: 23 December 2021

Accepted: 18 May 2022

Published: 05 July 2022

Citation:

Gray M, Zhao Z and Rosen GL (2022)
How Scalable Are Clade-Specific
Marker K-Mer Based Hash Methods
for Metagenomic Taxonomic
Classification?
Front. Sig. Proc. 2:842513.
doi: 10.3389/frsip.2022.842513

Efficiently and accurately identifying which microbes are present in a biological sample is important to medicine and biology. For example, in medicine, microbe identification allows doctors to better diagnose diseases. Two questions are essential to metagenomic analysis (the analysis of a random sampling of DNA in a patient/environment sample): How to accurately identify the microbes in samples and how to efficiently update the taxonomic classifier as new microbe genomes are sequenced and added to the reference database. To investigate how classifiers change as they train on more knowledge, we made sub-databases composed of genomes that existed in past years that served as “snapshots in time” (1999–2020) of the NCBI reference genome database. We evaluated two classification methods, Kraken 2 and CLARK with these snapshots using a real, experimental metagenomic sample from a human gut. This allowed us to measure how much of a real sample could confidently classify using these methods and as the database grows. Despite not knowing the ground truth, we could measure the concordance between methods and between years of the database within each method using a Bray-Curtis distance. In addition, we also recorded the training times of the classifiers for each snapshot. For all data for Kraken 2, we observed that as more genomes were added, more microbes from the sample were classified. CLARK had a similar trend, but in the final year, this trend reversed with the microbial variation and less unique k-mers. Also, both classifiers, while having different ways of training, generally are linear in time - but Kraken 2 has a significantly lower slope in scaling to more data.

Keywords: metagenomics, taxonomic classification, supervised classification, hash-based indexing, incremental learning, algorithm scalability, benchmarking

BACKGROUND

DNA sequencing has enabled the investigation of microbial communities using cultivation-independent, DNA/RNA-based approaches (Brul et al., 2010; Berg et al., 2020; Coenen, 2020). We can think of these microbial communities as microscopic civilizations, in which bacteria not only act independently but learn to cooperate and compete with each other, to gain more nutrients and resources, and that result in advanced time-course patterns of microbial proliferation and death (Figueiredo et al., 2020). As humans, we must take observations of microbiomes. While imaging is still too coarse for observing 10^{11} cells per Gram of colon content (Sender et al., 2016), sampling their

DNA from next-generation sequencing of microbes is commonly used, with many other ‘omic techniques emerging that sample measurements of the metatranscriptome, metaproteome, and metabolome (Creasy et al., 2021). Microbiomes are found everywhere on Earth, including soil, water, air, and animal hosts (Nemergut et al., 2013). Understanding microbiomes is the first step, with many potential engineering applications to follow (Woloszynek et al., 2016).

Signal processing has played an important role in metagenomic identification and taxonomic classification, which is the supervised labeling of a taxonomic class to a DNA/RNA sequencing read (Rosen and Moore, 2003; Rosen et al., 2009; Borrayo, 2014; Alshawaqfeh, 2017; Elworth et al., 2020). While taxonomic classification is the application that we cover in this paper, metagenomics is not limited only to this problem, and emerging techniques are proving useful for unsupervised “binning” of metagenomics reads (Kouchaki et al., 2019). Information-theoretic feature selection (Garbarine et al., 2011) and deep neural network sequence embeddings (Woloszynek et al., 2019), useful methods from signal processing, can be performed before metagenomic taxonomic classification to reduce feature dimensionality and computational complexity.

As of 2019, over 80 metagenomic taxonomic classification tools have been published (Gardner et al., 2019), while benchmarking efforts try to quantify the most representative ones (Ye et al., 2019). We have previously shown an in-depth case study of the naïve Bayes classifier’s (and its incremental version’s) accuracy and speed over the yearly growth of NCBI (Zhao et al., 2020). Now, for this study, we study clade-specific marker hash-based techniques, due to their popularity, efficiency/speed, and comparable sensitivity/precision when benchmarked against BLAST-based methods (Wood et al., 2014). These algorithms have been shown to be competitive algorithms on several benchmarks on real and simulated data (McIntyre et al., 2017; Sczyrba et al., 2017; Meyer et al., 2021). In 2017, a comparison of the two algorithms shows their performances are relatively similar, with CLARK tending to yield better relative abundance estimates than Kraken2, which can be due to more genomes in their curated database (McIntyre et al., 2017). While there are techniques like sourmash (Brown and Irber, 2016; Liu and Koslicki, 2022; LaPierre et al., 2020) that can sketch k-mer compositions, they do not perform well when the reference genome is missing from the database (dibsi-rnaseq, 2016). While Kraken2/CLARK has been shown to predict low-abundance false positives, it has been shown that a larger database can improve Kraken2 performance (LaPierre et al., 2020). Other techniques, such as LSHvec (Shi and Chen, 2021), which embeds sequences after a compression k-mers with a hash, may be able to transform some of the limitations of hash-based techniques using deep learning. Therefore, LaPierre et al. and McIntyre et al.’s study invite an investigation into how database composition can affect methods that use these efficient k-mer presence/absence to differentiate clades, and this study can give insight into how more recent hash-based techniques will perform.

It has been previously shown that database size influences the accuracy of Kraken and its Bayesian extension Bracken (Nasko

et al., 2018). While the study highlights the percentage of “unclassified” reads goes down as the database grows, it does not fully examine time to run the algorithms over varying size databases or how the final relative abundance result changes. As genomes in the databases increase, the representation of the organisms in the database may not always be uniform across the tree of life. With mutations, clade-identifying k-mers that may have been previously discriminating between taxa before, may be missing in updates, reducing the search capacity of these methods. These identifying k-mers will not be captured simply by looking at orthologs shared between genomes (Lan et al., 2014). Therefore, the size of the database and its growth may affect performance of the kmer-based algorithms in addition to runtime.

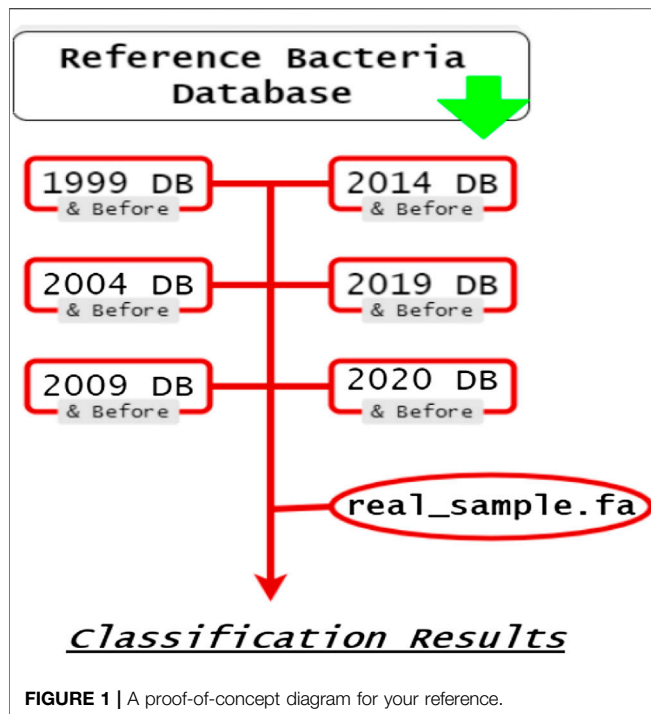
CLARK and Kraken 2

CLARK and Kraken 2 are both well known metagenomic classifiers, software that “reads” short sequences of DNA and attempts to accurately identify what organism they came from. Although CLARK and Kraken 2 are both clade-specific k-mer hash-based metagenomic classifiers, they operate in different, almost opposite, ways. Both software, like most classifiers, decompose the DNA sequences into smaller features called k-mers to make comparisons easier. Their k-mers are 31 nucleotides long by default. CLARK’s training step takes each k-mer and cycles through all the genomes in its database to see if any of them have that sequence. If more than one genome does, then the k-mer is ignored and the program moves on to the next one. Now when a query sequence is tested, for each k-mer in the query, if only one genome matches it, then that genome’s score of how many k-mers it matches the query is incremented. This approach prioritizes the calculation of unique k-mers, or k-mers that are only found in one genome to the query. After all the k-mers are cycled through, the genome with the highest unique k-mer score is deemed as the correct match. If the score is too low or there are genomes that tie, then the sample DNA is marked as unclassified (Ounit et al., 2015).

On the other hand, when Kraken 2 compares a k-mer in the query to the genomes in its database, for any k-mer match, the genome score is incremented by one. Kraken 2 doesn’t skip over k-mers that are shared by multiple genomes (Wood et al., 2014). Instead, it takes those into account. This approach prioritizes common k-mers, specifically the k-mers that the genomes have in common with the query DNA. The genome with the highest common k-mer score is deemed the correct match. In the event of a tie or if the scores do not meet Kraken 2’s default threshold (the genome has a confidence threshold of 0.65), then the sample DNA is marked as unclassified (Wood et al., 2019).

Goals

The goals of this paper are to examine the behaviors of metagenomic classifiers as the information in their databases increases over time: how much they classify, how they classify, and how fast they classify. While similar studies have been previously conducted, they are for other methods and for the study of Kraken 2, it was limited. For example, Nasko et al. (2018) examined Kraken’s performance for successive Refseq databases,



but the metrics were mostly for speed and amount classified (but not how the distribution of those classified changed). We wish to gain a more comprehensive insight into the scalability of k-mer based hash methods of metagenomic classifiers. We also wanted to compare two well-known classifiers, CLARK and Kraken 2, to see which one was more efficient and how both of them could improve to be useful into the future as more genomes are sequenced and added to their databases. Also, we previously benchmarked the naive Bayes classifier (Zhao et al., 2020) for its accuracy to classify (NBC classifies everything so the “amount” is negligible) and speed, and we will use the same dataset (devised on a yearly basis) in this study so that it can be fairly compared.

METHODOLOGY

Datasets

The build/train (sub-)databases are derived from the National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) bacterial genome database (Sayers et al., 2019) and the NCBI genbank assembly summary file for bacteria available at ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/assembly_summary.txt. The test data is taken from NCBI’s Sequence Read Archive (SRA ID: SRS105153) (Huttenhower et al., 2012) and is a human gut sample from the Human Microbiome Project (Nasko et al., 2018). We use experimental data because it is more likely to contain a true distribution of novel taxa.

Setup, Build, and Classify

The database snapshots from 1999 to 2020 were designed in (Zhao et al., 2020). Statistics about the database growth in

genomes and their lineages can be found in that paper’s Additional File 1: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7507296/bin/12859_2020_3744_MOESM1_ESM.pdf. The genomes from these lists were obtained in Kraken and usually a subset was found in CLARK’s downloaded database. In the supplementary additional file 2, we provide the Kraken/CLARK overlap and the additional genomes in Kraken (that were not found in CLARK’s database).

Kraken 2

Kraken 2’s default bacteria database was used to find the list of bacteria genomes. All uncompleted genomes were filtered out, leaving only the completed ones left in the list. Six lists (1999, 2004, 2009, 2014, 2019, and 2020) were then created, as shown in **Figure 1**. Each was filled with genomes that were sequenced in their respective years or before. For example, A bacteria genome sequenced in 2010 would be in the 2014, 2019, and 2020 list, but a genome sequenced in 2020 would only be present in the 2020 list. For Kraken 2, those genome lists were then used to create library. *fna* files that Kraken 2 uses in its databases (Wood et al., 2014). Those library. *fna* files were then used to create six sub-databases for Kraken 2 (1999, 2004, 2009, 2014, 2019, and 2020).

Creating the custom library. *fna* files required python programs: *summary.py* and *hive.py*. After this set up, each Kraken 2 sub-database was built and used to classify SRA ID: SRS105153, a file containing about 70 million reads (approximately 100-200bp in length per read) from a human gut sample (Huttenhower et al., 2012).

CLARK

CLARK’s custom sub-databases were built with the same lists as in **Figure 1**, but in a slightly different way. The difference is in how CLARK stored its genomes. When CLARK’s default bacteria database was downloaded, the genomes were stored in individual FASTA files. Some files had more than one genome written in it, but each file’s name corresponded to a GCF accession code. This made sorting the genomes from the default bacteria database into the custom databases much easier. After finding all the file paths for each of the GCF accession numbers, those files would then be copied into the “Custom/” folder of their corresponding custom sub-database.

$$T_{CB} = T_B + (C_{R1} - C_{R2}) \quad (1)$$

Where TBC is the entire CLARK build time, TB is the runtime of the initial stage of the build/training time of CLARK. It is then added to the second stage of the build time which is calculated by subtracting first run (which is the first stage build + classify times), CR1, minus solely the classify times, CR2. CR1 is the runtime for CLARK’s classification script “classify_metagenomes.sh” when it is run with a particular database for the first time, and CR2 is the runtime for CLARK’s classification script when it is run on that same database after the first time (second, third, etc. time).

Building a database with CLARK is not as straightforward as with Kraken 2. CLARK does its database building in two parts: the first part with its actual building script and the second part is built when the database is first used during classification. CLARK also

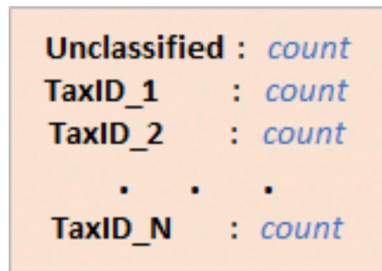


FIGURE 2 | The taxa in CLARK and Kraken 2's results and the number of reads that were classified as that taxa (count).

does not store the built database in a directory. Instead, if you want to use a different database or previous database, the build script must be run again, which includes both a build component and a classify component. Due to this combination of steps in the script, CLARK's build time was calculated (Eq. (1)). This was done by subtracting the second classification runtime (where only classification occurred) from the first (where the second part of database building was done). That difference was then added to the runtime of the build script (the first part of the building) to get the full runtime of CLARK's database building.

Parsing the Results

Kraken 2

For the Kraken 2 classification results, the text files were parsed line by line to gather information on whether that read was classified and what it was classified as. This information was stored into a Python dictionary, as well as a count variable that kept track of how many reads were classified as a particular taxa or how many were unclassified (Figure 2). Traceback was also performed to include counts of every taxonomic rank. For example, if a read was classified as genus X, then genus X's family, class, order, etc. would also be counted.

CLARK

Since CLARK's classification results were stored in .csv files, they were easy to parse. Each row in the "Assignment" column was read to ascertain what CLARK classified the read as. Traceback was also performed here, and the information was stored the same way Kraken 2's was (Figure 2).

Relative Abundance, Triangular Bray-Curtis, and Graphing

Calculating Relative Abundance

$$RA_i = \frac{C_i}{\sum_i C_i} \times 100\% \quad (2)$$

where RA_i is the relative abundance for a particular taxonomic class i . C_i is the number of times that a DNA read from class i is

observed in the sample divided by the number of all observations (DNA reads) from all classes.

A general equation (Eq. 2) was used to calculate each taxa's relative abundance in two different ways. The first way was to calculate the taxa's relative abundance within the set of reads that were given a classification label. This means that each read was assigned one of $\{C_0, C_1, \dots, C_{N-1}\}$, where N is the number of taxonomic units in a given taxonomic rank in the classification results, and summed then divided by the total reads on the taxonomic level (e.g. on the species level, each species is incremented by the count of each read assigned to that species and then divided by the total reads that classify on the species level). In other words, the count of each taxa was divided by the total number of classified reads, then multiplied by 100 to make it a percentage.

The second way was to calculate the taxa's relative abundance among the total number of reads. To illustrate the labeling of all reads, an unclassified category was added such that $i \in \{1, \dots, N, \text{unclassified}\}$ so that $C_{\text{unclassified}}$ is accounted for as a bar in the graph and in the denominator of the relative abundance calculation. These results were exported to excel file sheets for each taxonomic rank.

Graphing Relative Abundance

Each taxa's relative abundance is compared to a 3% threshold, meaning that any taxa that has a relative abundance above 3% of the sample is plotted in its own bar. Any taxa that do not meet these conditions are aggregated into the "Others" bar on their respective graph. The Percent Classified was calculated from the percent of unclassified reads and then plotted on top of the bar graph as a scatter plot.

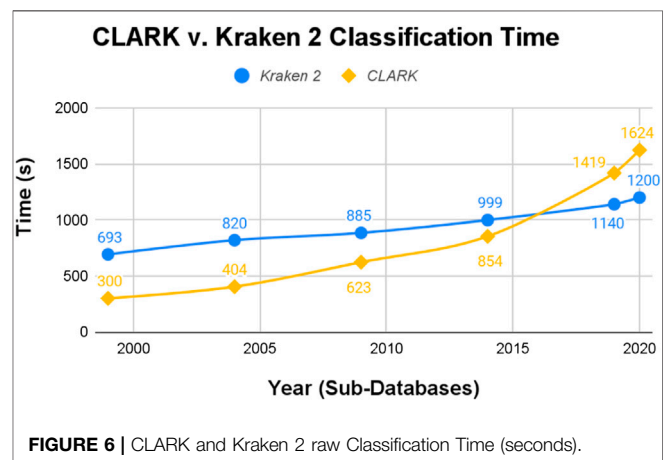
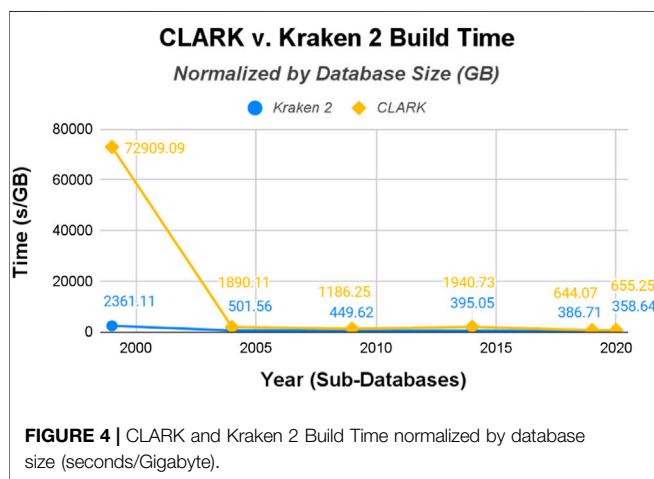
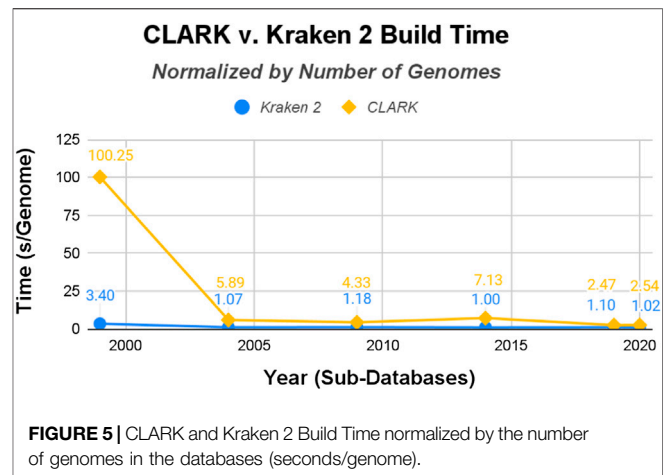
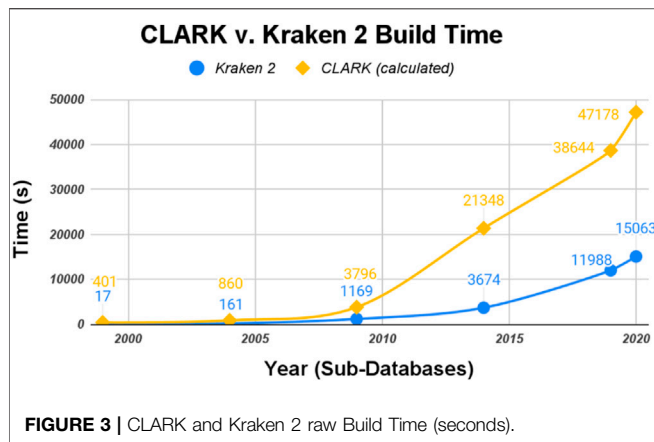
Calculating Pairwise Bray-Curtis Dissimilarity

$$\frac{\sum |u_i - v_i|}{\sum |u_i + v_i|} \quad (3)$$

Where u_i is the relative abundance of taxonomic class i in one comparison sample (e.g. 1999 database) and v_i relative abundance of a taxonomic class i in another sample (e.g. 2004 database). Each sum is summed over the total number of taxonomic classes.

The Bray-Curtis dissimilarity (Eq. 3) is commonly used in ecology to measure the differences between the community compositions of two populations. In this study, we calculate the Bray-Curtis dissimilarities between the classification results of the sub-databases. The calculation of the Bray-Curtis Dissimilarity was done by Scipy's `spatial.distance.braycurtis()` function, and the equation for it (Eq. 3) came from (Scipy, 2021). Using it in a pairwise fashion calculated a Bray-Curtis dissimilarity with every combination of sub-database (excluding duplicate pairs such as 1999 and 1999). This allows every sub-database's classification results to be compared to each other.

These values were arranged in a grid and used to create a heatmap of Bray-Curtis dissimilarity. The lower triangular dissimilarity is left blank because those values are redundant.



One heatmap shows how the classified part of each sub-database's results compare, while the other shows how the entire classification results of each sub-database compare. On heatmaps, 0 (zero) represents that the sub-databases are very similar, while 1 (one) represents that the sub-databases are very different (Figures 16–19).

RESULTS

Build/Training Time

Both normalizations follow the same trend: Kraken 2's database building procedure is faster than CLARK's. Just by raw numbers, shown in Figure 3, Kraken 2 had the fastest build time. It was somewhat complicated to measure CLARK's build time because of how its build and classify procedure is not separated in the first step (Eq. 1).

The raw data was then normalized with the size of each classifiers' sub-databases (in gigabytes shown in Figure 4 and the number of genomes shown in Figure 5). CLARK has an unusually large build time/GB for the smallest database (1999), and then the time per GB decreases drastically. Kraken 2's build time/GB for the 1999 database is also much larger than its build

time for the other five databases, but it is still 30x shorter than CLARK's build time for the 1999 database. Also, Kraken 2's build time for the other five databases are less than half that of CLARK's in time/GB and even more for time/genome. Overall, even when normalized to account for the difference in the size of databases and number of genomes, Kraken 2's database building procedure ran several times faster than CLARK's.

Classification Time

Conversely, CLARK's procedure is faster at classifying than Kraken 2's. Just by raw numbers, shown in Figure 6, CLARK had the shorter classification time for the 1999, 2004, 2009, and 2014 databases. Its classification time for the 2019 and 2020 databases, however, were longer than Kraken 2's.

Normalizing the raw data by gigabytes (see Figure 7) and genomes (see Figure 8), this time-trend remains similar. While both start out with particularly long runtimes for their classification procedures, Kraken 2's is substantially higher and remains that way, even after the drastic decrease after the 1999 database. But this time their runtimes are much closer in value than the build/training times. CLARK classifies several times faster for time/GB, but they are both similar in time/genome.

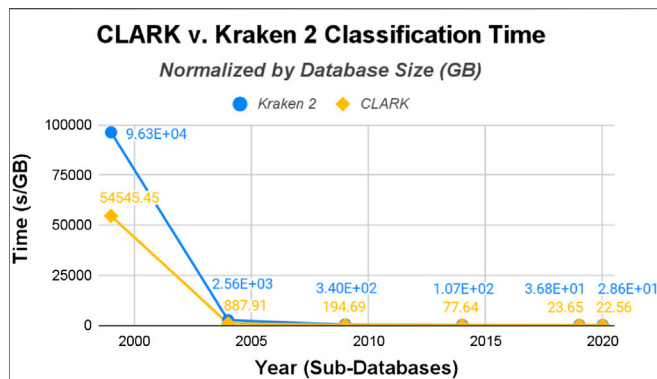


FIGURE 7 | CLARK and Kraken 2 Classification Time normalized by the size of the databases (seconds/Gigabyte).

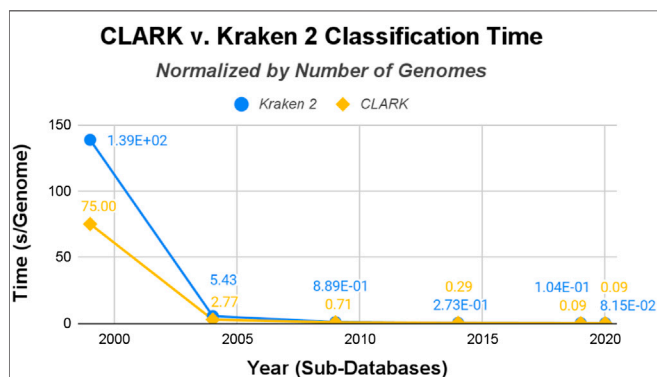


FIGURE 8 | CLARK and Kraken 2 Classification Time normalized by the number of genomes in the databases (seconds/genome).

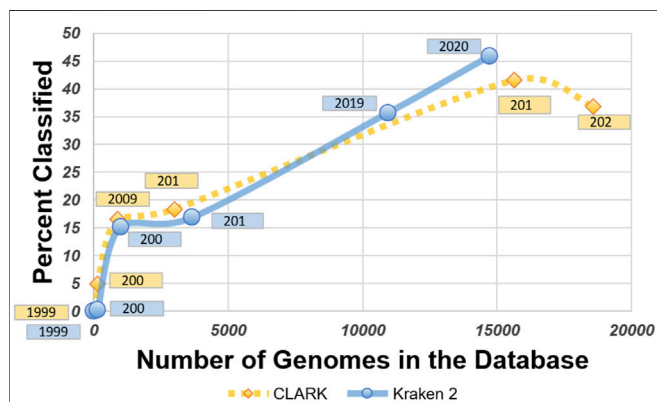


FIGURE 9 | Graph comparing the number of genomes in their databases with the percent classified for each sub-database for CLARK and Kraken 2.

Overall, the methods are designed to perform the classification procedure magnitudes faster than the build time, since users usually want results quickly and are willing to spend a one-time longer cost up-front.

Classification Results

Since there is no ground truth classification for the gut microbiome sample, there is no way to check how accurate CLARK or Kraken 2's classifications are, but we can examine how the number of reads classified changes as more genomes are added to their databases.

CLARK generally had a higher percentage of classified reads for all sub-databases except 2020, as shown in **Figure 9**. Even when Kraken 2 had more genomes in its database to reference, CLARK's percent-classified was still higher. In 2004, CLARK classifies about 5% of sequences while Kraken 2 classifies 1% (see relative abundance tables in **Supplementary Material**), and this difference compounded with the limitations of the databases causes a significant dissimilarity between the classifiers (shown later in **Figure 20**). Also, CLARK's percentage of classified reads dropped suddenly and drastically with the 2020 sub-database.

While Kraken 2's classification percentages seemed to increase steadily in an exponential curve, as shown in **Figure 11**, CLARK's had an unexpected decrease after 2019, as shown in **Figure 10**. **Figures 10, 11** show that CLARK and Kraken 2 classified reads in a similar fashion for genus level, in terms of quantity and identity. Since CLARK classified more than Kraken 2 in 2004, in **Figure 10**, it found *Bacteroides* as the first genus to rise above the 3% threshold (that we used for visualization). However, for 2020, CLARK only classified about 37% of the sample while Kraken 2 classified nearly 50% (see relative abundance tables in **Supplementary Material**). Also, *Bacteroides* and *Phocaeicola* are the dominant genera detected by both metagenomic classifiers. By 2020, for *Phocaeicola*, CLARK and Kraken 2's general relative abundance percentages were 10.66% and 13.47% respectively (see relative abundance tables in **Supplementary Material**). For *Bacteroides*, their percentages were 20.37% and 31.26% respectively (see relative abundance tables in **Supplementary Material**).

In **Figures 12, 13**, the differences between the genera that CLARK and Kraken 2 classified are shown. It is also notable to mention that CLARK and Kraken 2 did not detect *Bacteroides* to the same extent using the 1999 and 2004 sub-databases. This is probably due to CLARK's ability to detect *Bacteroides* given the limited database. Kraken 2 did not detect as many and therefore, other bacteria genera (e.g. *Bacteroidetes* such as *Porphyromonas*) were found in high abundance. Also in 2004, neither classifier detected *Phocaeicola* in any significant amount, probably due to the absence of that bacteria from the database.

What can be more contentious is the detection of *Alistipes* and *Faecalibacterium*. While *Faecalibacterium prausnitzii* is detected in the species level for 2009 and after for Kraken 2 (**Figure 15**), it is not detected in the genus level in 2020 (**Figure 13**). This is due to Kraken 2's ability to assign more reads at the genus level than the species level and while the *Faecalibacterium* has the same number of reads in each, it falls below our 3% threshold for the genus level. In fact, because Kraken 2 classifies less reads, there is more of a diversity of bacteria meeting this 3% threshold as shown in **Figures 13, 15**. However, for the 3 most abundant genera, the classifications tend to agree more when run on recent databases.

On the species level, shown in **Figures 14, 15**, CLARK and Kraken 2's classification results also differ slightly in what they

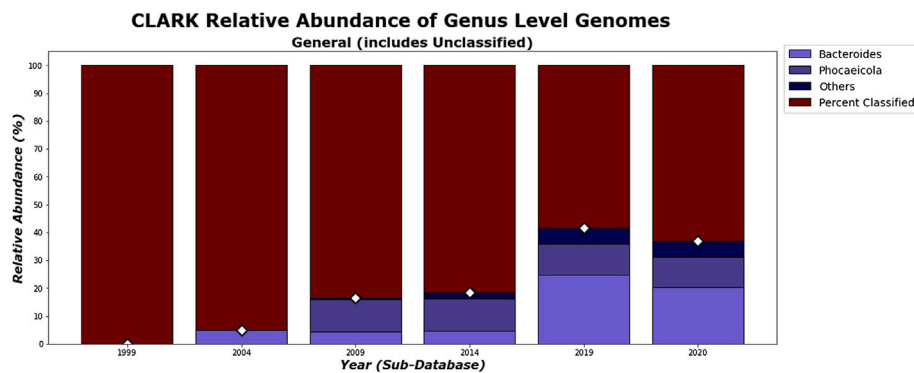


FIGURE 10 | CLARK's general relative abundance for Genus Level. Only taxa whose general relative abundance was at least 3% are shown as a colored bar here. A bar for the Unclassified group is included. The percentage of classified reads for each year are shown as diamond markers.

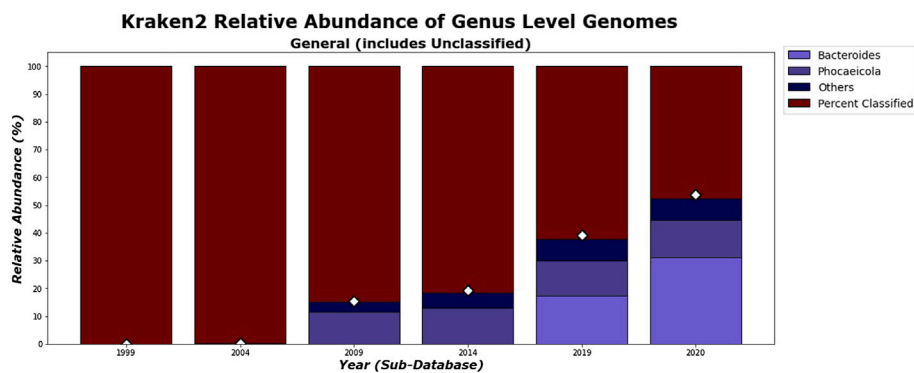


FIGURE 11 | Kraken 2's general relative abundance for Genus level. Only taxa whose general relative abundance was at least 3% are shown as a colored bar here. A bar for the unclassified group is included. The percentage of classified reads for each year are shown as diamond markers.

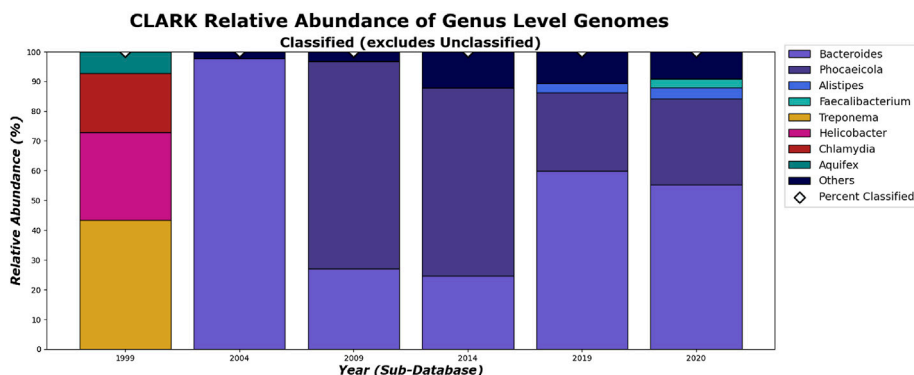


FIGURE 12 | CLARK's classified relative abundance for Genus Level. Only taxa whose classified relative abundance was at least 3% are shown as a colored bar here. Only bars for classified reads are included. The percentage of classified reads traced back to genus level for each year are shown as diamond markers.

classified. For example, only *Bacteroides* sp. M10 is found with Kraken 2, and this may be due to the different species in the different methods' databases. However, by 2020, the methods tend to be in more agreement on the sample composition. It is

also interesting to note that everything that CLARK classifies, it classifies on all levels (Figures 12, 14), while Kraken 2 has different percentages classified on each taxonomic level (Figures 13, 15.)

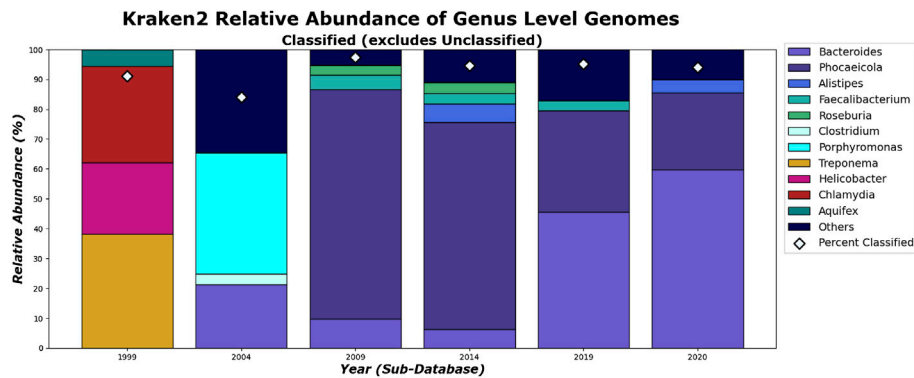


FIGURE 13 | Kraken 2's relative abundance for Genus level. Only taxa whose classified relative abundance was at least 3% are shown as a colored bar here. Only bars for the classified reads are included. The percentage of classified reads traced back to genus level for each year are shown as diamond markers.

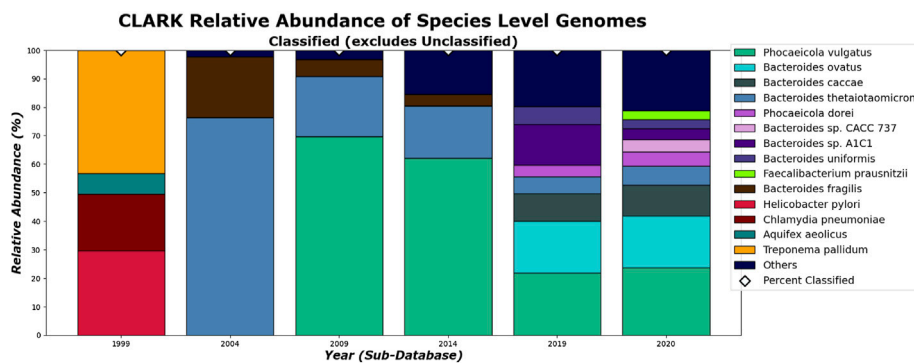


FIGURE 14 | CLARK's relative abundance for Species level. Only taxa whose classified relative abundance was at least 3% are shown as a colored bar here. Only bars for the classified reads are included. The percentage of classified reads traced back to species level for each year are shown as diamond markers.

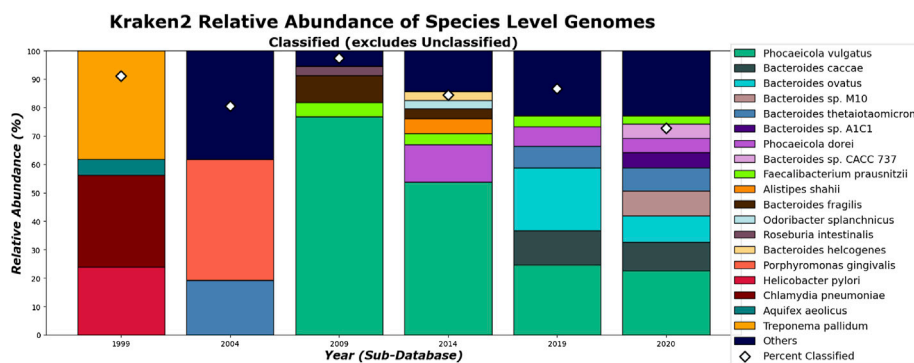


FIGURE 15 | Kraken 2's relative abundance for Species level. Only taxa whose classified relative abundance was at least 3% are shown as a colored bar here. Only bars for the classified reads are included. The percentage of classified reads traced back to species level for each year are shown as diamond markers.

We can see how increasing knowledge added to the training database changes the classification results over time—using the Bray-Curtis dissimilarity measure from the ecological literature to quantify ecosystem dissimilarity. As expected, the Bray-Curtis dissimilarity shows that the classification

results of the gut microbiome sample generally become less similar as the time increases between sub-database versions, shown in **Figures 16–19**. An exception is the Bray-Curtis dissimilarity between the 2009 and 2014 sub-databases of both CLARK and Kraken 2. That dissimilarity is even lower

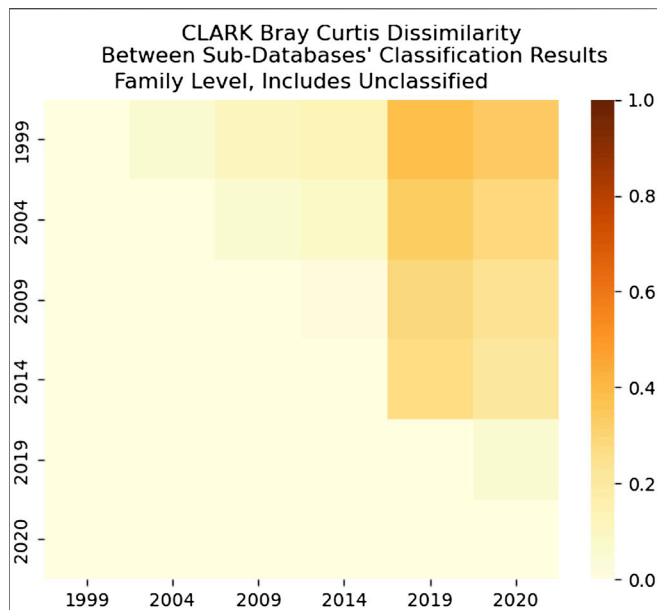


FIGURE 16 | CLARK's Bray-Curtis dissimilarity score for Family level. It is a comparison between each sub-databases' classification results for CLARK. It includes comparisons of what CLARK classified as well as what CLARK didn't classify for each year. It is interesting that 2009–2014 databases yield the most similar results on the family level (more similar than 2019–2020).

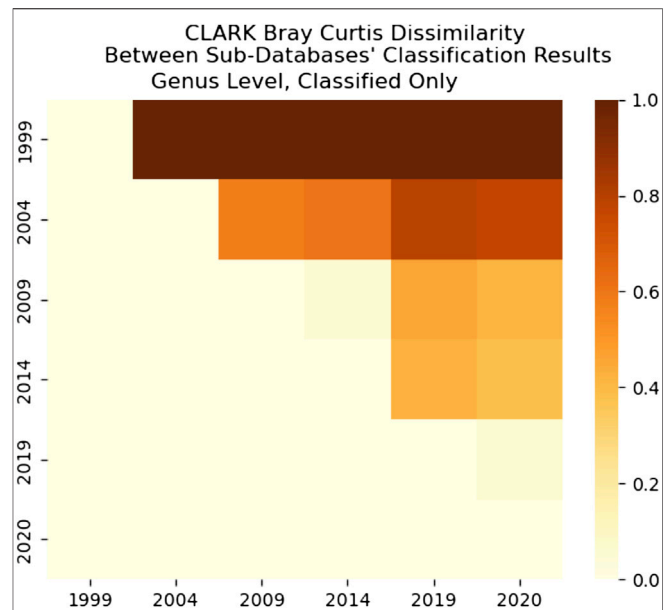


FIGURE 18 | CLARK's Bray-Curtis Dissimilarity score for Genus level. It is a comparison between each sub-databases' classification results. It only includes what CLARK classified. It is interesting that 1999 results are significantly different from any other years', while 2009–2014 and 2019–2020 are the most similar.

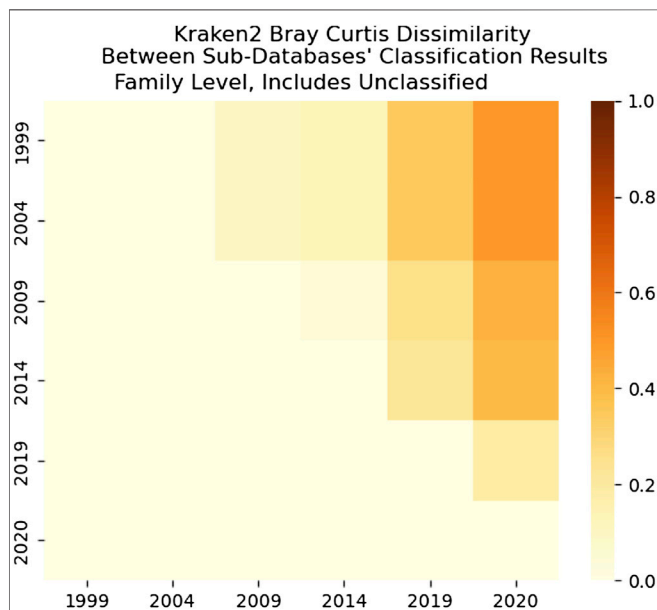


FIGURE 17 | Kraken 2's Bray-Curtis dissimilarity score for Family level. It is a comparison between each sub-databases' classification results for Kraken 2. It includes comparisons of what Kraken 2 classified as well as what it didn't classify for each year. Unlike CLARK, Kraken 2's most similar results are from the 1999 and 2004 databases.

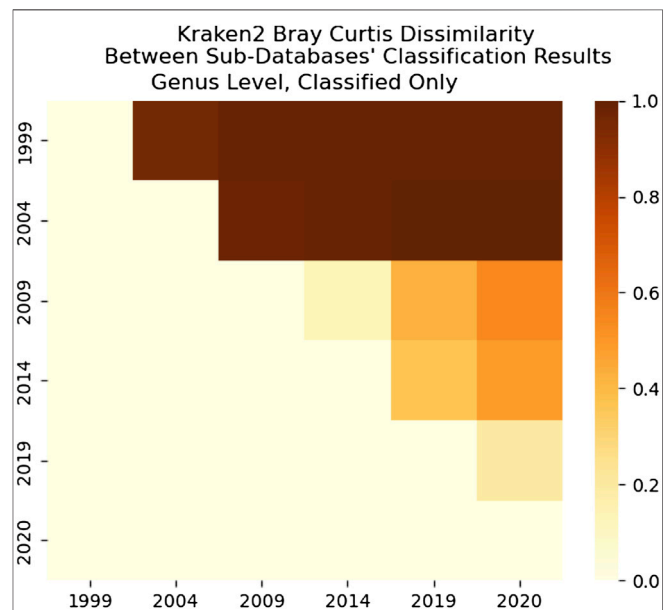


FIGURE 19 | Kraken 2's Bray-Curtis Dissimilarity score for Genus level between each sub-databases' classification results. The comparison shows that classified results are more similar for successive databases—although some are more similar than others (such as 1999–2004 and 2009–2014).

than the dissimilarity between the 2019 and 2020 sub-databases (see Bray-Curtis dissimilarity tables in **Supplementary Material**).

As we can see in **Figure 18**, the dissimilarity is greatest between the CLARK classifications on the genus level between 1999 and any other year, since only a handful of genera were

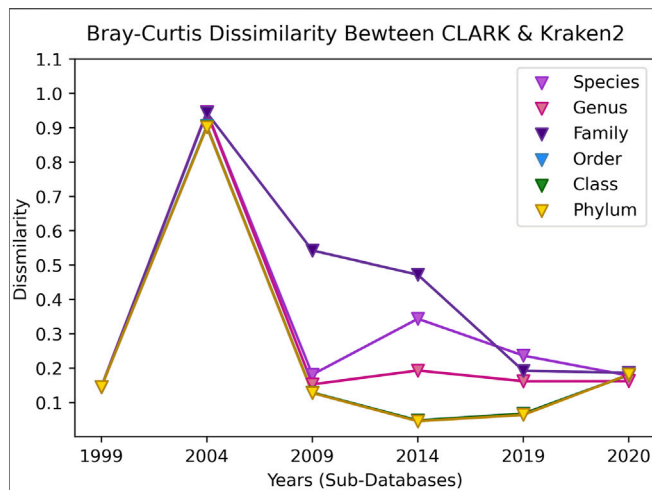


FIGURE 20 | The Bray-Curtis dissimilarity between the classified results of CLARK and Kraken 2 for each taxonomic level over time (increasing database size). Since only 4 organisms from 4 different phyla were classified in 1999, the similarity of the results is close. In 2004, a distinct issue is that CLARK classified significantly more reads than Kraken 2, which skewed the dissimilarity to be significant for all taxonomic levels. In 2009, the methods are more concordant, and for the rest of the years, the species-level classification is different while the phylum classification is more similar. The methods are pretty similar on all taxonomic levels once again in 2020.

known and CLARK had a large number of percent classified in 2004 (seen in **Figures 10, 12, 14**) compared to Kraken 2. For Kraken 2, since the percent classified did not increase for 2004, the 1999 and 2004 classification results are very similar, as seen in **Figures 17, 19**. This is also true for 2009 to 2014 for genus and family level classifications for both CLARK nor Kraken 2 (and result in more similar results than the transition from 2019 to 2020), as seen in **Figures 16–19**. This similarity reflects how the change in percent classified (for both CLARK and Kraken 2) between the 2009 to 2014 database years was the smallest change seen in all the years (seen in **Figures 10–15**). This can be due to the fact that the database additions did not add the gut microbes that they are in or relatives of those in this metagenomic sample, and those additions came later.

Finally, in **Figure 20**, we show the Bray-Curtis dissimilarity between CLARK and Kraken 2 for each year and taxonomic level. Interestingly, both make the same classifications in 1999 and are pretty similar. In 2004, the dissimilarity is mainly because CLARK classifies many more percentages of sequences than Kraken 2 (which the Bray-Curtis measure takes into consideration). CLARK classifying more sequences makes the methods more discordant at higher levels of taxonomic tree in 2009. Since the family level has many more classes than order, class, phyla, the Bray-Curtis values are very high for cases where taxa exists in one classifier but not the other, and this is much less likely with less classes at higher taxonomic levels. Taxa that are uniquely classified by each method also contribute to dissimilarity, but they are not the main contributors to the large dissimilarity value. Interestingly, after 2004, in time, both methods then become more concordant with increasing

knowledge, with some deviation in species and more concordance on the phylum level. Now, in the latest 2020 database update, the results are slightly more discordant than in 2009, despite having many more taxa classes, showing that the methods are able to agree when the **relevant** gut taxa that is “truly” in the sample is added to the database.

DISCUSSION

CLARK’s method of only comparing unique k-mers from a read to its target genomes may have been what aided its classification time but hindered its classification percentage. Because CLARK ignores any k-mer in a read that is shared between two or more targets (genomes in its database), it can work through data more quickly. This method seems to allow it to eliminate unlikely matches more efficiently. However, this also seems to make it harder to match k-mers uniquely to genomes that are closely related (ie. species level). The elimination of targets that have one of the read’s k-mers in common could cause CLARK to eliminate many genomes from being possible matches, most likely ending up with no more targets to compare and resulting in an unclassified read. This could explain why CLARK’s classification percentage decreased between the 2019 and 2020 databases so drastically: as more and more similar species were added, it became harder and harder for CLARK to match unique k-mers to them.

Despite CLARK having more genomes in its database to get through, it still classified faster than Kraken 2 under normalized circumstances. This could be due to how it decides which genome best matches the read. CLARK may be classifying faster because it only has to keep track of the unique k-mers in a read and compare its targets to those, while ignoring the common k-mers. While Kraken 2 has to keep track of all the common k-mers for every genome it is comparing.

In the future, a further study should be conducted with carefully designed mock communities or simulated communities with CAMISIM (Fritz et al., 2019) to make sure a carefully balanced novel/known set is contained in the training/test sets. However, much is still not known about the underlying k-mer distribution of novel organisms and their frequency.

CONCLUSION

In this paper, we compose a framework in which to compare metagenomic taxonomic classifiers, in terms of their computational time and classification agreement on a real metagenomic sample. We studied hash-based methods and found that a technique that eliminates common k-mers, CLARK, classifies more and faster (at a cost of longer training time) when trained on smaller and more diverse databases. However, the percent of the sample that it can classify starts to degrade for large databases. Kraken 2, on the other hand, gains percent classified and significantly takes less time building and classifying with more training data. Both methods’ agreement on classification labels tend to converge as the database knowledge in

each grows, and the database differences can cause some divergence between the two methods' classifications. The recommendation from our study is that Kraken 2 tends to scale better with more data. However, we recommend for future studies to extend this study and compare many methods' scalability in terms of time, percent classified, and agreement with the experimental framework that we introduce here.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

MG wrote the original draft, assisted in conceptualization, and implemented all the methodology, analysis, validation, and

visualization. ZZ organized and curated the data(bases) and assisted in methodology, analysis, validation, and visualization. GR conceptualized the study, acquired funding, supervised and coordinated the project, assisted with the methodology with analysis, methodology, and visualization and assisted in writing the original draft.

FUNDING

This work was supported by NSF grants #1919691, #1936791, and #2107108 that supported students and computer infrastructure needed for the project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frsip.2022.842513/full#supplementary-material>

REFERENCES

- Alshawaqfeh, M. K. (2017). Signal Processing and Machine Learning Techniques for Analyzing Metagenomic Data. Thesis. College Station, TX: Texas A&M. Available at: <https://oaktrust.library.tamu.edu/handle/1969.1/161461>.
- Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M.-C., Charles, T., et al. (2020). Microbiome Definition Re-Visited: Old Concepts and New Challenges. *Microbiome* 8 (1), 103. doi:10.1186/s40168-020-00875-0
- Borrayo, E., Mendizabal-Ruiz, E. G., Vélez-Pérez, H., Romo-Vázquez, R., Mendizabal, A. P., and Morales, J. A. (2014). Genomic Signal Processing Methods for Computation of Alignment-Free Distances from DNA Sequences. *PLOS ONE* 9 (11), e110954. doi:10.1371/journal.pone.0110954
- Brown, T., and Irber, L. (2016). Sourmash: A Library For Minhash Sketching of DNA. *J. Open Source Softw.* 1 (5), 27. doi:10.21105/joss.00027
- Brul, S., Kallemeyn, W., and Smits, G. (2010). Functional Genomics for Food Microbiology: Molecular Mechanisms of Weak Organic Acid Preservative Adaptation in Yeast. *CAB Rev.: Perspect. Agric. Vet. Sci. Nutrit. Nat. Resources* 3 (January), 1–14. doi:10.1079/PAVSNNR20083005
- Coenen, A. R., Hu, S. K., Luo, E., Muratore, D., and Weitz, J. S. (2020). A Primer for Microbiome Time-Series Analysis. *Front. Genet.* 11, 310. doi:10.3389/fgene.2020.00310
- Creasy, H. H., Felix, V., Aluvathingal, J., Crabtree, J., Ifeonu, O., Matsumura, J., et al. (2021). HMPDACC: A Human Microbiome Project Multi-Omic Data Resource. *Nucleic Acids Res.* 49 (D1), D734–D742. doi:10.1093/nar/gkaa996
- dibsi-rnaseq (2016). Sourmash Website. Available at: <https://dibsi-rnaseq.readthedocs.io/en/latest/kmersand-sourmash.html> (Accessed May 16, 2022).
- Elworth, R. A. L., Wang, Q., KotaKota, P. K., Barberan, C. J., Coleman, B., Balaji, A., et al. (2020). To Petabytes and Beyond: Recent Advances in Probabilistic and Signal Processing Algorithms and Their Application to Metagenomics. *Nucleic Acids Res.* 48 (10), 5217–5234. doi:10.1093/nar/gkaa265
- Figueiredo, A. R. T., and Kramer, J. (2020). Cooperation and Conflict within the Microbiota and Their Effects on Animal Hosts. *Front. Ecol. Evol.* 8, 132. doi:10.3389/fevo.2020.00132
- Fritz, A., Hofmann, P., Majda, S., Dahms, E., Dröge, J., Fiedler, J., et al. (2019). CAMISIM: Simulating Metagenomes and Microbial Communities. *Microbiome* 7, 17. doi:10.1186/s40168-019-0633-6
- Garbarine, E., DePasquale, J., Gadia, V., Polikar, R., and Rosen, G. (2011). Information-Theoretic Approaches to SVM Feature Selection for Metagenome Read Classification. *Comput. Biol. Chem.* 35 (3), 199–209. doi:10.1016/j.compbiolchem.2011.04.007
- Gardner, P. P., Watson, R. J., Draper, J. L., Finn, R. D., Morales, S. E., and Stott, M. B. (2019). Identifying Accurate Metagenome and Amplicon Software via a Meta-Analysis of Sequence to Taxonomy Benchmarking Studies. *PeerJ* 7, e6160. doi:10.7717/peerj.6160
- Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., et al. (2012). Structure, Function and Diversity of the Healthy Human Microbiome. *Nature* 486 (7402), 207–214. doi:10.1038/nature11234
- Kouchaki, S., Tapinos, A., and Robertson, D. L. (2019). A Signal Processing Method for Alignment-Free Metagenomic Binning: Multi-Resolution Genomic Binary Patterns. *Sci. Rep.* 9 (1), 2159. doi:10.1038/s41598-018-38197-9
- Lan, Y., Morrison, J. C., Hershberg, R., and Rosen, G. L. (2014). POGO-DB-a Database of Pairwise-Comparisons of Genomes and Conserved Orthologous Genes. *Nucl. Acids Res.* 42 (D1), D625–D632. doi:10.1093/nar/gkt1094
- LaPierre, N., Alser, M., Eskin, E., Koslicki, D., and Mangul, S. (2020). Metalign: Efficient Alignment-Based Metagenomic Profiling via Containment Min Hash. *Genome Biol.* 21, 242. doi:10.1186/s13059-020-02159-0
- Liu, S., and Koslicki, D. (2021). CMash: Fast, Multi-Resolution Estimation of K-Mer-Based Jaccard and Containment Indices. *Biorxiv*. doi:10.1101/2021.12.06.47143
- McIntyre, A. B. R., Ounit, R., Afshinnekoo, E., Prill, R. J., Hénaff, E., Alexander, N., et al. (2017). Comprehensive Benchmarking and Ensemble Approaches for Metagenomic Classifiers. *Genome Biol.* 18, 182. doi:10.1186/s13059-017-1299-7
- Meyer, F., Fritz, A., Deng, Z.-L., Koslicki, D., Gurevich, A., Robertson, G., et al. (2021). Critical Assessment of Metagenome Interpretation - The Second Round of Challenges. *Biorxiv*. Available at: <https://www.biorxiv.org/content/10.1101/2021.07.12.451567v1>.
- Nasko, D. J., Koren, S., Phillippy, A. M., and Treangen, T. J. (2018). RefSeq Database Growth Influences the Accuracy of K-Mer-Based Lowest Common Ancestor Species Identification. *Genome Biol.* 19, 165. doi:10.1186/s13059-018-1554-6
- Nemergut, D. R., Schmidt, S. K., Fukami, T., O'Neill, S. P., Bilinski, T. M., Stanish, L. F., et al. (2013). Patterns and Processes of Microbial Community Assembly. *Microbiol. Mol. Biol. Rev.* 77 (3), 342–356. doi:10.1128/MMBR.00051-12
- Ounit, R., Wanamaker, S., Close, T. J., and Lonardi, S. (2015). CLARK: Fast and Accurate Classification of Metagenomic and Genomic Sequences Using Discriminative K-Mers. *BMC Genomics* 16 (1), 236. doi:10.1186/s12864-015-1419-2
- Rosen, G. L., and Moore, J. D. (2003). "Investigation of Coding Structure in DNA," in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hong Kong.

- Rosen, G., Sokhansanj, B., Polikar, R., Bruns, M., Russell, J., Garbarine, E., et al. (2009). Signal Processing for Metagenomics: Extracting Information from the Soup. *Curr. Genomics* 10 (7), 493–510. doi:10.2174/138920209789208255
- Sayers, E. W., Agarwala, R., Bolton, E. E., Brister, J. R., Canese, K., Clark, K., et al. (2019). Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 47 (D1), D23–D28. doi:10.1093/nar/gky1069
- Scipy (2021). Scipy.Spatial.Distance.Braycurtis — SciPy v1.7.1 Manual. Available at: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.braycurtis.html> (Accessed December 1, 2021).
- Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., et al. (2017). Critical Assessment of Metagenome Interpretation-A Benchmark of Metagenomics Software. *Nat. Methods* 14, 1063–1071. doi:10.1038/nmeth.4458
- Sender, R., Fuchs, S., and Milo, R. (2016). Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell* 164 (3), 337–340. doi:10.1016/j.cell.2016.01.013
- Shi, L., and Chen, B. (2021). “LSHvec: A Vector Representation of DNA Sequences Using Locality Sensitive Hashing and Fasttext Word Embeddings,” in Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '21), New York, NY, USA, August 1–4, 2021 (Association for Computing Machinery). doi:10.1145/3459930.3469521
- Woloszynek, S., Pastor, S., Mell, J. C., Nandi, N., Sokhansanj, B., Rosen, G. L., et al. (2016). “Engineering Human Microbiota: Influencing Cellular and Community Dynamics for Therapeutic Applications,” in *International Review Of Cell And Molecular Biology* (Cambridge, MA: Academic Press), 324, 67–124. doi:10.1016/bs.ircmb.2016.01.003
- Woloszynek, S., Zhao, Z., Chen, J., Gail, L., Woloszynek, S., Zhao, Z., et al. (2019). 16S rRNA Sequence Embeddings: Meaningful Numeric Feature Representations of Nucleotide Sequences that Are Convenient for Downstream Analyses. *PLoS Comput. Biol.* 15 (2), e1006721. doi:10.1371/journal.pcbi.1006721
- Wood, D. E., and Salzberg, S. L. (2014). Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments. *Genome Biol.* 15 (3), R46. doi:10.1186/gb-2014-15-3-r46
- Wood, D. E., Lu, J., Ben, L., Wood, D. E., Lu, J., and Langmead, B. (2019). Improved Metagenomic Analysis with Kraken 2. *Genome Biol.* 20 (1), 257. doi:10.1186/s13059-019-1891-0
- Ye, S. H., Siddle, K. J., Park, D. J., and Sabeti, P. C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell.* 178 (4), 779–794. doi:10.1016/j.cell.2019.07.010
- Zhao, Z., Cristian, A., Rosen, G., Zhao, Z., Cristian, A., and Rosen, G. (2020). Keeping Up with the Genomes: Efficient Learning of Our Increasing Knowledge of the Tree of Life. *BMC Bioinforma.* 21, 412. doi:10.1186/s12859-020-03744-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Gray, Zhao and Rosen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Adaptive Discrete Motion Control for Mobile Relay Networks

Spilios Evmorfos^{1†}, Dionysios Kalogieras^{2†} and Athina Petropulu^{1*†}

¹Electrical and Computer Engineering, Rutgers, The State University of New Jersey, New Brunswick, NJ, United States, ²Electrical Engineering, Yale University, New Haven, CT, United States

OPEN ACCESS

Edited by:

Monica Bugallo,
Stony Brook University, United States

Reviewed by:

Francesco Palmieri,
University of Campania Luigi Vanvitelli,
Italy

Stefania Colonnese,
Sapienza University of Rome, Italy

*Correspondence:

Athina Petropulu
athinap@rutgers.edu

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Signal Processing for
Communications,
a section of the journal
Frontiers in Signal Processing

Received: 01 February 2022

Accepted: 01 June 2022

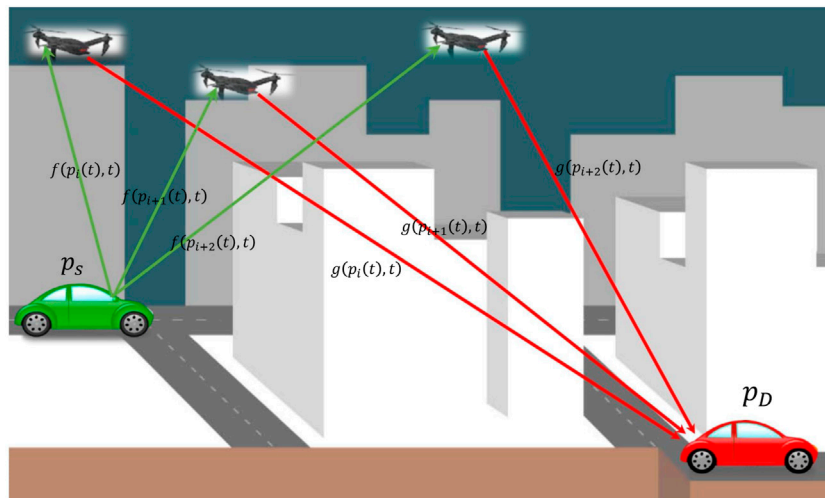
Published: 06 July 2022

Citation:

Evmorfos S, Kalogieras D and
Petropulu A (2022) Adaptive Discrete
Motion Control for Mobile
Relay Networks.
Front. Sig. Proc. 2:867388.
doi: 10.3389/frsip.2022.867388

We consider the problem of joint beamforming and discrete motion control for mobile relaying networks in dynamic channel environments. We assume a single source-destination communication pair. We adopt a general time slotted approach where, during each slot, every relay implements optimal beamforming and estimates its optimal position for the subsequent slot. We assume that the relays move in a 2D compact square region that has been discretized into a fine grid. The goal is to derive discrete motion policies for the relays, in an adaptive fashion, so that they accommodate the dynamic changes of the channel and, therefore, maximize the Signal-to-Interference + Noise Ratio (SINR) at the destination. We present two different approaches for constructing the motion policies. The first approach assumes that the channel evolves as a Gaussian process and exhibits correlation with respect to both time and space. A stochastic programming method is proposed for estimating the relay positions (and the beamforming weights) based on causal information. The stochastic program is equivalent to a set of simple subproblems and the exact evaluation of the objective of each subproblem is impossible. To tackle this we propose a surrogate of the original subproblem that pertains to the Sample Average Approximation method. We denote this approach as model-based because it adopts the assumption that the underlying correlation structure of the channels is completely known. The second method is denoted as model-free, because it adopts no assumption for the channel statistics. For the scope of this approach, we set the problem of discrete relay motion control in a dynamic programming framework. Finally we employ deep Q learning to derive the motion policies. We provide implementation details that are crucial for achieving good performance in terms of the collective SINR at the destination.

Keywords: relay networks, discrete motion control, stochastic programming, dynamic programming, deep reinforcement learning



GRAPHICAL ABSTRACT |

1 INTRODUCTION

In distributed relay beamforming networks, spatially distributed relays synergistically support the communication between a source and a destination (Havary-Nassab et al., 2008a; Li et al., 2011; Liu and Petropulu, 2011). The concepts of distributed beamforming hold the promise of extending the communication range and of minimizing the transmit power that is being wasted by being scattered to unwanted directions (Barriac et al., 2004).

Intelligent node mobility has been studied as a means of improving the Quality-of-Service (QoS) in communications. In (Chatzipanagiotis et al., 2014), the interplay of relay motion control and optimal transmit beamforming is considered with the goal of minimizing the relay transmit power, subject to a QoS-related constraint. In (Kalogerias et al., 2013), optimal relay positioning in the presence of an eavesdropper is considered, aiming to maximize the secrecy rate. In the context of communication-aware robotics, motion has been controlled with the goal of maintaining in-network connectivity (Yan and Mostofi, 2012; Yan and Mostofi, 2013; Muralidharan and Mostofi, 2017).

In this work, we examine the problem of optimizing the sequence of relay positions (relay trajectory) and the beamforming weights so that some SINR-based metric is maximized at the destination. The assumption that we adopt is that the channel evolves as a stochastic process that exhibits spatiotemporal correlations. Intrinsically, optimal relay positioning requires the knowledge of the Channel State Information (CSI) in all candidate positions at a future time instance. This is almost impossible to achieve since the channel varies with respect to time and space. Nonetheless, since the channel exhibits spatiotemporal correlations (induced by the shadowing propagation effect (Goldsmith, 2005; MacCartney et al., 2013) that is prominent in urban environments), it can

be, explicitly or implicitly, predicted. We follow two different directions, when it comes to the discrete relay motion control.

The first direction (Kalogerias and Petropulu, 2018; Kalogerias and Petropulu, 2016) (we call it model-based) pertains to the formulation of a stochastic program that computes the beamforming weights and the subsequent relay positions, so that some SINR-based metric at the destination is maximized, subject to a total relay power budget, assuming the availability of causal CSI information. This 2-stage problem is equivalent to a set of 2-stage subproblems that can be solved in distributed fashion, one by each relay. The objective of each subproblem is impossible to be analytically evaluated, so an efficient approximation is proposed. This approximation acts as a surrogate to the initial objective. The surrogate relies on the Sample Average Approximation (SAA) (Shapiro et al., 2009). The term “model-based” is not to be confused with model-based reinforcement learning. We just use it because this method (or direction rather) assumes complete knowledge of the underlying correlation structure of the channels, so it is helpful formalism to distinguish this method from the second approach that makes no particular assumption for the channel statistics.

The second direction (Evmorfos et al., 2021a; Evmorfos et al., 2021b; Evmorfos et al., 2022) tackles the problem of discrete relay motion control from a dynamic programming viewpoint. We formulate the Markov Decision Process (MDP), that is induced by the problem of controlling the motion. Finally, we employ deep Q learning (Mnih et al., 2015) to find relay motion policies that maximize the sum of SINRs at the destination over time. We propose a pipeline for adapting deep Q learning for the problem at hand. We experimentally show that Multilayer Perceptron Neural Networks (MLPs) cannot capture high frequency components in natural signals (in low-dimensional domains). This phenomenon, referred to as “Spectral Bias” (Jacot et al., 2018) has been observed in several contexts, and also arises as an issue in the adaptation of deep Q learning for the relay motion

control. We present an approach to tackle spectral bias, by parameterizing the Q function with a Sinusoidal Representation Network (SIREN) (Sitzmann et al., 2020).

Our intentions for this work lie in two directions. First, we attempt to compare two methods for relay motion control in urban communication environments. The two methods constitute two different viewpoints in terms of tackling the problem. The first method assumes complete knowledge on the underlying statistics of the channels (model-based) Kalogerias and Petropulu, (2018). The second method is completely model-free in the sense that it drops all assumptions for knowledge of the channel statistics and employs deep reinforcement learning to control the relay motion Evmorfos et al. (2022). In addition to the head-to-head comparison, we propose a slight variation of the model-free method that deviates from the one in Evmorfos et al. (2022) by augmenting the state with the addition of the timestep as an extra feature. This variation is more robust than the previous one, especially when the shadowing component of the urban environment is particularly strong.

Notation: We denote the matrices and vectors by bold uppercase and bold lowercase letters, respectively. The operators $(\cdot)^T$ and $(\cdot)^H$ denote transposition and conjugate transposition respectively. Caligraphic letters will be used to denote sets and formal script letters will be used to denote σ -algebras. The ℓ_p -norm of $\mathbf{x} \in \mathbb{R}^n$ is $\|\mathbf{x}\|_p \triangleq (\sum_{i=1}^n |x(i)|^p)^{1/p}$, for all $\mathbb{N} \ni p \geq 1$. For $\mathbb{N} \ni N \geq 1$, \mathbb{S}^N , \mathbb{S}_{+}^N will denote the sets of symmetric and symmetric positive (semidefinite) matrices, respectively. The finite N -dimensional identity operator will be denoted as \mathbf{I}_N . Additionally, we define $\mathfrak{J} \triangleq \sqrt{-1}$, $\mathbb{N}^+ \triangleq \{1, 2, \dots\}$, $\mathbb{N}_n^+ \triangleq \{1, 2, \dots, n\}$, $\mathbb{N}_n \triangleq \{0\} \cup \mathbb{N}_n^+$ and $\mathbb{N}_n^m \triangleq \mathbb{N}_n^+ \setminus \mathbb{N}_{m-1}^+$, for positive naturals $n > m$.

2 PROBLEM FORMULATION

2.1 System Model

Consider a scenario where source S, located at position $\mathbf{p}_S \in \mathbb{R}^2$, wishes to communicate with user D, located at $\mathbf{p}_D \in \mathbb{R}^2$ but does not have enough power to do so, or due to the topography, cannot communicate in a line-of-sight (LoS) fashion. Therefore, R single-antenna, trusted mobile relays are enlisted to support the communication. The relays are deployed over a two-dimensional space, which is partitioned into $M \times M$ imaginary grid cells. Time evolves in a time-slotted fashion, where T is the slot duration, and t denotes the current time slot. In every time slot, a grid cell can be occupied by at most one relay.

Source S transmits symbol $s(t) \in \mathbb{C}$, where $\mathbb{E}[|s(t)|^2] = 1$, using power $\sqrt{P_S} > 0$. Let us drop for notational simplicity the relay position dependence on t . The signal received by relay R_r , located at $\mathbf{p}_r(t)$, $r = 1, \dots, R$, equals

$$x_r(t) = \sqrt{P_S} f_r(\mathbf{p}_r, t) s(t) + n_r(t),$$

where f_r denotes the flat fading channel from S to relay R_r , and $n_r(t)$ denotes reception noise at relay R_r , with $\mathbb{E}[|n_r(t)|^2] = \sigma^2$, $r = 1, \dots, R$.

Each relay operates in an Amplify-and-Forward (AF) fashion, i.e., it transmits received signal, $x_r(t)$, multiplied by weight $w_r(t) \in \mathbb{C}$. Due to the relays' simultaneous transmissions, the destination D receives

$$y(t) = \sum_{r=1}^R g_r(\mathbf{p}_D, t) w_r(t) x_r(t) + n_D(t),$$

where g_r denotes the flat fading channel from relay R_r to destination D, and $n_D(t)$ denotes reception noise at D. We assume here that $\mathbb{E}[|n_D(t)|^2] = \sigma_D^2$ $y(t)$ can be rewritten as

$$y(t) = \underbrace{\sum_{r=1}^R g_r(\mathbf{p}_D, t) w_r(t) \sqrt{P_S} f_r(\mathbf{p}_r, t) s(t)}_{\text{desired signal}} + \underbrace{\sum_{r=1}^R g_r(\mathbf{p}_D, t) w_r(t) n_r(t)}_{\text{noise}} + n_D(t) \\ \triangleq y_{\text{signal}}(t) + y_{\text{noise}}(t),$$

where $y_{\text{signal}}(t)$ is the received signal component and $n_D(t)$ represents noise at the destination.

In the following, we will use the vector $\mathbf{p}(t) \triangleq [\mathbf{p}_1^T(t) \mathbf{p}_2^T(t) \dots \mathbf{p}_R^T(t)]^T$, to collect the positions of all relays at time t .

2.2 Channel Model

The channel evolves in time and space and can be described in statistical terms. In particular, during time slot t , the channel between the source and a relay positioned at $\mathbf{p}_r \in \mathbb{R}^2$ can be modeled as the product of four components (Heath, 2017), i.e.,

$$f_r(\mathbf{p}_r, t) \triangleq f_r^{PL}(\mathbf{p}_r) f_r^{SH}(\mathbf{p}_r, t) f_r^{MF}(\mathbf{p}_r, t) e^{j2\pi\phi(t)}, \quad (1)$$

where $f_r^{PL}(\mathbf{p}_r) \triangleq \|\mathbf{p}_r - \mathbf{p}_S\|_2^{-\ell/2}$ is the path-loss component with path-loss exponent ℓ ; $f_r^{SH}(\mathbf{p}_r, t)$ the shadow fading component; $f_r^{MF}(\mathbf{p}_r, t)$ the multi-path fading component; and $e^{j2\pi\phi(t)}$, with ϕ uniformly distributed in $[0, 1]$, a phase term. A similar model holds for the relay-destination channel $g_r(\mathbf{p}_r, t)$.

The logarithm of the squared channel magnitude of Eq. 1 converts the multiplicative channel model into an additive one, i.e.,

$$F_r(\mathbf{p}_r, t) \triangleq 10 \log_{10}(|f_r(\mathbf{p}_r, t)|^2) \\ \triangleq \alpha_r^f(\mathbf{p}_r) + \beta_r^f(\mathbf{p}_r, t) + \xi_r^f(\mathbf{p}_r, t),$$

with

$$\alpha_r^f(\mathbf{p}_r) \triangleq -\ell 10 \log_{10}(\|\mathbf{p}_r - \mathbf{p}_S\|_2), \\ \beta_r^f(\mathbf{p}_r, t) \triangleq 10 \log_{10}(|f_r^{SH}(\mathbf{p}_r, t)|^2) \sim \mathcal{N}(0, \eta^2), \quad \text{and} \\ \xi_r^f(\mathbf{p}_r, t) \triangleq 10 \log_{10}(|f_r^{MF}(\mathbf{p}_r, t)|^2) \sim \mathcal{N}(\rho, \sigma_\xi^2),$$

where η^2 is the shadowing power, and ρ, σ_ξ^2 are the mean and variance of multipath fading component, respectively.

The multipath fading component, $\xi_r^f(\mathbf{p}_r, t)$, varies fast in time and space, and is typically modeled as i. i. d. between different positions and times. On the other hand, the shadowing component, $\beta_r^f(\mathbf{p}_r, t)$, induced by relatively large and slowly moving objects in the path of the signal, exhibits correlation between any two positions \mathbf{p}_i and \mathbf{p}_j , and between any two time slots t_a and t_b , as (Kalogerias and Petropulu, 2018)

$$\mathbb{E}[\beta_r^f(\mathbf{p}_i, t_a)\beta_r^f(\mathbf{p}_j, t_b)] = \tilde{\Sigma}^f(\mathbf{p}_i, \mathbf{p}_j)e^{-\frac{|t_a-t_b|}{c_2}},$$

where

$$\tilde{\Sigma}^f(\mathbf{p}_i, \mathbf{p}_j) \triangleq \eta^2 e^{-\|\mathbf{p}_i - \mathbf{p}_j\|_2 / c_1} \in \mathbb{R}^{M^2 \times M^2},$$

with c_1 denoting the correlation distance, and c_2 the correlation time. Similar correlations hold for similarly $\beta_r^g(\mathbf{p}_i, t)$.

Further, $\beta_r^f(\mathbf{p}_i, t)$ and $\beta_r^g(\mathbf{p}_i, t)$ exhibit correlations as

$$\mathbb{E}[\beta_r^f(\mathbf{p}_i, t_a)\beta_r^g(\mathbf{p}_j, t_b)] = \tilde{\Sigma}^{fg}(\mathbf{p}_i, \mathbf{p}_j)e^{-\frac{|t_a-t_b|}{c_2}},$$

where

$$\tilde{\Sigma}^{fg}(\mathbf{p}_i, \mathbf{p}_j) = \tilde{\Sigma}^f(\mathbf{p}_i, \mathbf{p}_j)e^{-\frac{\|\mathbf{p}_i - \mathbf{p}_j\|_2}{c_3}}$$

and c_3 denoting the correlation distance of the source-destination channel (Kalogerias and Petropulu, 2018).

2.3 Joint Scheduling of Communications and Controls

Let us assume the same carrier for all communication tasks, and employ a basic joint communication/decision making TDMA-like protocol. At each time slot $t \in \mathbb{N}_{N_T}^+$, the following actions are taken:

1. The source broadcasts a pilot signal to all relays, based on which the relays estimate their channels to the source.
2. The destination also broadcasts pilots, which the relays use to estimate their channels relative to the destination.
3. Then, based on the estimated channels, the relays beamform in AF mode. Here we assume perfect CSI estimation.
4. Based on the CSI that has been received up to that point, a decision is made on where the relays need to go to, and relay motion controllers are determined to steer the relays to those positions.

The above steps are repeated for N_T time slots. Let us assume that the relays pass their estimated CSI to the destination via a dedicated low-rate channel. This simplifies information decoding at the destination (Gao et al., 2008; Proakis and Salehi, 2008).

Concerning relay motion, we assume that the relays obey the differential equation (Kalogerias and Petropulu, 2018)

$$\dot{\mathbf{p}}(\tau) \equiv \mathbf{u}(\tau), \quad \forall \tau \in [0, T],$$

where $\mathbf{u} \triangleq [\mathbf{u}_1 \dots \mathbf{u}_R]^T$, with $\mathbf{u}_i: [0, T]$ being the motion controller of relay $i \in \mathbb{N}_R^+$. Assuming the relays may move only after their controls have been determined and their movement must be completed before the start of the next time slot, we can write (Kalogerias and Petropulu, 2018)

$$\mathbf{p}(t) \equiv \mathbf{p}(t-1) + \int_{\Delta\tau_{t-1}} \mathbf{u}_{t-1}(\tau) d\tau, \quad \forall t \in \mathbb{N}_{N_T}^2,$$

with $\mathbf{p}(1) \equiv \mathbf{p}_{init}$, and where $\Delta\tau_t \subset \mathbb{R}$ and \mathbf{u}_t denote the time interval that the relays are allowed to move in, and the respective relay controller, in each time slot $t \in \mathbb{N}_{N_T-1}^+$. It holds that

$\mathbf{u}(\tau) \equiv \sum_{t \in \mathbb{N}_{N_T-1}^+} \mathbf{u}_t(\tau) 1_{\Delta\tau_t}(\tau)$, where τ belongs in the first $N_T - 1$ time slots. In each time slot t , the length of $\Delta\tau_t$, $|\Delta\tau_t|$, must be small enough, so that the shadowing correlation at adjacent time slots is strong enough. These correlations are controlled by parameter γ , which can be function of the slot width. Thus, relay velocity must be of the order of $(|\Delta\tau_t|)^{-1}$. For simplicity, here we assume that the relays are not resource constrained when they move and they are only limited by their transmission power.

To determine the relay motion controller $\mathbf{u}_{t-1}(\tau)$, $\tau \in \Delta\tau_{t-1}$, given a goal position vector at time slot t , $\mathbf{p}^o(t)$, it suffices to decide on a path in \mathcal{S}^R , such that the points $\mathbf{p}^o(t)$ and $\mathbf{p}(t-1)$ are connected in at most time $|\Delta\tau_{t-1}|$. Assuming the simplest path, i.e., a straight line between $\mathbf{p}_i^o(t)$ and $\mathbf{p}_i(t-1)$, for all $i \in \mathbb{N}_R^+$, the relay controllers at time slot $t-1 \in \mathbb{N}_{N_T-1}^+$ is

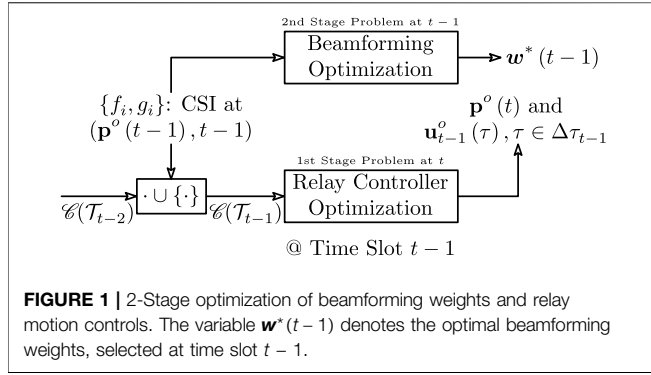
$$\mathbf{u}_{t-1}^o(\tau) \triangleq \frac{1}{|\Delta\tau_{t-1}|} (\mathbf{p}^o(t) - \mathbf{p}(t-1)), \quad \forall \tau \in \Delta\tau_{t-1}.$$

Based on the above, the motion control problem can be formulated in terms of specifying the relay positions at the next time slot, given the relay positions at the current time slot and the estimated CSI. We assume here for simplicity that there exists some path planning and collision avoidance mechanism, the derivation of which is out of the scope of this paper.

For simplicity and tractability, we are assuming that the channel is the same for every position *within* each grid cell, and for the duration of each time slot. In other words, we are essentially adopting a *time-space block fading model*, at least for motion control purposes. This is a valid approximation of reality as the grid cell size and the time slot duration become smaller, at the expense of more stringent resource constraints at the relays, and faster channel sensing capability. Under this setting, *communication and relay control can indeed happen simultaneously* within each time slot, with the understanding that at the start of the next time slot, each relay must have completed their motion (starting at the previous time slot—also see our discussion earlier in this section—). In this way, our approach is valid in a practical setting where communication needs to be continuous and uninterrupted.

Additionally, we are assuming that the relays move sufficiently slowly, such that the local spatial and temporal changes of the wireless channel due to relay motion itself are negligible, e.g., Doppler shift effects. Then, spatial and temporal variations in channel quality are only due to changes in the physical environment, which happen at a much slower rate than that of actual communication. Note that this is a standard requirement for achieving a high communication rate, whatsoever.

We see that there is a natural interplay between relay velocity and the relative rate of change of the communication channel (Kalogerias and Petropulu (2018)). The challenge here is to identify a fair tradeoff between a reasonable relay velocity, grid size and a time slot, which would enable simultaneously faithful channel prediction and feasible and effective motion control (adhering to potential relay motion constraints). The width of the communication time slot depends on the spatial characteristics



of the terrain, which varies with each application. This also determines the sampling rate employed for identifying the parameters of the adopted channel model. In theory, for a given relay velocity, the relays could move to any position up to which the channel remains correlated. However, as the per time slot rate of communications depends on the relay velocity (characterizing system throughput), the relays should move to much smaller distances within the slot.

In the following we use $\mathcal{C}(\mathcal{T}_t)$ to denote the set of channel gains observed by the relays, along their trajectories $\mathcal{T}_t \triangleq \{\mathbf{p}(1) \dots \mathbf{p}(t)\}$, $t \in \mathbb{N}_{N_T}^+$. Then, \mathcal{T}_t may be recursively updated as $\mathcal{T}_t \equiv \mathcal{T}_{t-1} \cup \{\mathbf{p}(t)\}$, for all $t \in \mathbb{N}_{N_T}^+$, with $\mathcal{T}_0 \triangleq \emptyset$. In a more precise sense, $\{\mathcal{C}(\mathcal{T}_t)\}_{t \in \mathbb{N}_{N_T}^+}$ will also denote the filtration generated by the CSI observed at the relays, along \mathcal{T}_t , interchangeably. In other words, $\mathcal{C}(\mathcal{T}_t)$ denotes the information (i.e., the σ -algebra) generated by the CSI observed up to and including time slot t and $\mathbf{p}(1) \dots \mathbf{p}(t)$, for all $t \in \mathbb{N}_{N_T}^+$. By convention, we define $\mathcal{C}(\mathcal{T}_0) \equiv \mathcal{C}(\{\emptyset\})$ (i.e., as the trivial σ -algebra $\mathcal{C}(\mathcal{T}_0) \triangleq \{\emptyset, \Omega\}$), and we refer to time $t \equiv 0$, as a *dummy time slot*.

2.4 Spatially Controlled SINR Maximization at the Destination

Next, we present the first stage of the 2-stage generic formulation. The 2-stage approach optimizes network QoS by optimally selecting beamforming weights *and* relay positions, on a *per time slot* basis. In this subsection, we focus on the calculation of the beamforming weights. The calculation of the weights at each step remains the same both for the stochastic programming (model-based) method and the dynamic programming (model-free) method.

Optimization of Beamforming Weights: At time slot $t \in \mathbb{N}_{N_T}^+$, given CSI in $\mathcal{C}(\mathcal{T}_t)$, we formulate the problem (Havary-Nassab et al., 2008b; Zheng et al., 2009)

$$\begin{aligned} & \underset{\mathbf{w}(t) \triangleq [w_1(t), \dots, w_R(t)]^T}{\text{maximize}} && \frac{\mathbb{E}\{P_S(t) | \mathcal{C}(\mathcal{T}_t)\}}{\mathbb{E}\{P_{I+N}(t) | \mathcal{C}(\mathcal{T}_t)\}}, \\ & \text{subject to} && \mathbb{E}\{P_R(t) | \mathcal{C}(\mathcal{T}_t)\} \leq P_c \end{aligned} \quad (2)$$

where $P_R(t)$, $P_S(t)$ and $P_{I+N}(t)$ denote the random instantaneous power at the relays, the power of the signal component and the power of the interference plus noise at the

destination, respectively, and where $P_c > 0$ denotes the total relay transmission power budget. Based on the mutual independence of source and destination CSI, (Eq. 2) can be expressed as (Havary-Nassab et al., 2008b)

$$\begin{aligned} & \underset{\mathbf{w}(t)}{\text{maximize}} && \frac{\mathbf{w}^H(t) \mathbf{R}(\mathbf{p}(t), t) \mathbf{w}(t)}{\sigma_D^2 + \mathbf{w}^H(t) \mathbf{Q}(\mathbf{p}(t), t) \mathbf{w}(t)}, \\ & \text{subject to} && \mathbf{w}^H(t) \mathbf{D}(\mathbf{p}(t), t) \mathbf{w}(t) \leq P_c \end{aligned} \quad (3)$$

where, dropping the dependence on $(\mathbf{p}(t), t)$ or t for brevity,

$$\mathbf{D} \triangleq P_0 \text{diag}([|f_1|^2 |f_2|^2 \dots |f_R|^2]^T) + \sigma^2 \mathbf{I}_R \in \mathbb{S}_{++}^R,$$

$$\mathbf{R} \triangleq P_0 \mathbf{h} \mathbf{h}^H \in \mathbb{S}_+^R, \text{ with } \mathbf{h} \triangleq [f_1 g_1 f_2 g_2 \dots f_R g_R]^T \text{ and}$$

$$\mathbf{Q} \triangleq \sigma^2 \text{diag}([|g_1|^2 |g_2|^2 \dots |g_R|^2]^T) \in \mathbb{S}_{++}^R.$$

The optimization problem of Eq. 3 is *always feasible, as long as P_c is nonnegative*, and the optimal value of Eq. 3 can be expressed in closed form as (Havary-Nassab et al., 2008b)

$$\begin{aligned} V_t & \equiv V(\mathbf{p}(t), t) \\ & \triangleq P_c \lambda_{\max}((\sigma_D^2 \mathbf{I}_R + P_c \mathbf{D}^{-1/2} \mathbf{Q} \mathbf{D}^{-1/2})^{-1} \mathbf{D}^{-1/2} \mathbf{R} \mathbf{D}^{-1/2}), \end{aligned}$$

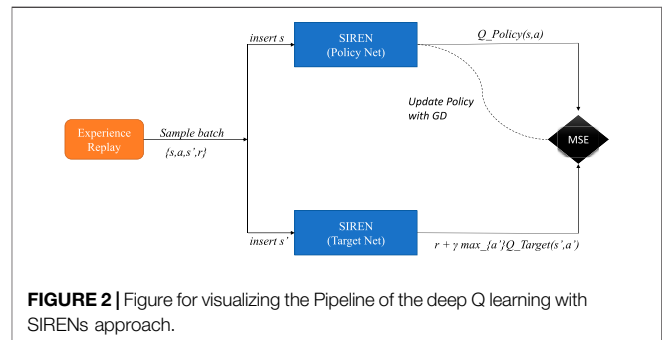
for all $t \in \mathbb{N}_{N_T}^+$, which can be further written as (Zheng et al., 2009)

$$\begin{aligned} V_t & \equiv \sum_{i \in \mathbb{N}_R^+} \frac{P_c P_0 |f(\mathbf{p}_i(t), t)|^2 |g(\mathbf{p}_i(t), t)|^2}{P_0 \sigma_D^2 |f(\mathbf{p}_i(t), t)|^2 + P_c \sigma^2 |g(\mathbf{p}_i(t), t)|^2 + \sigma^2 \sigma_D^2} \\ & \triangleq \sum_{i \in \mathbb{N}_R^+} V_i(\mathbf{p}_i(t), t), \quad \forall t \in \mathbb{N}_{N_T}^+. \end{aligned}$$

The above analytical expression of the optimal value V_t in terms of relay positions and their corresponding channel magnitudes will be key in our subsequent development.

3 STOCHASTIC PROGRAMMING FOR MYOPIC RELAY CONTROL

During time slot $t-1$, we need to determine the relay positions for time slot t , so that we achieve the maximum V_t . However, at time slot $t-1$, we only know $\mathcal{C}(\mathcal{T}_{t-1})$, which does not include information on the CSI that will be experienced during time slot t . Therefore, exactly optimizing the relay positions at the next time slot seems to be an impossible task.



Since deterministic optimization of V_t with respect to $\mathbf{p}(t)$ is not possible to be carried out during time slot $t - 1$, we can alternatively optimize a projection of V_t onto the space of all measurable functions of $\mathcal{C}(\mathcal{T}_{t-1})$ (Kalogerias and Petropulu, 2018). Since, for every $\mathbf{p}(t) \in \mathcal{S}^R$, V_t is of finite variance, we can consider orthogonal projections. In other words, we can consider the Minimum Mean-Square Error (MMSE) predictor of V_t given the available information $\mathcal{C}(\mathcal{T}_{t-1})$. We can then optimize the $\mathbb{E}\{V_t | \mathcal{C}(\mathcal{T}_{t-1})\}$ with respect to the point $\mathbf{p}(t)$, which results in the 2-stage stochastic program (Shapiro et al., 2009)

$$\begin{aligned} & \underset{\mathbf{p}(t)}{\text{maximize}} \quad \mathbb{E} \left\{ V_t \equiv \sum_{i \in \mathbb{N}_R^+} V_i(\mathbf{p}_i(t), t) \middle| \mathcal{C}(\mathcal{T}_{t-1}) \right\}, \\ & \text{subject to} \quad \mathbf{p}(t) \in \mathcal{C}(\mathbf{p}^o(t-1)) \end{aligned} \quad (4)$$

to be solved at time slot $t - 1 \in \mathbb{N}_{N_T-1}^+$, where $\mathbf{p}^o(1) \in \mathcal{S}^R$ is the initial positions of the relays and $\mathcal{C}(\mathbf{p}^o(t-1)) \subseteq \mathcal{S}^R$ denotes spatially feasible neighborhood around point $\mathbf{p}^o(t-1) \in \mathcal{S}^R$, which is the optimal decision vector determined at time slot $t - 2 \in \mathbb{N}_{N_T-2}$. For example, \mathcal{C} may be such that it does not allow the relays to collide with each other, or with other obstacles in space at their next slot positions. In general, \mathcal{C} depends on t , but here, for simplicity that dependence is not shown.

The map $\mathcal{C}(\cdot)$ is typically referred to as *finite-valued multifunction*, and we write $\mathcal{C}: \mathcal{S}^R \rightrightarrows \mathcal{S}^R$ (Shapiro et al., 2009). Additionally, problems (4) and (3) are referred to as the *first-stage problem* and the *second-stage problem*, respectively (Shapiro et al., 2009). The block diagram of the above described process is shown in **Figure 1**.

As compared to traditional AF beamforming for a static case, our spatially controlled system described above, uses the same CSI as in the stationary case, to predict the optimal beamforming performance in its vicinity in the MMSE sense, and moves to the optimally selected location. The prediction here relies on the aforementioned spatiotemporal channel model. Of course, this requires a sufficiently slowly varying channel relatively to relay motion, which can be guaranteed if the motion is constrained within small steps.

3.1 Motion Policies & the Interchangeability Principle

To assist in the process of understanding the techniques to solve **Eq. 4**, we make note of an important *variational* property of **Eq. 4**, related to the *long-term performance* of the proposed spatially controlled beamforming system. Our discussion pertains to the employment of the so-called *Interchangeability Principle (IP)* (Bertsekas and Shreve, 1978; Bertsekas, 1995; Rockafellar and Wets, 2004; Shapiro et al., 2009; Kalogerias and Petropulu, 2017), also known as the *Fundamental Lemma of Stochastic Control (FLSC)* (Astrom, 1970; Speyer and Chung, 2008) (Kalogerias and Petropulu, 2018). The IP refers conditions that allow the interchange of expectation and maximization or minimization in general stochastic programs.

A version of the IP for the first-stage problem of **(4)** is established in (Kalogerias and Petropulu, 2017). Specifically, the IP implies that **(4)** is exchangeable by the variational problem (Kalogerias and Petropulu, 2017)

$$\begin{aligned} & \underset{\mathbf{p}(t)}{\text{maximize}} \quad \mathbb{E}\{V_t\} \\ & \text{subject to} \quad \mathbf{p}(t) \in \mathcal{C}(\mathbf{p}^o(t-1)) \\ & \quad \mathbf{p}(t) \text{ is } \mathcal{C}(\mathcal{T}_{t-1})\text{-measurable} \end{aligned} \quad (5)$$

to be solved at each $t - 1 \in \mathbb{N}_{N_T-1}^+$. Upon comparing **Eq. 5** and the original problem **Eq. 4** one can see that, the former problem includes optimization of the *unconditional expectation* of V_t over all (measurable) mappings of the variables generating $\mathcal{C}(\mathcal{T}_{t-1})$ to $\mathcal{C}(\mathbf{p}^o(t-1))$. “This implies that, in **Eq. 5**, $\mathbf{p}(t)$ is a function of all CSI and motion controls up to and including time slot $t - 1$, whereas, in **Eq. 4**, $\mathbf{p}(t)$ is a *point*, since all variables generating $\mathcal{C}(\mathcal{T}_{t-1})$ are *fixed before decision making*. Aligned with the literature, any feasible decision $\mathbf{p}(t)$ in **Eq. 5** will be called an (*admissible*) *policy*, or a *decision rule*. *Exchangeability* of **Eqs. 4, 5** is understood in the sense that the optimal value of **Eq. 5**, which is a number, coincides with the *expectation* of the optimal value of **Eq. 4**, which is a measurable function of $\mathcal{C}(\mathcal{T}_{t-1})$ (and fixed for every realization of the variables generating $\mathcal{C}(\mathcal{T}_{t-1})$). In other words, maximization is *interchangeable* with integration, in the sense that” (Kalogerias and Petropulu, 2017)

$$\sup_{\mathbf{p}(t) \in \mathcal{D}_t} \mathbb{E}\{V_t\} \equiv \mathbb{E} \left\{ \sup_{\mathbf{p}(t) \in \mathcal{C}(\mathbf{p}^o(t-1))} \mathbb{E}\{V_t | \mathcal{C}(\mathcal{T}_{t-1})\} \right\},$$

for all $t \in \mathbb{N}_{N_T}^2$, where \mathcal{D}_t denotes the set of feasible decisions for **(Eq. 5)**. Furthermore, due to our assumption that the control space \mathcal{S} is finite, the IP guarantees that an optimal solution to the original stochastic program **(Eq. 4)** is also feasible and thus, optimal, for **(Eq. 5)**.

$$\mathbf{m}_{1:t-1} \triangleq [\mathbf{F}^T(1) \mathbf{G}^T(1) \dots \mathbf{F}^T(t-1) \mathbf{G}^T(t-1)]^T \in \mathbb{R}^{2R(t-1) \times 1} \quad (6)$$

$$\boldsymbol{\mu}_{1:t-1} \triangleq [\boldsymbol{\alpha}_S(\mathbf{p}(1)) \boldsymbol{\alpha}_D(\mathbf{p}(1)) \dots \boldsymbol{\alpha}_S(\mathbf{p}(t-1)) \boldsymbol{\alpha}_D(\mathbf{p}(t-1))]^T \ell \in \mathbb{R}^{2R(t-1) \times 1} \quad (7)$$

$$\mathbf{c}_{1:t-1}^{F(G)}(\mathbf{p}) \triangleq [\mathbf{c}_1^{F(G)}(\mathbf{p}) \dots \mathbf{c}_{t-1}^{F(G)}(\mathbf{p})] \in \mathbb{R}^{1 \times 2R(t-1)} \quad (8)$$

$$\mathbf{c}_k^{F(G)}(\mathbf{p}) \triangleq \left[\mathbb{E}\{\sigma_{S(D)}(\mathbf{p}, t) \sigma_S^i(k)\} \right]_{j \in \mathbb{N}_R^+} \left[\mathbb{E}\{\sigma_{S(D)}(\mathbf{p}, t) \sigma_D^j(k)\} \right]_{j \in \mathbb{N}_R^+} \in \mathbb{R}^{1 \times 2R}, \forall k \in \mathbb{N}_{t-1}^+ \quad (9)$$

$$\boldsymbol{\Sigma}_{1:t-1} \triangleq \begin{bmatrix} \boldsymbol{\Sigma}(1, 1) & \dots & \boldsymbol{\Sigma}(1, t-1) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}(t-1, 1) & \dots & \boldsymbol{\Sigma}(t-1, t-1) \end{bmatrix} \in \mathbb{S}_{++}^{2R(t-1)} \quad (10)$$

3.2 Near-Optimal Beamformer Motion Control

One can readily observe that the problem of **(4)** is separable. Given that, for each $t \in \mathbb{N}_{N_T-1}^+$, decisions taken and CSI collected so far are available to all relays, **(4)** can be solved in a distributed fashion at the relays, with the i th relay being responsible for solving the problem (Kalogerias and Petropulu, 2018)

$$\begin{aligned} & \underset{\mathbf{p}}{\text{maximize}} \quad \mathbb{E}\{V_I(\mathbf{p}, t) \mid \mathcal{C}(\mathcal{T}_{t-1})\}, \\ & \text{subject to} \quad \mathbf{p} \in C_i(\mathbf{p}^o(t-1)), \end{aligned} \quad (11)$$

at each $t-1 \in \mathbb{N}_{N_T-1}^+$, where $C_i: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ denotes the corresponding section of \mathcal{C} , for each $i \in \mathbb{N}_R^+$. Note that no local exchange of intermediate results is required among relays; given the available information, each relay independently solves its own subproblem. It is also evident that apart from the obvious difference in the feasible set, the optimization problems at each of the relays are identical.

However, the objective of problem **Eq. 11** is impossible to obtain analytically, and it is necessary to resort to some well behaved and computationally efficient *surrogates*. Next, we present a *near-optimal* such approach. The said approach relies on *global* function approximation techniques, and achieves excellent empirical performance.

The proposed approximation to the stochastic program (11) will be based on the following technical, though simple, result.

Lemma 1 (Big Expectations) (Kalogerias and Petropulu, 2018) *Under the assumptions of the wireless channel model, it is true that, at any $\mathbf{p} \in \mathcal{S}$,*

$$\left[\begin{array}{c} F(\mathbf{p}, t) \\ G(\mathbf{p}, t) \end{array} \right] \Big| \mathcal{C}(\mathcal{T}_{t-1}) \sim \mathcal{N}(\boldsymbol{\mu}_{t|t-1}^{F,G}(\mathbf{p}), \boldsymbol{\Sigma}_{t|t-1}^{F,G}(\mathbf{p})),$$

for all $t \in \mathbb{N}_{N_T}^2$, and where we define

$$\begin{aligned} \boldsymbol{\mu}_{t|t-1}^{F,G}(\mathbf{p}) &\triangleq [\boldsymbol{\mu}_{t|t-1}^F(\mathbf{p}) \boldsymbol{\mu}_{t|t-1}^G(\mathbf{p})]^T, \\ \boldsymbol{\mu}_{t|t-1}^F(\mathbf{p}) &\triangleq \alpha_S(\mathbf{p})\boldsymbol{\ell} + \mathbf{c}_{1:t-1}^F(\mathbf{p})\boldsymbol{\Sigma}_{1:t-1}^{-1}(\mathbf{m}_{1:t-1} - \boldsymbol{\mu}_{1:t-1}) \in \mathbb{R}, \\ \boldsymbol{\mu}_{t|t-1}^G(\mathbf{p}) &\triangleq \alpha_D(\mathbf{p})\boldsymbol{\ell} + \mathbf{c}_{1:t-1}^G(\mathbf{p})\boldsymbol{\Sigma}_{1:t-1}^{-1}(\mathbf{m}_{1:t-1} - \boldsymbol{\mu}_{1:t-1}) \in \mathbb{R} \text{ and} \\ \boldsymbol{\Sigma}_{t|t-1}^{F,G}(\mathbf{p}) &\triangleq \begin{bmatrix} \eta^2 + \sigma_\xi^2 & \eta^2 e^{-\frac{\|\mathbf{p}_S - \mathbf{p}_D\|_2}{\delta}} \\ \eta^2 e^{-\frac{\|\mathbf{p}_S - \mathbf{p}_D\|_2}{\delta}} & \eta^2 + \sigma_\xi^2 \end{bmatrix} \\ &\quad - \begin{bmatrix} \mathbf{c}_{1:t-1}^F(\mathbf{p}) \\ \mathbf{c}_{1:t-1}^G(\mathbf{p}) \end{bmatrix} \boldsymbol{\Sigma}_{1:t-1}^{-1} \begin{bmatrix} \mathbf{c}_{1:t-1}^F(\mathbf{p}) \\ \mathbf{c}_{1:t-1}^G(\mathbf{p}) \end{bmatrix}^T \in \mathbb{S}_{++}^2, \end{aligned}$$

with $\mathbf{m}_{1:t-1}$, $\boldsymbol{\mu}_{1:t-1}$, $\mathbf{c}_{1:t-1}^F(\mathbf{p})$, $\mathbf{c}_{1:t-1}^G(\mathbf{p})$, $\mathbf{c}_k^F(\mathbf{p})$, $\mathbf{c}_k^G(\mathbf{p})$ and $\boldsymbol{\Sigma}_{1:t-1}$ defined as in (6), (7), (8), (9), and (10) respectively, for all $(\mathbf{p}, t) \in \mathcal{S} \times \mathbb{N}_{N_T}^2$. Further, for every choice of $(m, n) \in \mathbb{Z} \times \mathbb{Z}$, the conditional correlation of the fields $|f(\mathbf{p}, t)|^m$ and $|g(\mathbf{p}, t)|^n$ relative to $\mathcal{C}(\mathcal{T}_{t-1})$ may be expressed in closed form as

$$\begin{aligned} & \mathbb{E}\{|f(\mathbf{p}, t)|^m |g(\mathbf{p}, t)|^n \mid \mathcal{C}(\mathcal{T}_{t-1})\} \\ & \equiv 10^{(m+n)p/20} \exp\left(\frac{\log(10)}{20} \left[\begin{array}{c} m \\ n \end{array} \right]^T \boldsymbol{\mu}_{t|t-1}^{F,G}(\mathbf{p}) + \left(\frac{\log(10)}{20}\right)^2 \left[\begin{array}{c} m \\ n \end{array} \right]^T \boldsymbol{\Sigma}_{t|t-1}^{F,G}(\mathbf{p}) \left[\begin{array}{c} m \\ n \end{array} \right]\right), \end{aligned}$$

at any $\mathbf{p} \in \mathcal{S}$ and for all $t \in \mathbb{N}_{N_T}^2$.

The detailed description of the proposed technique for efficiently approximating our base problem (11) now follows.

Sample Average Approximation (SAA): This is a direct Monte Carlo approach, where, at worst, existence of a sampling, or pseudosampling mechanism at each relay is assumed, capable of generating samples from a bivariate Gaussian measure. We may then observe that the objective of **Eq. 11** can be represented, for all $t \in \mathbb{N}_{N_T}^2$, via a Lebesgue integral as

$$\mathbb{E}\{V_I(\mathbf{p}, t) \mid \mathcal{C}(\mathcal{T}_{t-1})\} = \int_{\mathbb{R}^2} r(\mathbf{x}) \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{t|t-1}^{F,G}(\mathbf{p}), \boldsymbol{\Sigma}_{t|t-1}^{F,G}(\mathbf{p})) d\mathbf{x},$$

for any choice of $\mathbf{p} \in \mathcal{S}$, where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}): \mathbb{R}^2 \rightarrow \mathbb{R}_{++}$ denotes the bivariate Gaussian density, with mean $\boldsymbol{\mu} \in \mathbb{R}^{2 \times 1}$ and covariance $\boldsymbol{\Sigma} \in \mathbb{S}_{++}^{2 \times 2}$, and the function $r: \mathbb{R}^2 \rightarrow \mathbb{R}_{++}$ is defined as

$$r(\mathbf{x}) \triangleq \frac{P_c P_0 10^{\rho/10} [\exp(x_1 + x_2)]^\zeta}{P_0 \sigma_D^2 [\exp(x_1)]^\zeta + P_c \sigma^2 [\exp(x_2)]^\zeta + 10^{-\frac{\rho}{10}} \sigma^2 \sigma_D^2},$$

for all $\mathbf{x} \equiv (x_1, x_2) \in \mathbb{R}^2$, where $\zeta \triangleq \log(10)/10$. By a simple change of variables, it is also true that

$$\mathbb{E}\{V_I(\mathbf{p}, t) \mid \mathcal{C}(\mathcal{T}_{t-1})\} = \int_{\mathbb{R}^2} r\left(\sqrt{\boldsymbol{\Sigma}_{t|t-1}^{F,G}(\mathbf{p})} \mathbf{x} + \boldsymbol{\mu}_{t|t-1}^{F,G}(\mathbf{p})\right) \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{I}_2) d\mathbf{x},$$

for all $\mathbf{p} \in \mathcal{S}$ and $t \in \mathbb{N}_{N_T}^2$.

Now, for each relay $i \in \mathbb{N}_R^+$, at each $t \in \mathbb{N}_{N_T-1}^+$ and for some $S \in \mathbb{N}^+$, let $\{\mathbf{x}_{i,t}^j\}_{j \in \mathbb{N}_S^+}$ be a sequence of independent random elements in \mathbb{R}^2 , such that $\mathbf{x}_{i,t}^j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$, for all $j \in \mathbb{N}_S^+$. We also assume that all such sequences are mutually independent of the channel fields F and G . Then, by defining the sample average estimate

$$\mathbf{S}_S(\mathbf{p}, t) \triangleq \frac{1}{S} \sum_{j \in \mathbb{N}_S^+} r\left(\sqrt{\boldsymbol{\Sigma}_{t|t-1}^{F,G}(\mathbf{p})} \mathbf{x}_{i,t}^j + \boldsymbol{\mu}_{t|t-1}^{F,G}(\mathbf{p})\right),$$

the SAA of our initial problem **Eq. 11** is formulated as

$$\begin{aligned} & \underset{\mathbf{p}}{\text{maximize}} \quad \mathbf{S}_S(\mathbf{p}, t) \\ & \text{subject to} \quad \mathbf{p} \in C_i(\mathbf{p}^o(t-1)), \end{aligned} \quad (12)$$

at relay $i \in \mathbb{N}_R^+$, solved at each $t-1 \in \mathbb{N}_{N_T-1}^+$. A detailed analysis of the SAA problem **Eq. 12** is out of the scope of our discussion herein. Still, it is worth mentioning that the feasible of set of **Eq. 12** is finite, and therefore its optimal solution possesses various strong asymptotic guarantees in terms of convergence to the optimal solution of the original problem, as $S \rightarrow \infty$. For further details, see (Shapiro et al. (2009), Chapter 5).

On the downside, computing the objective of the SAA problem **Eq. 12** assumes availability of Monte Carlo samples, which could be restrictive in certain scenarios. Nevertheless, assuming mutual independence of the sequences $\{\mathbf{x}_{i,t}^j\}_j$, for each i and each t is not required. In fact, one could generate one sequence for all relays, per time slot, or even better, one sequence for all relays, for all time slots altogether. Such sampling schemes are legitimate, for two reasons. First, all SAAs of the form **Eq. 12** are solved independently for each relay and at each time slot. Second, Monte Carlo sampling is by construction statistically independent from the spatiotemporal channel fields F and G . As a result, such sampling schemes relax (in fact, eliminate) the need for (pseudo)random sampling at each individual relay. This makes them particularly attractive for practical purposes.

We denote this approach as SAA for the rest of the paper. The control flow of the SAA is presented in Algorithm 1.

Algorithm 1. SAA

Algorithm 1 SAA

```

Initialize Memory Buffer (MB) with fixed capacity  $D$ 
Initialize MB with  $D$  experiences - tuples of  $\{f_i, g_i, \mathbf{p}_i\}_{i=1}^D$ 
Set  $N_{\text{episodes}}$ 
set  $T_{\text{episode}}$ 
for all episodes  $N_{\text{episodes}}$  do
  set  $t$ 
  for all time steps of an episode  $T_{\text{episode}}$  do
    for all relays do
      Compute candidate positions  $\mathcal{P}_{tr} = \{\mathbf{p}_{nei}(t)\}$  of relay  $r$  (resp. grid boundaries and priority)
      Compute  $F_D$  and  $G_D$  based on all data in MB
      Compute  $\mu_{t+1|t}^{F_D, G_D}$  based on all data in MB
      Compute  $\Sigma_{t+1|t}^{F_D, G_D}$  based on all data in MB
      Compute  $S_S(\mathbf{p}_{nei}(t), t)$  for all  $\mathbf{p}_{nei}(t) \in \mathcal{P}_t$ 
      Choose  $\mathbf{p}(t+1) = \text{argmax}_{\mathbf{p}_{nei}(t)} S_S(\mathbf{p}_{nei}(t), t)$ 
      Observe  $f(\mathbf{p}(t+1))$  and  $g(\mathbf{p}(t+1))$ 
      Insert  $f(\mathbf{p}(t+1))$  and  $g(\mathbf{p}(t+1))$  and  $\mathbf{p}(t+1)$  in MB
    end for
    Synchronize and beamform to the destination
    Update  $t$  to be  $t+1$ 
  end for
end for

```

4 DEEP REINFORCEMENT LEARNING FOR ADAPTIVE DISCRETE RELAY MOTION CONTROL

4.1 Dynamic Programming for Relay Motion Control

The previously mentioned approach tackles the problem of relay motion control from a myopic perspective in the sense that the stochastic program is formulated so as to select the relay positions for the subsequent time slot with the goal of maximizing the collective SINR at the destination only for that particular slot.

The employment of reinforcement learning for the problem of discrete relay motion control entails that we reformulate the problem as a dynamic program. In this set up we want, at time slot $t-1$, to derive a motion policy (a methodology for choosing the relays' displacement) so as to maximize the discounted sum of V_t s (in expectation) from the subsequent time step t to the infinite horizon.

To formally pose that program we need to introduce a Markov Decision Process (MDP). The MDP is a tuple defined as $\{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma\}$ (Sutton and Barto, 2018):

The formulation of the dynamic program is as follows:

If γ is a discount factor, we can formulate the infinite horizon relay control problem as:

$$\begin{aligned}
 & \underset{\mathbf{u}(t), t \geq 0}{\text{maximize}} \quad \mathbb{E} \left\{ \sum_{t=1}^{\infty} \gamma^{t-1} \sum_{r=1}^R V_t(\mathbf{p}(t), t) \right\} \\
 & \text{subject to} \quad \begin{bmatrix} \mathbf{C}(t) \\ \mathbf{p}(t) \end{bmatrix} = \begin{bmatrix} e^{-1/c_2} \mathbf{C}(t-1) + \mathbf{W}(t) \\ \mathbf{p}(t-1) + \mathbf{u}(t) \end{bmatrix}, \quad (13) \\
 & \quad \mathbf{u}(t) \in \mathcal{A} \text{ is a function of } \mathcal{C}(\mathcal{T}_{t-1})
 \end{aligned}$$

where $\mathbf{u}(t)$ is the control at time t (essentially determining the relay displacement), and the driving noise $\mathbf{W}(t)$ is distributed as $\mathcal{N}(0, (1 - e^{-2/c_2}) \Sigma_C)$ and $\mathbf{C}(0) \sim \mathcal{N}(0, \Sigma_C)$. Σ_C is the covariance matrix for all channels (source and destination) for all the cells in the grid. The said covariance matrix is explicitly defined in (Kalogerias and Petropulu, 2017) and admits a particular form

if the channels evolve according to the spatiotemporal Gaussian process defined in 2.2.

Now, either the above problem defines a MDP or POMDP is dependent on the history $\mathcal{C}(\mathcal{T}_t)$. In particular, if $\mathcal{C}(\mathcal{T}_t)$ is generated by the whole state vector at each time slot then it is easy to see that problem Eq. 13 is fully observable, since all CSI generated by the environment is available for the relays to exploit for deciding upon the subsequent displacement.

On the other hand, if $\mathcal{C}(\mathcal{T}_t)$ is generated by the relay decisions together with only their local observations by their trajectories, then problem Eq. 13 becomes partially observable. Specifically, partial observability may be thought of as a dynamic observation selection process, which only reveals CSI pertaining to the trajectory of each relay, keeping the rest of the CSI hidden from the decision making process.

4.2 Deep Q Learning for Discrete Relay Motion Control

The employment of deep Q learning for relay motion control expels the need for making particular assumption for the underlying correlation structure of the channels.

Taking into account the (12) one can infer that we can construct a single policy that is learned by the collective experience of all the agents/relays and it constitutes the single policy that the movement of all relays strictly adhere to. In that spirit, we instantiate one neural network to parameterize the state-action value function (Q) and it is being trained on the experiences of all the relay. The motion policy is ϵ -greedy with respect to the estimation of the Q function.

Initially, we adopt the deep Q learning algorithm as described in (Mnih et al., 2015) and illustrated in Figure 2. Even though, as we pointed out in the previous subsection, the state of the MDP is the concatenation of the relay position $\mathbf{p} = \mathbf{s}$ and the channels $f(\mathbf{p}, t)$ and $g(\mathbf{p}, t)$, we follow a slightly different approach in the adoption of deep Q learning. In particular, the input to the neural network is the concatenation of the position $\mathbf{p} = [x, y]$ and the time step t . We should note at this point that augmenting the neural network input with the timestamp of the transition is a differentiation between the algorithm presented in this current work and the solution proposed in Evmorfos et al. (2022). This alternative, even though does not affect the implementation much, provides measurable improvements in cases where the power of the shadowing is strong. The reward r is the contribution of the relay to the SINR at the destination during the respective time step (V_t). At each time slot the relay selects an action $a \in \mathcal{A}_{\text{full}}$.

In general, Q learning with rich function approximators such as neural networks requires some heuristics for stability. The first such heuristic is the *Experience Replay* (Mnih et al., 2015). Each tuple of experience for a relay, namely $\{\text{state}, \text{action}, \text{next state}, \text{reward}\} \equiv \mathbf{s}, a, \mathbf{s}', r$, is stored in a memory. This memory we denote as Experience Replay. For the neural network updates, we sample uniformly a batch of experiences from the Experience Replay and use that batch to perform gradient descent to estimate the Q function (and subsequently the decision-making policy).

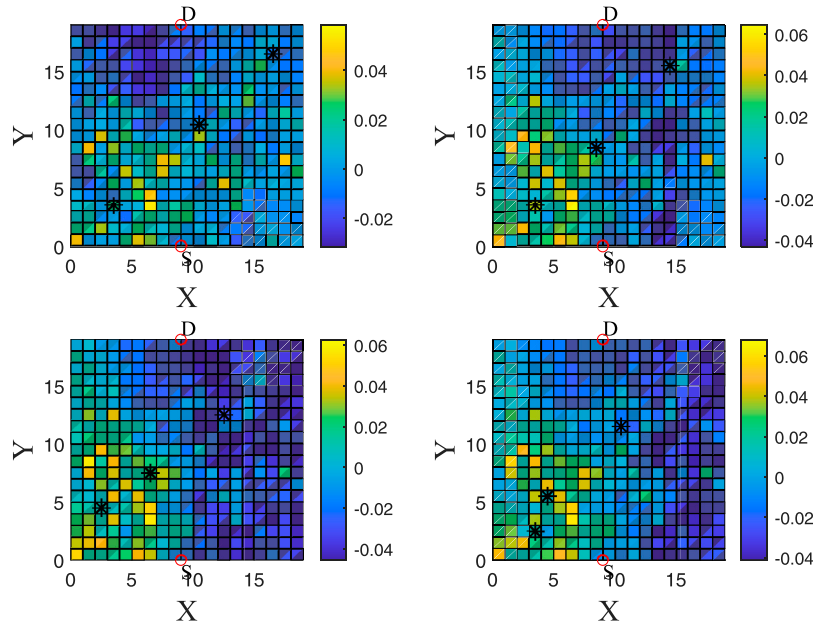


FIGURE 3 | This is a heatmap for visualizing a trajectory of the relays. We can see the V_t for all grid cells for four different time steps (each time step has a 2-time-slot difference with the previous and the next). One can see the positions of the relays for every time slot. The relays are moving towards better and better positions (larger V_t s).

The second heuristic is the *Target Network* (Mnih et al., 2015). The Target Network ($Q_{target}(s', a'; \theta^-)$) provides the estimation for the targets (labels) for the updates of the *Policy Network* ($Q_{policy}(s', a'; \theta^+)$), i.e., the network used for estimating the Q function. The two networks share (typically) the same architecture. We do not update the Target Network's weights with any optimization scheme, but, after a predefined number of training steps, the weights of the Policy Network are copied to the Target Network. This provides stationary targets for the weight updates and brings the task of the Q function approximation closer to a supervised learning paradigm.

Therefore, at each update step we sample a batch of experiences from the Experience Replay and use the batch to perform gradient descent on the loss:

$$\mathcal{L} = \left(Q_{policy}(s, a; \theta^+) - \left(r + \gamma \max_{a'} Q_{target}(s', a'; \theta^-) \right) \right)^2.$$

At each step, the Policy Network's weights are updated according to:

$$\theta_{t+1}^+ = \theta_t^+ + \lambda (Y_t - Q_{policy}(s, a; \theta_t^+)) \nabla_{\theta_t^+} Q(s, a; \theta_t^+).$$

where,

$$Y_t = r + \gamma \max_{a'} Q_{target}(s', a'; \theta_t^-)$$

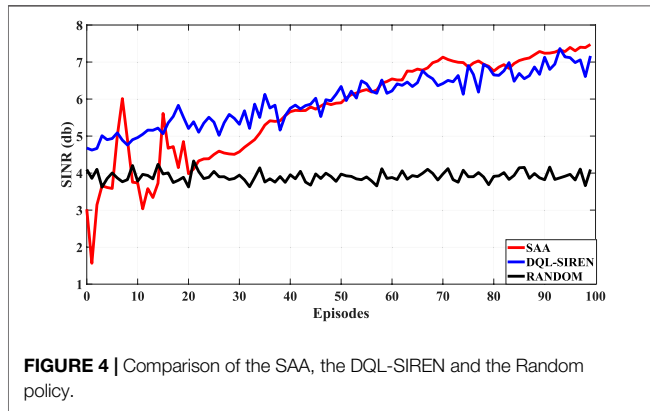
The parameter λ is the learning rate. The parameter γ is a scalar called the *discount factor* and *gamma* (0, 1). The choice for the discount factor pertains to a trade off between the importance assigned to long term rewards and the importance assigned to short term rewards. The parameters $a, a' \in \mathcal{A}_{full}$ correspond to the action chosen during the current state and the action chosen

for the next state (the state during the next time slot). Also, s and s' correspond to the current state and the next state respectively. The general pipeline of the deep Q learning algorithm is defined in **Figure 3**.

When the relays move (they do not stay in the same grid cell for two consecutive slots), they require additional energy consumption. In some cases though, the displacement to a neighboring grid cell does not correspond to significant improvement in terms of the cumulative SINR at the destination. Therefore, to account for the energy used for the application, we choose to not perform the ϵ -greedy policy directly on the estimates $Q_{policy}(s, a; \theta^+)$ of the Q function, but we decrease the estimates for all actions a , except for the action *stay*, by a small percentage μ . In that way we prohibit the relay displacement if this action does not correspond to a significant increase in the expectation of the cumulative sum of rewards (SINR). How significant this displacement action should be for it to be performed pertains to the choice of μ . For our simulations, in the subsequent sections, we choose μ to be 1%.

4.3 Sinusoidal Representation Networks for Q Function Parameterization

There have been many recent works which convincingly claim that coordinate-based Multilayer Perceptron Neural Networks (MLPs), i.e., MLPs that map a vector of coordinates to a low-dimensional natural signal, fail to learn high frequency components of the said signal. This constitutes a phenomenon that is called the spectral bias in machine learning literature (Jacot et al., 2018; Cao et al., 2019). The work in (Sitzmann et al., 2020)



examines the amelioration of spectral bias for MLPs. The inadequacy of MLPs for such inductive biases is bypassed by introducing a variation of the conventional MLP architecture with sinusoid ($\sin(\cdot)$) as activation function between layers. This MLP alternative was termed *Sinusoidal Representation Networks* (SIRENs), and was shown, both theoretically and experimentally, to effectively tackle the spectral bias.

The sinusoid is a periodic function which is quite atypical as a choice for activation function in neural networks. The authors in (Sitzmann et al., 2020) propose the employment of weight initialization framework so that the distribution of activations is retained during training and convergence is achieved without the network oscillating.

In particular, if we assume an intermediate layer of the neural network with input $\mathbf{x} \in \mathbb{R}^n$, then the output is an affine transformation using the weights \mathbf{w} passed through the sinusoid activation, therefore the output is $\sin(\mathbf{w}^T \mathbf{x} + b)$. Since the layer is not the first layer of the network, the input \mathbf{x} is arcsine distributed. With these assumptions it was shown in (Sitzmann et al., 2020) that, if the elements of \mathbf{w} , namely w_i , are initialized from a uniform distribution $w_i \sim \mathcal{U}(-\sqrt{\frac{6}{n}}, \sqrt{\frac{6}{n}})$, then $\mathbf{w}^T \mathbf{x} \sim \mathcal{N}(0, 1)$ as n grows. Therefore one should initialize the weights of all intermediate layers with $w_i \sim \mathcal{U}(-\sqrt{\frac{6}{n}}, \sqrt{\frac{6}{n}})$. The neurons of the first layer are initialized with the use of a scalar hyperparameter ω_0 , so that the output of the first layer, $\sin(\omega_0 \mathbf{W} \mathbf{x} + b)$ spans multiple periods over $[-1, 1]$. \mathbf{W} is a matrix whose elements correspond to the weights of the first layer.

When we adopt the deep Q learning approach for discrete relay motion control, we basically train a neural network (MLP) to learn a low-dimensional natural signal from coordinates, namely the state-action value function $Q(s, a)$. The Q function, $Q(s, a)$, represents the sum of SINR at the destination that the relays are expected to achieve for an infinite time horizon, starting from the respective position s and performing action a . The Policy Network, being a coordinate MLP may not be able to converge for the high frequency components of the underlying Q function that arise from the fact that the channels exhibit very abrupt spatiotemporal variations.

Therefore we propose that both the Policy and the Target Networks are SIRENs. The control flow of the algorithm we

propose is given in Algorithm 2. We denote this as DQL-SIREN, which stands for *Deep Q Learning with Sinusoidal Representation Networks*.

Algorithm 2. DQL-SIREN

Algorithm 2 DQL-SIREN

```

Initialize Experience Replay (ER)
Initialize  $\theta^-$  and  $\theta^+$ 
set update frequency
for all episodes do
  for all relays do
    input  $s = [x, y, t]$  to  $Q_{policy}$ 
    get  $Q_{policy}(s, a; \theta^+)$   $\forall a$ 
    subtract  $\mu = 1\%$  from  $Q_{policy}(s, a; \theta^+)$   $\forall a \neq \text{stay}$ 
     $\epsilon$ -greedy choice of  $a$ ,
      respecting grid boundaries and priority
    observe next state  $s'$  and reward  $r$ 
    store  $\{s, a, s', r\}$  to ER
     $s = s'$ 
  end for
  sample a batch of tuples  $\{s, a, s', r\}$  from ER
  for all tuples in the batch do
    input  $s$  to  $Q_{policy}$ , get  $Q_p = Q_{policy}(s, a; \theta^+)$ 
    input  $s'$  to  $Q_{target}$ , get  $Q_t = Q_{target}(s', a; \theta^-)$ 
     $\mathcal{L} = (Q_p - (r + \gamma \max_{a'} Q_t))^2$ 
    update  $\theta^+$  with gradient descent on  $\mathcal{L}$ 
    if steps % update frequency == 0 then
      copy the weights:  $\theta^+ \rightarrow \theta^-$ 
    end if
  end for
end for

```

5 SIMULATIONS

We test our proposed schemes by simulating a 20, \times , 20 m grid. All the grid cells are $1m \times 1m$. The number of agents/relays that assist the single source destination communication pair is $R = 3$. For every time slot the position of each relay is constrained within the boundaries of the gridded region and also constrained to adhere to a predetermined relay movement priority. Only one relay can occupy a grid cell per time slot. The center of the relay/agent and the center of the respective grid cell coincide.

When it comes to the shadowing part of our assumed channel model, we define a threshold θ which quantifies the distance in time and space where the shadowing component is important and can be taken into account for the construction of the motion policy. We assume that the shadowing power $\eta^2 = 15$ and the autocorrelation distance is $c_1 = 10m$ and the autocorrelation time is $c_2 = 20sec$. The variances of noises at the relays and destination are fixed as $\sigma^2 \equiv \sigma_D^2 \equiv 1$. The source and destination are fixed at $\mathbf{p}_S \equiv [10 \ 0]^T$ and $\mathbf{p}_D \equiv [10 \ 20]^T$.

Each one of the relays can move 1 grid cell/time slot and the size of each cell is $1m \times 1m$ (as mentioned before). The time slot length is set to be 0.6sec. Therefore the calculation of the channel and the decision of the movement for each relay should take up an amount of time that is strictly less than the duration of the time interval.

5.1 Specifications for the DQL-SIREN and the SAA

Regarding the DQL-SIREN, we employ SIRENs for both the Policy and the Target Networks. Each SIREN is comprised by three dense layers (350 neurons for each layer) and the learning rate is $1e - 4$.

TABLE 1 | Table of comparison between the two methods regarding key features.

Features	SAA	DQL-SIREN
Channel statistics	known (model-based)	unknown (model-free)
Robustness w.r.t seeds	extremely robust	slight variation bt seeds
Memory size	150 transitions	3,000 transitions
Horizon	myopic policies	long horizon policies (for γ close to 1)
Exploration	not required	required (ϵ -greedy)
Best SINR achieved	7.4 db	7.2 db

The Experience Replay size is 3,000 tuples and we begin every experiment with 300 transitions derived by a completely random policy before the start of training for all the deep Q learning approaches. The ϵ of the ϵ -greedy policy is initialized to be 1 but it is steadily decreased until it gets to 0.1 This is a very typical regime in RL. It is a very simple way to handle the dilemma between exploration and exploitation in RL, where we begin by giving emphasis to exploration first and then gradually exploration is traded for exploitation. We copy the weights of the Policy Network to the weights of the Target Network every 100 steps of training. The batch size is chosen to be 128 (even though the methods work reliably for different batch sizes ranging from 64 to 512) and the discount factor γ is chosen to be 0.99. We want to mention that small values for γ translate to a more myopic agent (an agent that assigns significance to short term rewards at the expense of long term/delayed rewards). On the other hand, values of γ closer to 1 correspond to agents that assign almost equal value to long term rewards and short term rewards. For the deep Q learning methods that we have proposed, we noticed that for low values of γ convergence and performance is impeded, something that we attribute to the interplay of Q learning and neural network employment rather than to the nature of the underlying MDP.

We set the ω_0 for the DQL-SIREN to 5 (the performance of the algorithm is robust for different values of the said parameter). Finally, we use the Adam optimizer for updating the network weights.

When it comes to the SAA, the sample size is set to 150 for the experiments.

5.2 Synthesized Data and Simulations

We create synthetic CSI data that adhere to the channel statistics described in 2.2.

In **Figure 4**, we plot the average SINR at the destination (in dB scale) achieved by the cooperation of all three relays, per episode, for 100 episodes, where every episode is comprised by 30 steps. The transmission power of the source is $P_S = 57\text{dbm}$ and the relay transmission power budget is $P_R = 57\text{dBm}$. The assumed channel parameters are set as $\ell = 2.3$, $\rho = 3$, $\eta^2 = 15$, $\sigma_\xi^2 = 3$, $c_1 = 10$, $c_2 = 20$, $c_3 = 0.5$. The variance of the noise at the relays and destination are $\sigma_D^2 = \sigma^2 = 0.5$.

We generate $3,000 = 100, \times, 30$ instances of the source-relay and relay destination channels for the whole grid $(20, \times, 20)$. Every 30 time steps we initialize the relays to random positions in the grid and let them move. We plot the average SINR for every 30 steps of the algorithms.

5.3 Simulation Results and Discussion

We present the results of our simulations in **Figure 4**. As we stated before, the results correspond to the average SINR at the destination for 100 episodes. Each episode consists of 30 time steps. The runs correspond to the average over six different seeds.

We compare three different policies. The first one is the Random policy, where each relay chooses the displacement for the next step at random. The second policy is the DQL-SIREN that solves the dynamic program (maximization of the discounted sum of V_I s for every relay from the current time step to the infinite horizon). The third policy is the myopic SAA that corresponds to the stochastic program and optimizes each individual relay's V_I for the subsequent slot.

As one can see that both the SAA and the DQL-SIREN perform significantly better than the Random policy (they both achieve an average SINR of approximately 7 db in contrast to the Random policy that achieves about 4 db). **Table 1** contains a head-to-head comparison of the SAA and the DQL-SIREN approaches regarding some qualitative and some quantitative features.

The convergence of the DQL-SIREN is faster than that of SAA. This is reasonable since, when it comes to the SAA approach, for the first five episodes there have not been collected enough samples (150). Both SAA and DQL-SIREN perform approximately the same in terms of average SINR. Towards the end of the experiments there is a small gap between the two (with the SAA performing slightly better). This can be attributed to the ϵ -greedy policy of the DQL-SIREN, where ϵ never goes to zero (choosing a random action a small percentage of the time for maintaining exploration).

There are some interesting inferences that one can make, based on the simulations. First of all, even though the SAA is myopic and only attempts to maximize the SINR for the subsequent time slot, works quite well in the sense of the aggregated statistic of the average SINR. This is a clear indication that, for the formulated problem, being greedy translates to performing adequately in the sense of cumulative reward.

Of course this peculiarity stands true only when the statistics of the channels are completely known and do not change significantly during the operation time. Apparently, in such a scenario, the phenomenon of delayed rewards is not much prevalent.

6 CONCLUSION

In this paper, we examine the discrete motion control for mobile relays facilitating the communication between a source and a destination. We compare two different approaches to tackle the problem. The first approach employs stochastic programming for scheduling the relay motion. This approach is myopic meaning that it seeks to maximize the SINR at the destination, only at the subsequent time slot. In addition, the stochastic programming approach makes specific assumption for the statistics of the channel evolution. The second approach is a deep reinforcement learning approach that is not myopic meaning that its goal is to maximize the discounted sum of SINR at the destination from the subsequent slot to an infinite time horizon. Additionally, the second approach makes no particular assumptions for the channel statistics. We test our methods in synthetic channel data produced in accordance to a known model for spatiotemporally varying channels. Both methods perform similarly and achieve significant improvement in comparison to a standard random policy for relay motion. We also provide a head-to-head comparison of the two approaches regarding various key qualitative and quantitative

features. As future work, we plan on extending the current methods for scenarios with multiple source-destination communication pairs and, possibly, include the existence of eavesdroppers.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

Work supported by ARO under grant W911NF2110071.

REFERENCES

- Astrom, K. J. (1970). *Introduction to Stochastic Control Theory*, 70. New York: Academic Press.
- Barriac, G., Mudumbai, R., and Madhow, U. (2004). "Distributed Beamforming for Information Transfer in Sensor Networks," in Third International Symposium on Information Processing in Sensor Networks, 2004 (IEEE), 81–88. doi:10.1145/984622.984635
- Bertsekas, D. (1995). *Dynamic Programming & Optimal Control*. 4th edn., II. Belmont, Massachusetts: Athena Scientific.
- Bertsekas, D. P., and Shreve, S. E. (1978). *Stochastic Optimal Control: The Discrete Time Case*, 23. New York: Academic Press.
- Cao, Y., Fang, Z., Wu, Y., Zhou, D.-X., and Gu, Q. (2019). Towards Understanding the Spectral Bias of Deep Learning. *arXiv preprint arXiv:1912.01198*.
- Chatzipanagiotis, N., Liu, Y., Petropulu, A., and Zavlanos, M. M. (2014). Distributed Cooperative Beamforming in Multi-Source Multi-Destination Clustered Systems. *IEEE Trans. Signal Process.* 62, 6105–6117. doi:10.1109/tsp.2014.2359634
- Evmorfos, S., Diamantaras, K., and Petropulu, A. (2021a). "Deep Q Learning with Fourier Feature Mapping for Mobile Relay Beamforming Networks," in 2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 126–130. doi:10.1109/SPAWC51858.2021.9593138
- Evmorfos, S., Diamantaras, K., and Petropulu, A. (2021b). "Double Deep Q Learning with Gradient Biasing for Mobile Relay Beamforming Networks," in 2021 55th Asilomar Conference on Signals, Systems, and Computers, 742–746. doi:10.1109/ieeeconf53345.2021.9723405
- Evmorfos, S., Diamantaras, K., and Petropulu, A. (2022). Reinforcement Learning for Motion Policies in Mobile Relaying Networks. *IEEE Trans. Signal Process.* 70, 850–861. doi:10.1109/TSP.2022.3141305
- Gao, F., Cui, T., and Nallanathan, A. (2008). On Channel Estimation and Optimal Training Design for Amplify and Forward Relay Networks. *IEEE Trans. Wirel. Commun.* 7, 1907–1916. doi:10.1109/TWC.2008.070118
- Goldsmith, A. (2005). *Wireless Communications*. Cambridge University Press.
- Havary-Nassab, V., Shahbazpanahi, S., Grami, A., and Zhi-Quan Luo, Z.-Q. (2008a). Distributed Beamforming for Relay Networks Based on Second-Order Statistics of the Channel State Information. *IEEE Trans. Signal Process.* 56, 4306–4316. doi:10.1109/tsp.2008.925945
- Havary-Nassab, V., Shahbazpanahi, S., Grami, A., and Zhi-Quan Luo, Z.-Q. (2008b). Distributed Beamforming for Relay Networks Based on Second-Order Statistics of the Channel State Information. *IEEE Trans. Signal Process.* 56, 4306–4316. doi:10.1109/TSP.2008.925945
- Heath, R. W. (2017). *Introduction to Wireless Digital Communication: A Signal Processing Perspective*. Prentice-Hall.
- Jacot, A., Gabriel, F., and Hongler, C. (2018). "Neural Tangent Kernel: Convergence and Generalization in Neural Networks," in Advances in Neural Information Processing Systems. Editors S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc).31.
- Kalogerias, D. S., Chatzipanagiotis, N., Zavlanos, M. M., and Petropulu, A. P. (2013). "Mobile Jammers for Secrecy Rate Maximization in Cooperative Networks," in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference, 2901–2905. doi:10.1109/ICASSP.2013.6638188
- Kalogerias, D. S., and Petropulu, A. P. (2016). "Mobile Beamforming Amp; Spatially Controlled Relay Communications," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6405–6409. doi:10.1109/ICASSP.2016
- Kalogerias, D. S., and Petropulu, A. P. (2017). Spatially Controlled Relay Beamforming: 2-stage Optimal Policies. *Arxiv*.
- Kalogerias, D. S., and Petropulu, A. P. (2018). Spatially Controlled Relay Beamforming. *IEEE Trans. Signal Process.* 66, 6418–6433. doi:10.1109/tsp.2018.2875896
- Li, J., Petropulu, A. P., and Poor, H. V. (2011). Cooperative Transmission for Relay Networks Based on Second-Order Statistics of Channel State Information. *IEEE Trans. Signal Process.* 59, 1280–1291. doi:10.1109/TSP.2010.2094614
- Liu, Y., and Petropulu, A. P. (2011). On the Sumrate of Amplify-And-Forward Relay Networks with Multiple Source-Destination Pairs. *IEEE Trans. Wirel. Commun.* 10, 3732–3742. doi:10.1109/twc.2011.091411.101523
- MacCartney, G. R., Zhang, J., Nie, S., and Rappaport, T. S. (2013). Path Loss Models for 5G Millimeter Wave Propagation Channels in Urban Microcells. *Globecom*, 3948–3953. doi:10.1109/glocom.2013.6831690
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level Control through Deep Reinforcement Learning. *nature* 518, 529–533. doi:10.1038/nature14236
- Muralidharan, A., and Mostofi, Y. (2017). "First Passage Distance to Connectivity for Mobile Robots," in Proceedings of the American Control Conference (IEEE), 1517–1523. doi:10.23919/ACC.2017.7963168
- Proakis, J. G., and Salehi, M. (2008). *Digital Communications*. McGraw-Hill.
- Rockafellar, R. T., and Wets, R. J.-B. (2004). *Variational Analysis*, 317. Springer Science & Business Media.
- Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2009). *Lectures on Stochastic Programming*. 2nd edn. Society for Industrial and Applied Mathematics.

- Sitzmann, V., Martel, J., Bergman, A., Lindell, D., and Wetzstein, G. (2020). Implicit Neural Representations with Periodic Activation Functions. *Adv. Neural Inf. Process. Syst.* 33, 7462–7473.
- Speyer, J. L., and Chung, W. H. (2008). *Stochastic Processes, Estimation, and Control*. Siam.
- Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT press.
- Yan, Y., and Mostofi, Y. (2013). Co-optimization of Communication and Motion Planning of a Robotic Operation under Resource Constraints and in Fading Environments. *IEEE Trans. Wirel. Commun.* 12, 1562–1572. doi:10.1109/twc.2013.021213.120138
- Yan, Y., and Mostofi, Y. (2012). Robotic Router Formation in Realistic Communication Environments. *IEEE Trans. Robot.* 28, 810–827. doi:10.1109/TRO.2012.2188163
- Zheng, G., Wong, K.-K., Paulraj, A., and Ottersten, B. (2009). Collaborative-Relay Beamforming with Perfect CSI: Optimum and Distributed Implementation. *IEEE Signal Process. Lett.* 16, 257–260. doi:10.1109/LSP.2008.2010810

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Evmorfos, Kalogieras and Petropulu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Bayesian Nonparametric Learning and Knowledge Transfer for Object Tracking Under Unknown Time-Varying Conditions

Omar Alotaibi and Antonia Papandreou-Suppappola*

School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ, United States

OPEN ACCESS

Edited by:

Hagit Messer,
Tel Aviv University, Israel

Reviewed by:

Allan De Freitas,
University of Pretoria, South Africa
Le Yang,
University of Canterbury, New Zealand

*Correspondence:

Antonia Papandreou-Suppappola
papandreou@asu.edu

Specialty section:

This article was submitted to
Statistical Signal Processing,
a section of the journal
Frontiers in Signal Processing

Received: 03 February 2022

Accepted: 26 April 2022

Published: 06 July 2022

Citation:

Alotaibi O and
Papandreou-Suppappola A (2022)
Bayesian Nonparametric Learning and
Knowledge Transfer for Object
Tracking Under Unknown Time-
Varying Conditions.
Front. Sig. Proc. 2:868638.
doi: 10.3389/frsip.2022.868638

We consider the problem of a primary source tracking a moving object under time-varying and unknown noise conditions. We propose two methods that integrate sequential Bayesian filtering with transfer learning to improve tracking performance. Within the transfer learning framework, multiple sources are assumed to perform the same tracking task as the primary source but under different noise conditions. The first method uses Gaussian mixtures to model the measurement distribution, assuming that the measurement noise intensity at the learning sources is fixed and known a priori and the learning and primary sources are simultaneously tracking the same source. The second tracking method uses Dirichlet process mixtures to model noise parameters, assuming that the learning source measurement noise intensity is unknown. As we demonstrate, the use of Bayesian nonparametric learning does not require all sources to track the same object. The learned information can be stored and transferred to the primary source when needed. Using simulations for both high- and low-signal-to-noise ratio conditions, we demonstrate the improved primary tracking performance as the number of learning sources increases.

Keywords: Bayesian nonparametric methods, machine learning, transfer learning, Gaussian mixture model, Dirichlet process mixture model

1 INTRODUCTION

Most statistical signal processing algorithms for tracking moving objects rely on physics-based models of the motion dynamics and on functions that relate sensor observations to the unknown object parameters (Bar-Shalom and Fortmann, 1988; Arulampalam et al., 2002). Any uncertainty in the motion dynamics or the tracking environment is most often characterized using probabilistic models with fixed parameters. However, when the operational or environmental conditions change during tracking, it is difficult to timely update the model parameters to better fit the new conditions. Some of the algorithm assumptions may no longer hold during such changes, resulting in loss of tracking performance. For example, radar performance has been shown to decrease when processing echo returns from rain and fog conditions due to changes in signal-to-noise ratio (SNR) (Hawkins and La Plant, 1959). As a result, unexpected changes in weather conditions will affect the accuracy of estimating the position of a moving target. Such a degradation in performance could be avoided if new information becomes available to help adapt the tracking algorithm.

Recent advances in sensing technology and increases in data availability have mandated the use of statistical models driven by sensors and data and thus the integration of machine learning into signal

processing algorithms (Mitchell, 1997; Hastie et al., 2016; Qiu et al., 2016; Rojo-Álvarez et al., 2018; Little, 2019; Lang et al., 2020; Theodoridis, 2020). For example, Gaussian mixtures, have been extensively used for data clustering or density estimation (Fraleigh and Raftery, 2002; Baxter, 2011; Reynolds, 2015). Different machine learning methods have been used, for example, to overcome limitations due to various assumptions on the sensing environment and to solve complex inference problems. Transfer learning is a machine learning method used to transfer and apply knowledge that is learned from previous tasks to solve a current task (Pan and Yang, 2010; Torrey and Shavlik, 2010; Karbalayghareh et al., 2018; Kouw and Loog, 2019; Papež and Quinn, 2019). This method is particularly advantageous when the data provided for inference is not sufficient or is difficult to label (Jaini et al., 2017). Transfer learning has been integrated into various signal processing applications, including trajectory tracking and radioactive particle tracking (Pereida et al., 2018; Lindner et al., 2022). Whereas many machine learning methods are applicable to learning a set of parameters of parametric models, Bayesian nonparametric methods allow for probability models from infinite dimensional families. They provide the flexibility to learn from current (and adapt to new) measurements as well as to integrate prior knowledge within the problem formulation (Ferguson, 1973; Antoniak, 1974; Hjort et al., 2010; Orbanz and Teh, 2010; Müller and Mitra, 2013; Xuan et al., 2019). Bayesian nonparametric methods have been adopted in tracking applications to model uncertainty directly from sensor observations. Dirichlet process mixtures were used to learn unknown probability density functions (PDFs) of noisy measurements (Escobar and West, 1995; Caron et al., 2008; Rabaoui et al., 2012); hierarchical Dirichlet process priors were used to learn an unknown number of dynamic modes (Fox et al., 2011); Dirichlet process mixture models were used to cluster an unknown number of statistically dependent measurements by estimating their joint density (Moraffah et al., 2019); and the dependent Dirichlet process was applied to learn the time-varying number and label of objects, together with measurement-to-object associations (Moraffah and Papandreou-Suppappola, 2018).

In this article, we propose tracking methods that integrate learning methodologies with sequential Bayesian filtering to track an object moving under unknown and time-varying noise conditions. We consider a primary tracking source whose task is to estimate the unknown dynamic state of the object using measurements whose noise characteristics are unknown and time-varying. Within the transfer learning framework, the primary source acquires prior knowledge from multiple learning sources that perform a similar tracking task but under different conditions. The first approach considers learning sources that use measurements with fixed and known noise intensity values and that simultaneously track the same object as the primary source. The Gaussian mixtures are used to model the measurement likelihood distribution at each learning source, and the model parameters are transferred to the primary source as prior knowledge. At the primary source, the unknown measurement likelihood distribution is estimated at each time step by modeling the transferred information as a finite mixture whose weights are learned using conjugate priors (Alotaibi and Papandreou-Suppappola, 2020). The method is also integrated with track-before-detect filtering for

tracking in high noise conditions. As the many assumptions made by this method can limit its applicability, we consider a second approach for tracking in more realistic and complex scenarios. This method considers learning sources with unknown noise intensity and exploits Bayesian nonparametric learning by modeling noise parameters using Dirichlet process mixtures. The mixture parameters are learned using conjugate priors, whose hyperparameters are modeled to provide estimates of the unknown noise intensity. The learned models are stored and made available to the primary source when needed (Alotaibi and Papandreou-Suppappola, 2021). Both proposed methods are extended to perform under high noise conditions by integrating track-before-detect filtering with transfer learning.

2 MATERIALS AND METHODS

2.1 Overview of Learning Methods

2.1.1 Transfer Learning

Transfer learning (TL) differs from other machine learning methods in that the data involved can originate from different tasks or have different domains. It aims to improve the performance of a primary source task by utilizing information learned from multiple learning sources that may perform the same or similar tasks but under different conditions (Arnold et al., 2007; Pan and Yang, 2010; Torrey and Shavlik, 2010; Weiss et al., 2016; Karbalayghareh et al., 2018; Kouw and Loog, 2019; Papež and Quinn, 2019). This is specifically important when sufficient data is not available at the primary source or when labeling the data is problematic. The inductive TL method assumes that the primary and secondary learning sources perform different but related tasks under the same conditions. On the other hand, the transductive TL method assumes that the same task is performed by both the primary source and the learning sources but under different conditions (Arnold et al., 2007; Pan and Yang, 2010). In particular, the learning sources use labeled data in order to adapt and learn a predictive distribution that can then be used by the primary source to learn the same predictive distribution but with unlabeled data. It is also important to determine which of the learned information to transfer to the primary source to optimize performance.

2.1.2 Gaussian Mixture Modeling

The unknown probability density function (PDF) of a noisy measurement vector \mathbf{z}_k at time step k is often estimated using the Gaussian mixture model (GMM). This is a probabilistic model that assumes all measurements originate from a mixture of M Gaussian components, and the m th component PDF $\text{ND}(\mathbf{z}_k; \boldsymbol{\mu}_{m,k}, C_{m,k})$ is characterized by the mean vector $\boldsymbol{\mu}_{m,k}$ and the covariance matrix $C_{m,k}$, $m = 1, \dots, M$. The model is given by¹ (Fraleigh and Raftery, 2002; Reynolds, 2015):

¹Throughout the paper, we use boldface lower case letters for row vectors, upper case letters for matrices, and boldface upper case Greek letters for sets. **Supplementary Appendix A** defines all acronyms and mathematical symbols used in the paper.

$$p(\mathbf{z}_k | \phi_k) = \sum_{m=1}^M b_{m,k} \text{ND}(\mathbf{z}_k; \boldsymbol{\mu}_{m,k}, C_{m,k}). \quad (1)$$

where $\phi_k = [\Phi_{1,k} \dots \Phi_{M,k}]$ is the GMM parameter vector and $\Phi_{m,k} = \{b_{m,k}, \boldsymbol{\mu}_{m,k}, C_{m,k}\}$. The GMM parameters are learned using the Dirichlet distribution conjugate prior for the weight $b_{m,k}$ and the normal inverse Wishart distribution (NIWD) conjugate prior for $\boldsymbol{\mu}_{m,k}, C_{m,k}$.

2.1.3 Dirichlet Process Mixture Modeling

A commonly used Bayesian nonparametric model for random probability measures in an infinite dimensional space is the Dirichlet process (DP) (Ferguson, 1973; Sethuraman, 1994). The DP G defines a prior in the space of probability distributions and is distributed according to $\text{DP}(\alpha, G_0)$, where $\alpha > 0$ is the concentration parameter and G_0 is the base distribution. The DP G is discrete, consisting of a countably infinite number of independent and identically distributed parameter sets Θ_k randomly drawn from the continuous G_0 (Sethuraman, 1994). The DP can be used to estimate the PDF of measurement \mathbf{z}_k , with statistically exchangeable samples, as follows:

$$p(\mathbf{z}_k) = \int p(\mathbf{z}_k | \Theta_{1:k}) dG(\Theta_{1:k}). \quad (2)$$

It can also be used for clustering using mixture models. Specifically, \mathbf{z}_k forms a cluster if $p(\mathbf{z}_k | \Theta_k)$ is parameterized by the same parameter set Θ_k drawn from $\text{DP}(\alpha, G_0)$. The DP mixture (DPM) model is a mixture model with a countably infinite number of clusters. Given DP parameter sets $\Theta_{1:k-1}$, the predictive distribution of Θ_k , drawn from the DP for clustering, is given by the Pólya urn representation

$$p(\Theta_k | \Theta_{1:k-1}, \alpha, G_0, \Psi) = \frac{\alpha}{k-1 + \alpha} G_0(\Theta_k; \Psi) + \frac{1}{k-1 + \alpha} \sum_{i=1}^{k-1} \delta(\Theta_k - \Theta_i). \quad (3)$$

For a multivariate normal G_0 , $\Theta_k = \{\boldsymbol{\mu}_k, C_k\}$ consists of the Gaussian mean $\boldsymbol{\mu}_k$ and covariance C_k . The NIWD conjugate prior with hyperparameter $\Psi = \{\boldsymbol{\mu}_0, \kappa, \Sigma, \nu\}$ is used to model the distribution of Θ_k .

2.2 Formulation of Object Tracking

2.2.1 Dynamic State Space Representation

We consider tracking a moving object with an unknown state parameter \mathbf{x}_k using measurement \mathbf{z}_k at each time step $k, k = 1, \dots, K$. The dynamic system is described by the state-space representation.

$$\mathbf{x}_k = g(\mathbf{x}_{k-1}) + \mathbf{v}_{k-1} \Rightarrow p(\mathbf{x}_k | \mathbf{x}_{k-1}), \quad (4)$$

$$\mathbf{z}_k = h(\mathbf{x}_k) + \mathbf{w}_k \Rightarrow p(\mathbf{z}_k | \mathbf{x}_k), \quad (5)$$

where \mathbf{w}_k is the measurement noise vector and \mathbf{v}_k is a random vector that accounts for modeling errors. The function $g(\mathbf{x}_k)$ models the transition of the unknown state parameters between time steps, and $h(\mathbf{x}_k)$ provides the relationship between the

measurement and the unknown state. The unknown state is obtained by estimating the state posterior PDF $p(\mathbf{x}_k | \mathbf{z}_k)$ (Kalman, 1960; Bar-Shalom and Fortmann, 1988). This can be achieved using recursive Bayesian filtering that involves two steps. The prediction step obtains an estimate of the posterior PDF using the transition PDF $p(\mathbf{x}_k | \mathbf{x}_{k-1})$ in Eq. 4 and the posterior PDF $p(\mathbf{x}_{k-1} | \mathbf{z}_{k-1})$ from the previous time step. The update step amends the predicted estimate using the measurement likelihood $p(\mathbf{z}_k | \mathbf{x}_k)$ in Eq. 5. Assuming that the probabilistic models for \mathbf{v}_k in Eq. 4 and \mathbf{w}_k in Eq. 5 are known, the posterior PDF can be estimated recursively. Such methods include the Kalman filter (KF), which assumes linear system functions and Gaussian processes, and sequential Monte Carlo methods such as particle filtering (Doucet et al., 2001; Arulampalam et al., 2002).

2.2.2 Tracking With Transfer Learning

We integrate transductive TL in our tracking formulation (see Section 2.1.1), where a primary source and L learning sources perform the same task of tracking a moving object. For ease of notation, the primary source object state and measurement vectors are denoted by \mathbf{x}_k and \mathbf{z}_k , as in Eqs. 4 and 5, respectively; the corresponding ones for the ℓ th learning source, $\ell = 1, \dots, L$, are denoted by $\mathbf{x}_{\ell,k}$ and $\mathbf{z}_{\ell,k}$. The primary source is tracking under time-varying conditions, resulting in measurements with an unknown noise intensity $\xi_k \in \Xi_p$ at time step k in Eq. 5. Note that $\Xi_p \in \mathbb{R}^+$ is a set of discrete levels of noise intensity values. The primary tracking is expected to benefit from knowledge transferred from the L learning sources, provided that the ℓ th source measurement noise intensity $\xi_k^{(\ell,L)}$, $\ell = 1, \dots, L$, takes values from the set $\Xi \in \mathbb{R}^+$ that has common values with Ξ_p . This prior knowledge is in the form of learned probabilistic models of the measurement noise distribution from each learning source. At the primary source, the transferred models are integrated into a finite mixture whose weights are learned using Dirichlet priors.

2.2.3 Tracking Under Low Signal-To-Noise Ratio Conditions

The measurements in Eq. 5 provided for tracking differ depending on the SNR. For high SNRs, the object is assumed present at all times and the measurements correspond to estimated information from generalized matched filtering. However, when the SNR is low, unthresholded measurements are processed by integrating the track-before-detect (TBD) approach with Bayesian sequential methods (Tonissen and Bar-Shalom, 1988; Salmond and Birch, 2001; Boers and Driessen, 2004; Ebenezer and Papandreou-Suppappola, 2016). TBD incorporates a binary object existence indicator λ_k and models the object existence as a two-state Markov chain. The new formulation depends on the probability $P_d = \Pr(\lambda_k = 0 | \lambda_{k-1} = 1)$, which is the probability that the object is not detected at time step k given that it was detected at time $k - 1$. The transition PDF is given by

$$p(\mathbf{x}_k, \lambda_k | \mathbf{x}_{k-1}, \lambda_{k-1}) = \begin{cases} p(\mathbf{x}_k | \mathbf{x}_{k-1}) (1 - P_d), & \lambda_k = \lambda_{k-1} = 1 \\ p_b(\mathbf{x}_k) P_b, & \lambda_k = 1, \lambda_{k-1} = 0 \end{cases} \quad (6)$$

where $P_b = \Pr(\lambda_k = 1 \mid \lambda_{k-1} = 0)$ and $p_b(\mathbf{x}_k)$ is the initial PDF of the object state when detected. The measurement likelihood is given by

$$p(\mathbf{z}_k \mid \mathbf{x}_k, \lambda_k) = \begin{cases} p(\mathbf{z}_k \mid \mathbf{x}_k), & \lambda_k = 1 \\ p(\mathbf{w}_k), & \lambda_k = 0. \end{cases} \quad (7)$$

2.3 Tracking With Transfer Learning and Gaussian Mixture Model Modeling

Following the tracking formulation in Section 2.2 within the TL framework (see Section 2.1.1), we propose an approach to track a moving object under time-varying measurement noise conditions as our primary source task. It is assumed that L other sources are simultaneously tracking the same object but using measurements obtained from different sensors. The approach models the measurement likelihood PDF of each learning source using Gaussian mixtures and transfers the learned model parameters to the primary source to improve its tracking performance. The TL-GMM tracking method, summarized in Algorithm 1, discussed next, for high and low SNR conditions.

Algorithm 1. TL-GMM Recursive Tracking Algorithm

L learning sources with input measurements $\mathbf{z}_{\ell,k}$, $\ell = 1, \dots, L$

for $\ell = 1 : L$ **do**

- Sample state $\mathbf{x}_{\ell,k}$ using $p(\mathbf{x}_{\ell,k} \mid \mathbf{x}_{\ell,k-1})$ in (4)
- for** $m = 1 : M$ **do**
 - Draw NIWD hyperparameter set $\Upsilon_{m,\ell,k}$ and Dirichlet distribution hyperparameter $\gamma_{m,\ell,k}$
 - Sample Gaussian mixture component $\mu_{m,\ell,k}, C_{m,\ell,k} \sim \text{NIWD}(\mu_{m,\ell,k}, C_{m,\ell,k} \mid \Upsilon_{m,\ell,k})$
 - Draw mixture weight $b_{m,\ell,k}$ from Dirichlet distribution prior $\text{Dir}(b_{m,\ell,k} \mid \gamma_{m,\ell,k})$
 - Model measurement $\mathbf{z}_{\ell,k} \mid \mathbf{x}_{\ell,k} \sim \text{ND}(\mathbf{z}_{\ell,k} \mid \mu_{m,\ell,k}, C_{m,\ell,k})$ in (8)
 - Form parameter set $\Phi_{m,\ell,k} = \{b_{m,\ell,k}, \mu_{m,\ell,k}, C_{m,\ell,k}\}$
- end for**
- Form GMM parameter vector $\phi_{\ell,k} = [\Phi_{1,\ell,k} \dots \Phi_{M,\ell,k}]$ from prior $p(\phi_{\ell,k})$ in (9)
- Obtain measurement likelihood $p(\mathbf{z}_{\ell,k} \mid \mathbf{x}_{\ell,k}, \phi_{\ell,k})$ using (8)
- Compute posterior PDF $p(\mathbf{x}_{\ell,k}, \phi_{\ell,k} \mid \mathbf{z}_{\ell,k})$ using (10)
- Return learned model parameter vector $\phi_{\ell,k}$ for the ℓ th learning source

end for

Primary source with input measurement \mathbf{z}_k and transferred parameters $\phi_{\ell,k}$, $\ell = 1, \dots, L$

- Sample state $\mathbf{x}_k \sim p(\mathbf{x}_k \mid \mathbf{x}_{k-1})$ using (4)
- for** $\ell = 1 : L$ **do**
 - Model measurement $\mathbf{z}_k \mid \mathbf{x}_k \sim p(\mathbf{z}_k \mid \mathbf{x}_k, \phi_{\ell,k})$
 - Draw Dirichlet distribution hyperparameter $\tilde{\gamma}_{m,\ell,k}$
 - Draw mixture weight $d_{\ell,k}$ from Dirichlet distribution prior $\text{Dir}(d_{\ell,k} \mid \tilde{\gamma}_{\ell})$
- end for**
- Form measurement PDF $p(\mathbf{z}_k \mid \mathbf{x}_k, \mathbf{d}_k)$ in (11)
- Compute posterior PDF $p(\mathbf{x}_k, \mathbf{d}_k \mid \mathbf{z}_k)$ in (12)
- Return estimated state vector \mathbf{x}_k and weight basis vector \mathbf{d}_k

2.3.1 TL-GMM Tracking Method

2.3.1.1 Multiple Source Learning With TL-GMM

The task of the ℓ th learning source is to estimate the posterior PDF of the object state $\mathbf{x}_{\ell,k}$ at time step k , given measurement $\mathbf{z}_{\ell,k}$, following Eqs. 4 and 5. The measurement noise $\mathbf{w}_{\ell,k}$ is assumed to have a zero-mean Gaussian distribution with a known and constant intensity level $\xi^{(\ell,L)} \in \Xi_L$; though not necessary, we assume that each source has a unique intensity value. The state is recursively estimated using the posterior PDF. It is first predicted using the prior PDF $p(\mathbf{x}_{\ell,k} \mid \mathbf{x}_{\ell,k-1})$. Given the measurement $\mathbf{z}_{\ell,k}$, the measurement likelihood PDF is estimated using Gaussian mixtures, as in Eq. 1.

$$p(\mathbf{z}_{\ell,k} \mid \mathbf{x}_{\ell,k}, \phi_{\ell,k}) = \sum_{m=1}^M b_{m,\ell,k} \text{ND}(\mathbf{z}_{\ell,k} \mid \mu_{m,\ell,k}, C_{m,\ell,k}). \quad (8)$$

The m th component has a mean $\mu_{m,\ell,k}$ and a covariance matrix $C_{m,\ell,k}$ and is weighted by the mixing parameter $b_{m,\ell,k}$, $m = 1, \dots, M$. As the measurement noise intensity $\xi^{(\ell,L)}$ is assumed to be known at the learning sources, the noise covariance can be used to initialize each GMM component with an equal probability $b_{m,\ell,1} = 1/M$. The GMM parameter vector $\phi_{\ell,k} = [\Phi_{1,\ell,k} \dots \Phi_{M,\ell,k}]$, with $\Phi_{m,\ell,k} = \{b_{m,\ell,k}, \mu_{m,\ell,k}, C_{m,\ell,k}\}$, is learned using conjugate priors. The weight $b_{m,\ell,k}$ uses the Dirichlet distribution (Dir) prior with hyperparameter $\gamma_{m,\ell,k}$, and the Gaussian mean $\mu_{m,\ell,k}$ and covariance $C_{m,\ell,k}$ use the NIWD prior with hyperparameter set $\Upsilon_{m,\ell,k}$. The resulting prior is

$$p(\phi_{\ell,k}) = \text{Dir}(\mathbf{b}_{\ell,k} \mid \gamma_{\ell,k}) \prod_{m=1}^M \text{NIWD}(\mu_{m,\ell,k}, C_{m,\ell,k} \mid \Upsilon_{m,\ell,k}), \quad (9)$$

where $\mathbf{b}_{\ell,k} = [b_{1,\ell,k} \dots b_{M,\ell,k}]$ and $\gamma_{\ell,k} = [\gamma_{1,\ell,k} \dots \gamma_{M,\ell,k}]$, and the posterior PDF is

$$p(\mathbf{x}_{\ell,k}, \phi_{\ell,k} \mid \mathbf{z}_{\ell,k}) \propto p(\mathbf{z}_{\ell,k} \mid \mathbf{x}_{\ell,k}, \phi_{\ell,k}) p(\mathbf{x}_{\ell,k} \mid \mathbf{x}_{\ell,k-1}) p(\phi_{\ell,k}) p(\mathbf{x}_{\ell,k-1}, \phi_{\ell,k-1} \mid \mathbf{z}_{\ell,k-1}). \quad (10)$$

The derivation steps are provided in **Supplementary Appendix B**.

2.3.1.2 Primary Source Tracking With TL-GMM

From the TL formulation in Section 2.2, the primary source measurement noise \mathbf{w}_k in Eq. 5 is assumed to have a zero-mean Gaussian with a covariance matrix $C_k = \xi_k C$, with noise intensity $\xi_k \in \Xi_p$. At each time step k , the primary source receives the modeled prior hyperparameter sets $\phi_{\ell,k}$, $\ell = 1, \dots, L$, in Eq. 9, from each of the L learning sources and uses them to model the primary measurement likelihood PDF as

$$p(\mathbf{z}_k \mid \mathbf{x}_k, \mathbf{d}_k) = \sum_{\ell=1}^L d_{\ell,k} p(\mathbf{z}_k \mid \mathbf{x}_k, \phi_{\ell,k}) \\ = \sum_{\ell=1}^L d_{\ell,k} \sum_{m=1}^M b_{m,\ell,k} \text{ND}(\mathbf{z}_k \mid \mu_{m,\ell,k}, C_{m,\ell,k}), \quad (11)$$

where $\mathbf{d}_k = [d_{1,k} \dots d_{L,k}]$. As the PDF in Eq. 11 is a collection of PDFs and mixing weights (Lindsay, 1995; Baxter, 2011), it can be viewed as a finite mixture model. The weight $d_{\ell,k}$ is learned using a Dirichlet distribution conjugate prior with the hyperparameter $\tilde{\gamma}_{\ell}$. This learning step allows for the best matched learning sources to be exploited at different time steps. The posterior PDF is thus given by

$$p(\mathbf{x}_k, \mathbf{d}_k \mid \mathbf{z}_k) \propto p(\mathbf{z}_k \mid \mathbf{x}_k, \mathbf{d}_k) p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) p(\mathbf{d}_k) p(\mathbf{x}_{k-1}, \mathbf{d}_{k-1} \mid \mathbf{z}_{k-1}), \quad (12)$$

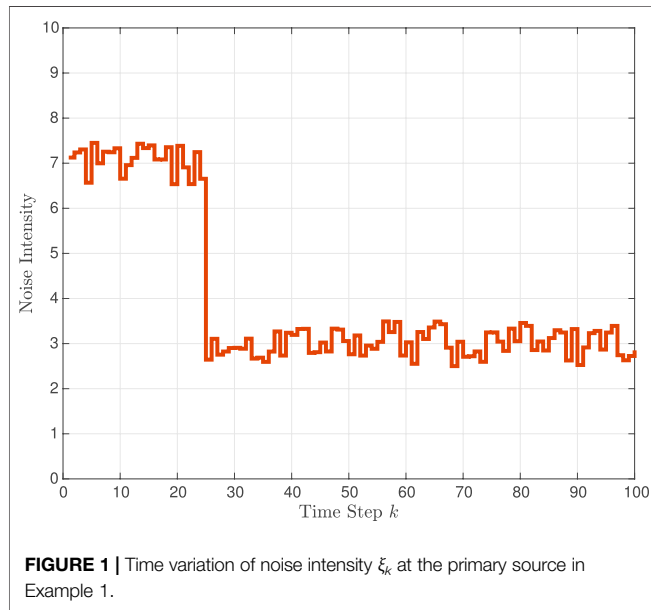
where $p(\mathbf{x}_k \mid \mathbf{x}_{k-1})$ is given in Eq. 4 and $p(\mathbf{x}_{k-1}, \mathbf{d}_{k-1} \mid \mathbf{z}_{k-1})$ is the posterior from the previous time step.

2.3.2 TL-GMM Tracking With Track-Before-Detect

When tracking under low SNR conditions, the measurement likelihood PDF in Eq. 7 for the L learning sources depends on the binary object existence indicator $\lambda_{\ell,k}$. Following the GMM model in Eq. 8 for the TBD formulation, the measurement likelihood for the ℓ th learning source, $\ell = 1, \dots, L$, is

TABLE 1 | Noise intensity values from set Ξ_L for L learning sources in Examples 1–5, where Ξ_p is the set of the primary source noise intensity values.

Learning source noise intensity $\xi^{(\ell,L)}$, $\ell = 1, \dots, L$											
L	Ξ_L	$\xi^{(1,L)}$	$\xi^{(2,L)}$	$\xi^{(3,L)}$	$\xi^{(4,L)}$	$\xi^{(5,L)}$	$\xi^{(6,L)}$	$\xi^{(7,L)}$	$\xi^{(8,L)}$	$\xi^{(9,L)}$	$\xi^{(10,L)}$
1	{4, 5}	4.4									
2	{5, 9}	8.2	5.8								
4	{2, 10}	1.5	6.3	4.2	9.4						
10	{1, 10}	6.1	9.2	3.2	4.5	7.0	2.6	3.9	8.4	1.8	7.7
5	{1, 10}	2.1	7.4	3.2	9.0	1.5					
5	{5, 10}	5.5	7.8	6.0	9.4	9.0					
5	{4, 10}	8.1	4.6	9.7	7.7	5.9					
10	{1, 10}	2.8	5.7	8.8	5.0	7.1	9.4	6.5	8.9	5.3	3.4
10	{5, 10}	7.3	5.2	8.7	5.2	9.8	8.7	9.7	7.6	6.2	5.9
3	{12, 18}	12.1	17.1	13.8							
2	{6, 10}	6	10								
3	{1, 9}	6	9	1							
5	{2, 10}	2	4	6	8	10					
10	{1, 10}	4	3	6	8	2	5	1	7	10	9
5	{1, 7}	1	2	4	6	7					
5	{1, 10}	2	4	6	8	10					



$$p(\mathbf{z}_{\ell,k} | \mathbf{x}_{\ell,k}, \lambda_{\ell,k}, \phi_{\ell,k}) = \begin{cases} \sum_{m=1}^M b_{m,\ell,k} \text{ND}(\mathbf{z}_{\ell,k} | \boldsymbol{\mu}_{m,\ell,k}, C_{m,\ell,k}), & \lambda_{\ell,k} = 1 \\ \mathbf{w}_{\ell,k}, & \lambda_{\ell,k} = 0 \end{cases} \quad (13)$$

The GMM model in Eq. 13 is used to obtain the posterior PDF, following Eq. 10, as

$$p(\mathbf{x}_{\ell,k}, \lambda_{\ell,k}, \phi_{\ell,k} | \mathbf{z}_{\ell,k}) \propto p(\mathbf{z}_{\ell,k} | \mathbf{x}_{\ell,k}, \lambda_{\ell,k}, \phi_{\ell,k}) p(\mathbf{x}_{\ell,k}, \lambda_{\ell,k} | \mathbf{x}_{\ell,k-1}, \lambda_{\ell,k-1}) \cdot p(\phi_{\ell,k}) p(\mathbf{x}_{\ell,k-1}, \lambda_{\ell,k-1}, \phi_{\ell,k-1} | \mathbf{z}_{\ell,k-1}),$$

where $p(\mathbf{x}_{\ell,k}, \lambda_{\ell,k} | \mathbf{x}_{\ell,k-1}, \lambda_{\ell,k-1})$ is given in Eq. 6 and $p(\phi_{\ell,k})$ in Eq. 9. The PDF $p(\mathbf{x}_{\ell,k-1}, \lambda_{\ell,k-1}, \phi_{\ell,k-1} | \mathbf{z}_{\ell,k-1})$ is obtained from the previous time step with probability $(1 - P_d)$ when $\lambda_{\ell,k-1} = 1$

and is otherwise set to its initial value. When tracking at the primary source, following Eq. 11, the measurement PDF is

$$p(\mathbf{z}_k | \mathbf{x}_k, \lambda_k, \mathbf{d}_k) = \begin{cases} \sum_{\ell=1}^L d_{\ell,k} \lambda_{\ell,k} \sum_{m=1}^M b_{m,\ell,k} \text{ND}(\mathbf{z}_k | \boldsymbol{\mu}_{m,\ell,k}, C_{m,\ell,k}), & \lambda_k = 1 \\ \mathbf{w}_k, & \lambda_k = 0 \end{cases}$$

The posterior PDF is, thus, given by

$$p(\mathbf{x}_k, \lambda_k, \mathbf{d}_k | \mathbf{z}_k) \propto p(\mathbf{z}_k | \mathbf{x}_k, \lambda_k, \mathbf{d}_k) p(\mathbf{x}_k, \lambda_k | \mathbf{x}_{k-1}, \lambda_{k-1}) p(\mathbf{d}_k) p(\mathbf{x}_{k-1}, \lambda_{k-1}, \mathbf{d}_{k-1} | \mathbf{z}_{k-1}).$$

2.4 Tracking With Transfer Learning and Bayesian Nonparametric Modeling

The TL-GMM method not only assumes that the learning sources have known noise intensity, but it also requires both the primary and learning sources to be simultaneously tracking the same object. We instead consider the more realistic scenario, where each of the learning sources is tracking under unknown noise intensity conditions. Our proposed approach is based on integrating TL with Bayesian nonparametric (BNP) methods to allow for modeling of the multiple source measurements without the assumption of parametric models. The learned model parameters are stored and acquired as needed as prior knowledge for the primary tracking source to improve its performance when tracking under time-varying noise intensity conditions. The TL-BNP approach is discussed next and summarized in Algorithm 2.

2.4.1 Multiple Source Learning Using TL-BNP

Within the TL framework, the ℓ th learning source, $\ell = 1, \dots, L$, is tracking a moving object using measurements embedded in zero-mean Gaussian noise with unknown intensity $\xi^{(\ell,L)}$. Using the DPM model in Eq. 2 with base distribution G_ℓ for the ℓ th source, the DP model parameter set $\Theta_{\ell,k} = \{\boldsymbol{\mu}_{\ell,k}, C_{\ell,k}\}$ provides the mean $\boldsymbol{\mu}_{\ell,k}$ and covariance $C_{\ell,k}$ of the Gaussian mixed PDF $p(\mathbf{z}_{\ell,k} | \Theta_{\ell,k}, \Psi_\ell)$.

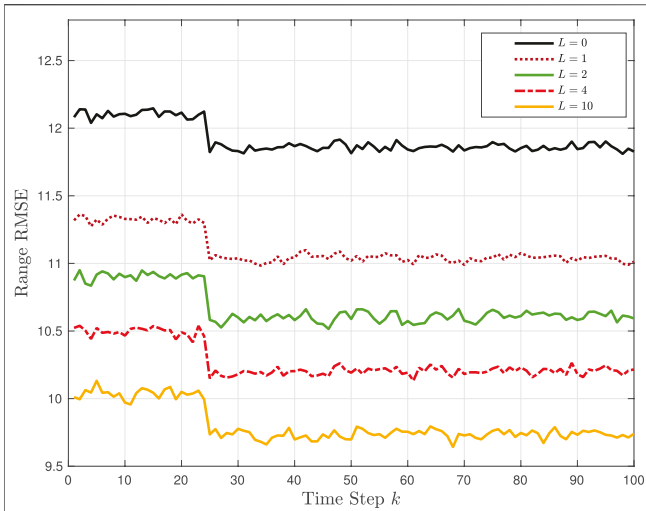


FIGURE 2 | TL-GMM tracking in Example 1: Range RMSE performance without transfer learning ($L = 0$) and with $L = 1, 2, 4, 10$ learning sources.

Parameter set $\Theta_{\ell,k}$ is learned using the NIWD conjugate prior with hyperparameter set $\Psi_\ell = \{\mu_{0,\ell}, \kappa_\ell, \Sigma_\ell, \nu_\ell\}$, which can be computed using Markov chain Monte Carlo methods such as Gibbs sampling (West, 1992; Neal, 2000; Rabaoui et al., 2012). In (Rabaoui et al., 2012), navigation performance under hard reception conditions was improved by estimating NIWD hyperparameters in an efficient Rao-Blackwellized particle filter (RBPf) implementation. In (Gómez-Villegas et al., 2014), the sensitivity to added perturbations on prior hyperparameters was demonstrated using the Kullback–Leibler divergence measure.

Algorithm 2. TL-BNP Recursive Tracking Algorithm

L learning sources, initialize $\mathbf{x}_{\ell,0}$, $\Theta_{\ell,0}$, $\ell = 1, \dots, L$
 At time step k , input measurement $\mathbf{z}_{\ell,k}$ and use $\mathbf{x}_{\ell,k-1}$ and $\Theta_{\ell,k-1}$ from time step $k-1$
for PF particles, $i = 1 : N_s$ **do**
 – Draw state particles $\mathbf{x}_{\ell,k}^{(i)}$ from $p(\mathbf{x}_{\ell,k}^{(i)} | \mathbf{x}_{\ell,k-1}^{(i)})$ in (4)
 – Draw NIWD hyperparameters $\Psi_\ell^{(i)}$ from $p(\Psi_\ell^{(i)} | \Theta_{\ell,k-1}^{(i)})$ in (14)
 – Draw DP parameter particles $\Theta_{\ell,k}^{(i)} = \{\mu_{\ell,k}^{(i)}, C_{\ell,k}^{(i)}\}$ from $p(\Theta_{\ell,k}^{(i)} | \Theta_{\ell,k-1}^{(i)}, \Psi_\ell^{(i)})$ in (3)
 – Obtain Gaussian likelihood $p(\mathbf{z}_{\ell,k} | \mathbf{x}_{\ell,k}^{(i)}, \Theta_{\ell,k}^{(i)}, \Psi_\ell^{(i)})$ in (17)
 – Compute weight $w_{\ell,k}^{(i)}$ using (16)
 – Normalize weights $\tilde{w}_{\ell,k}^{(i)} = w_{\ell,k}^{(i)} / \sum_{i=1}^{N_s} w_{\ell,k}^{(i)}$
 – Resample particles $\mathbf{x}_{\ell,k}^{(i)}$ and $\Theta_{\ell,k}^{(i)}$, $i = 1, \dots, N_s$
end for
 Compute posterior $p(\mathbf{x}_{\ell,k}, \Theta_{\ell,k}, \Psi_\ell | \mathbf{z}_{\ell,k})$ in (15)
 Return learned model $p(\Psi_\ell)$ for the ℓ th learning source
Primary source, transfer parameters Ψ_ℓ , $\ell = 1, \dots, L$, initialize \mathbf{x}_0
 At time step k , input measurement \mathbf{z}_k and use \mathbf{x}_{k-1} from time step $k-1$
 Sample state \mathbf{x}_k using $p(\mathbf{x}_k | \mathbf{x}_{k-1})$ in (4)
for $\ell = 1 : L$ **do**
 – Draw DP parameter $\Theta_{\ell,k}$ from $p(\Theta_{\ell,k} | \Psi_\ell)$ in (19)
 – Model measurement $p(\mathbf{z}_k | \Theta_{\ell,k}, \Psi_\ell)$
 – Draw mixture weight $d_{\ell,k}$ using $\text{Dir}(d_{\ell,k} | \hat{\gamma}_\ell)$
end for
 – Form mixture distribution $p(\Theta_k, \mathbf{d}_k | \mathbf{z}_k, \Psi)$ in (18)
 – Use PF to estimate $p(\mathbf{x}_k | \Theta_k, \mathbf{d}_k, \mathbf{z}_k, \Psi)$
 – Compute posterior PDF $p(\mathbf{x}_k, \Theta_k, \mathbf{d}_k | \mathbf{z}_k, \Psi)$ in (20)
 – Return estimated state vector \mathbf{x}_k and estimated parameter Θ_k

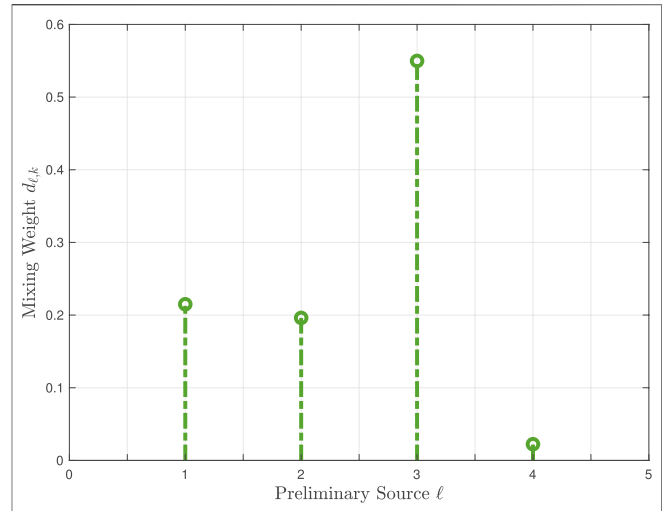


FIGURE 3 | Learned mixing weights $d_{\ell,k}$, for $k = 80$ and $\ell = 1, 2, 3, 4$ in Example 1.

Given measurement $\mathbf{z}_{\ell,k}$, the DP and NIWD model parameters are

$$p(\Theta_{\ell,k}, \Psi_\ell | \mathbf{z}_{\ell,k}) \propto p(\Theta_{\ell,k} | \Theta_{\ell,k-1}, \Psi_\ell) p(\Psi_\ell | \Theta_{\ell,k-1}). \quad (14)$$

The object tracking involves the estimation of the object state $\mathbf{x}_{\ell,k}$, DP model parameter set $\Theta_{\ell,k}$, and hyperparameter set Ψ_ℓ , given measurement $\mathbf{z}_{\ell,k}$. Their joint PDF $p(\mathbf{x}_{\ell,k}, \Theta_{\ell,k}, \Psi_\ell | \mathbf{z}_{\ell,k})$ is approximated using particle filtering (Arulampalam et al., 2002), as detailed in **Supplementary Appendix C**. At each time step k , N_s particles, $\mathbf{x}_{\ell,k-1}^{(i)}$ and $\Theta_{\ell,k-1}^{(i)}$,

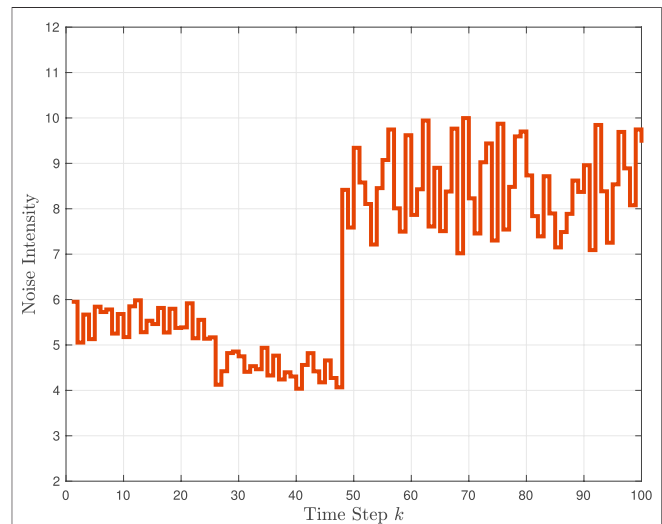


FIGURE 4 | Time variation of primary source noise intensity ξ_k in Example 2.

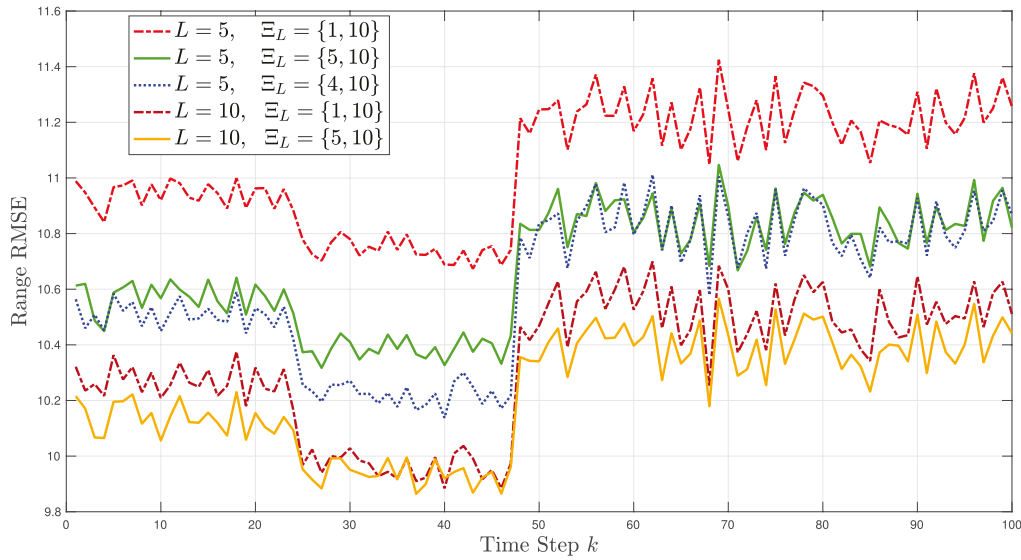


FIGURE 5 | TL-GMM tracking in Example 2: Range RMSE performance with $L = 5, 10$ learning sources with varying sets of noise intensity values Ξ_L .

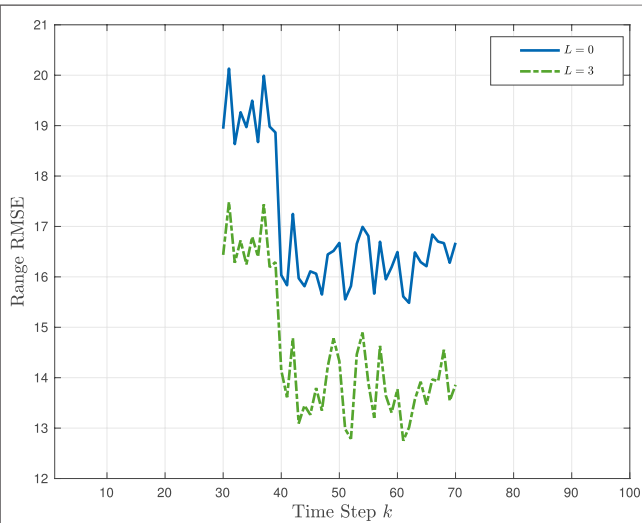


FIGURE 6 | TL-GMM tracking with TBD in Example 3: Range RMSE performance without transfer learning ($L = 0$) and with $L = 3$ learning sources.

$i = 1, \dots, N_s$, are sampled from a proposal distribution to obtain

$$p(\mathbf{x}_{\ell,k}, \Theta_{\ell,k}, \Psi_{\ell} | \mathbf{z}_{\ell,k}) \approx \sum_{i=1}^{N_s} w_{\ell,k}^{(i)} \delta(\mathbf{x}_{\ell,k} - \mathbf{x}_{\ell,k-1}^{(i)}) \delta(\Theta_{\ell,k} - \Theta_{\ell,k-1}^{(i)}). \quad (15)$$

The joint prior PDF $p(\mathbf{x}_{\ell,k}^{(i)} | \Theta_{\ell,k-1}^{(i)}, \Psi_{\ell}^{(i)}) = p(\mathbf{x}_{\ell,k}^{(i)} | \mathbf{x}_{\ell,k-1}^{(i)}) p(\Theta_{\ell,k}^{(i)} | \Theta_{\ell,k-1}^{(i)}, \Psi_{\ell}^{(i)})$ is selected as the proposal distribution, which assumes that the object state and model parameters are independent during prediction. Particles $\mathbf{x}_{\ell,k}^{(i)}$ are drawn from the state prior $p(\mathbf{x}_{\ell,k}^{(i)} | \mathbf{x}_{\ell,k-1}^{(i)})$ in Eq. 4. Particles $\Theta_{\ell,k}^{(i)}$ are independently drawn using the Pólya urn representation of the DP $p(\Theta_{\ell,k}^{(i)} | \Theta_{\ell,k-1}^{(i)}, \Psi_{\ell}^{(i)})$ in Eq. 3. Note that particles $\Psi_{\ell}^{(i)}$ are

drawn from $p(\Psi_{\ell}^{(i)} | \Theta_{\ell,k-1}^{(i)})$, that provides a probabilistic model for the hyperparameter set. The weights in Eq. 15 are updated using

$$w_{\ell,k}^{(i)} \propto w_{\ell,k-1}^{(i)} p(\mathbf{z}_{\ell,k} | \mathbf{x}_{\ell,k}^{(i)}, \Theta_{\ell,k}^{(i)}, \Psi_{\ell}^{(i)}), \quad (16)$$

where N is the number of $\mathbf{z}_{\ell,k}$ samples. The Gaussian likelihood is computed based on Eq. 5,

$$p(\mathbf{z}_k | \mathbf{x}_{\ell,k}^{(i)}, \Theta_{\ell,k}^{(i)}, \Psi_{\ell}^{(i)}) = \frac{1}{(2\pi)^{N/2} \sqrt{|C_{\ell,k}^{(i)}|}} \exp\left(-\frac{1}{2}(\mathbf{z}_{\ell,k} - h(\mathbf{x}_{\ell,k}^{(i)}))^T (C_{\ell,k}^{(i)})^{-1} (\mathbf{z}_{\ell,k} - h(\mathbf{x}_{\ell,k}^{(i)}))\right), \quad (17)$$

using covariance matrix $C_{\ell,k}^{(i)}$ from model parameter $\Theta_{\ell,k}^{(i)}$ in the Gaussian mixed PDF $p(\mathbf{z}_{\ell,k} | \Theta_{\ell,k}^{(i)}, \Psi_{\ell}^{(i)})$.

2.4.2 Primary Source Tracking With TL-BNP

The learned hyperparameter set $\Psi = \{\Psi_1, \dots, \Psi_L\}$ from the learning sources is stored and made available, when needed, to use as prior knowledge for the primary tracking task. Note that, unlike with the GMM-based transfer, the learning source tracking does not need to occur simultaneously as the primary tracking. Thus, at the primary source, Ψ is used to learn the unknown and time-varying measurement noise characteristics. Specifically,

$$p(\Theta_k, \mathbf{d}_k | \mathbf{z}_k, \Psi) = \sum_{\ell=1}^L d_{\ell,k} p(\Theta_{\ell,k} | \mathbf{z}_k, \Psi_{\ell}), \quad (18)$$

where weights $\mathbf{d}_k = [d_{1,k} \dots d_{L,k}]$ are learned with a Dirichlet distribution prior with hyperparameter $\tilde{\gamma}_{\ell}$, and $\Theta_{\ell,k}$ are sampled from the transferred learned parameters Ψ_{ℓ} . The PDF $p(\Theta_{\ell,k} | \mathbf{z}_k, \Psi_{\ell})$ is given by

$$p(\Theta_{\ell,k} | \mathbf{z}_k, \Psi_{\ell}) \propto p(\mathbf{z}_k | \Theta_{\ell,k}, \Psi_{\ell}) p(\Theta_{\ell,k} | \Psi_{\ell}). \quad (19)$$

The posterior PDF is given by

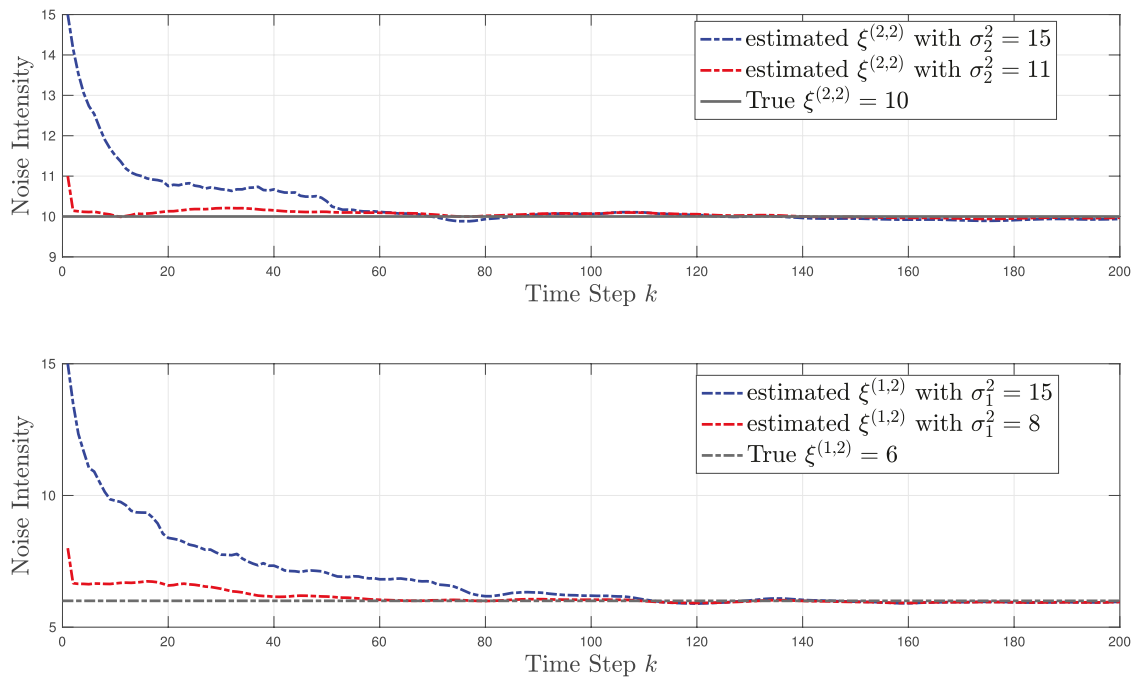


FIGURE 7 | TL-BNP tracking in Example 4: Modeling of unknown noise intensities $\xi^{(1,2)}$ and $\xi^{(2,2)}$ for $L = 2$ learning sources by varying the NIWD hyperparameter σ_L^2 .

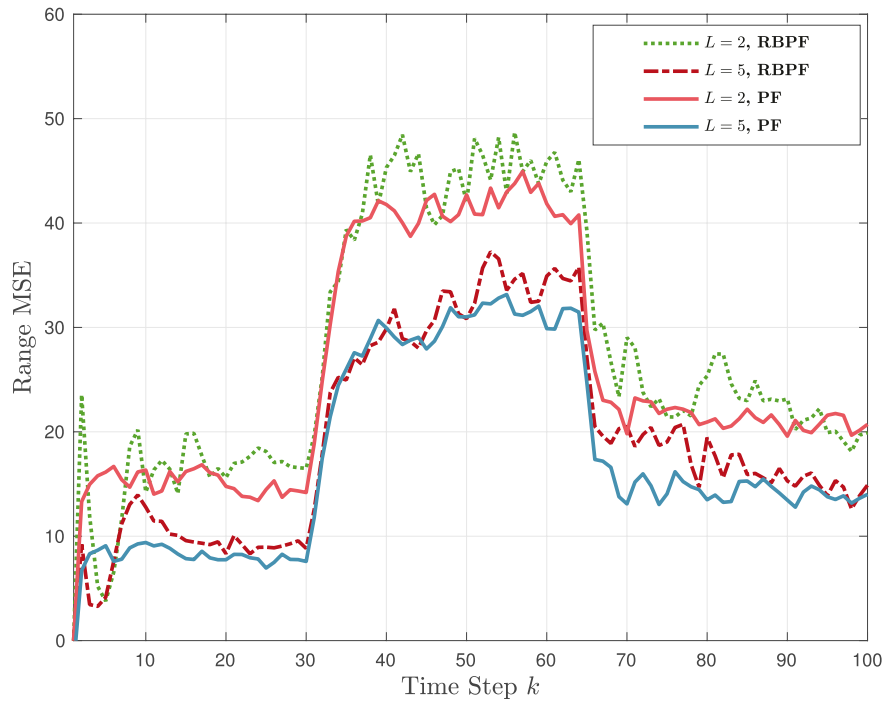
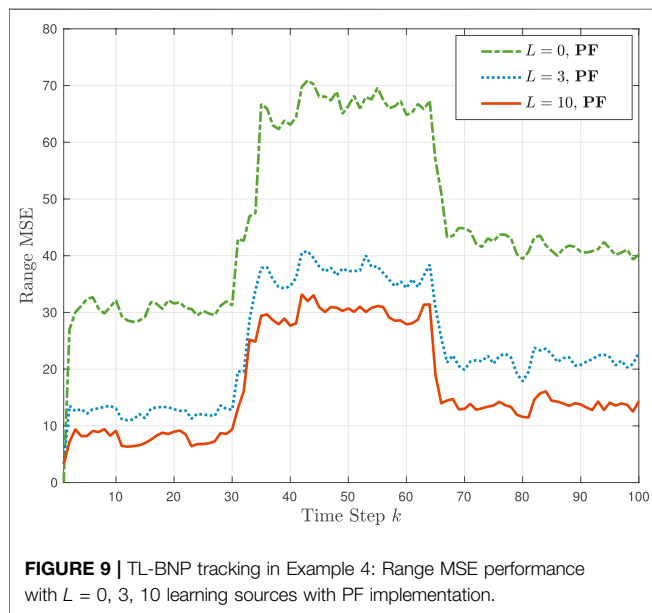


FIGURE 8 | TL-BNP tracking in Example 4: Range MSE performance with $L = 2, 5$ using two different implementations, PF and RBPF, at the learning sources.

$$p(\mathbf{x}_k, \mathbf{d}_k, \Theta_k | \mathbf{z}_k, \Psi) = p(\mathbf{x}_k | \Theta_k, \mathbf{d}_k, \mathbf{z}_k, \Psi) p(\Theta_k, \mathbf{d}_k | \mathbf{z}_k, \Psi), \quad (20)$$

with $p(\Theta_k, \mathbf{d}_k | \mathbf{z}_k, \Psi)$ in Eq. 19 and estimating $p(\mathbf{x}_k | \Theta_k, \mathbf{d}_k, \mathbf{z}_k, \Psi)$ with a PF.

Note that, similarly to the TL-GMM approach in Section 2.3.2, the TL-BPN can also be extended to incorporate the TBD framework for tracking under low SNR conditions.



3 RESULTS AND DISCUSSION

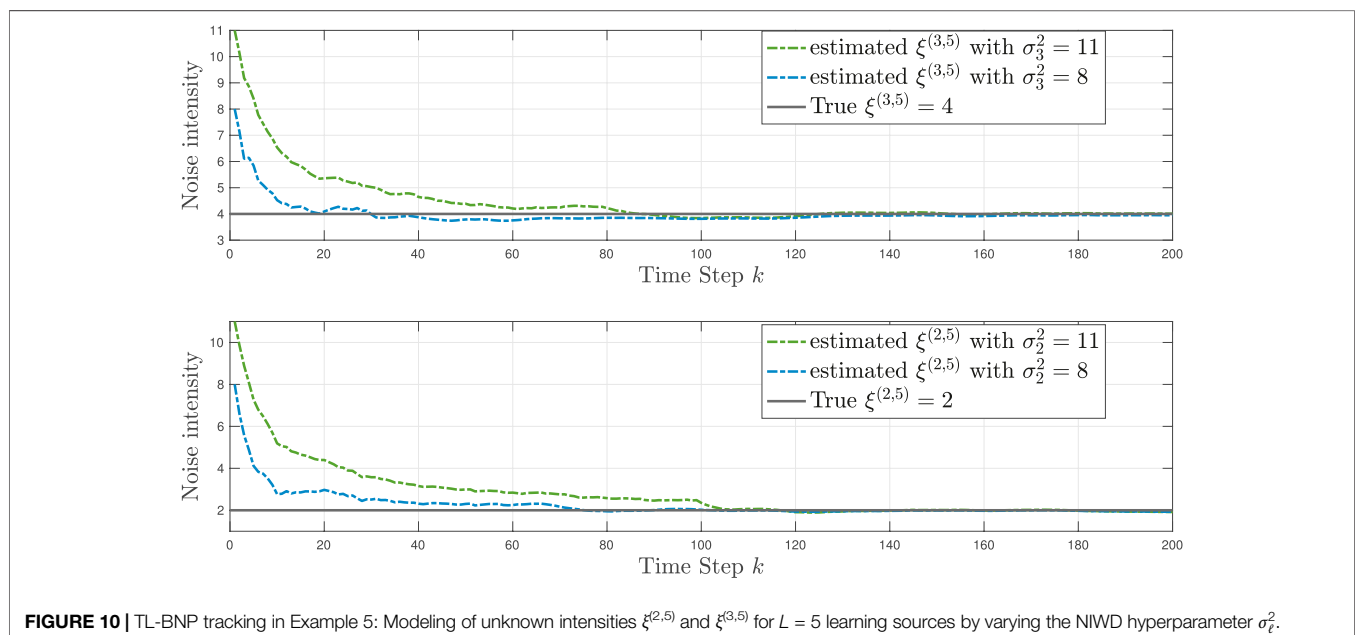
3.1 Simulation Settings

In this section, we simulate various scenarios of tracking a moving object under time-varying conditions to demonstrate and compare the performance of our two proposed methods. The methods are discussed in **Sections 2.3** and **2.4**, and we refer to them as the TL-GMM method (transfer learning and Gaussian mixture modeling) and the TL-BNP method (transfer learning and Bayesian nonparametric modeling), respectively. For both methods, the noise intensity ξ_k at the primary source is assumed to be unknown and time-varying. Note that our goal is not to

explicitly estimate the noise intensity ξ_k ; we model and learn the measurement noise intensity information in order to use it in estimating the unknown object state.

For all simulations, our goal is to estimate a moving object's two-dimensional (2-D) position that is denoted by the object state vector $\mathbf{x}_k = [x_k \ y_k]^T$, $k = 1, \dots, K$, where (x_k, y_k) are the Cartesian coordinates in meters. We assumed a simple first order Markov process for the state transition, $\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{v}_{k-1}$, and we selected a high variance of $\sigma_v^2 = 6$ for the zero-mean white Gaussian vector \mathbf{v}_k to emulate motion. The time between time steps is 1 s and the total number of time steps is $K = 100$. The sensor measurement vector \mathbf{z}_k at the primary source is assumed corrupted by additive zero-mean Gaussian noise with an unknown intensity ξ_k at time step k . For the ℓ th learning source, we generated a uniformly sampled intensity value $1 \leq \xi^{(\ell,L)} \leq 10$ for high SNR and $12 \leq \xi^{(\ell,L)} \leq 18$ for low SNR, $\ell = 1, \dots, L$. The measurement vector $\mathbf{z}_k = [r_k \ \zeta_k]$ consists of the object's range $r_k = \sqrt{x_k^2 + y_k^2}$ and bearing $\zeta_k = \arctan(y_k/x_k)$. For low SNR tracking using TBD filtering, the measurement vector \mathbf{z}_k in **Eq. 7** corresponds to unthresholded cross-ambiguity function measurements that are modeled as 2-D Gaussian resolution frames of range and bearing cells (Ebenezer and Papandreou-Suppappola, 2016). In **Eq. 6**, we set $P_d = P_b = 0.03$.

For the algorithm implementation, unless otherwise stated, we used 10,000 Monte Carlo runs. The sequential importance resampling PF was used for tracking in both approaches, with $N_s = 3,000$ particles. For GMM modeling, the number of Gaussian mixtures was set to $M = 10$ as we considered a maximum of $L = 10$ learning sources. Before receiving any measurements, the initial NIWD hyperparameter set for the GMM parameters was set to $\mathbf{Y}_{m,\ell,0} = \{[0, 0], 3, \text{diag}([1, 1], 3)\}$ in **Eq. 9**. For DPM modeling, we fixed the concentration parameter to $\alpha = 0.1$ the base distribution G_0 as Gaussian in



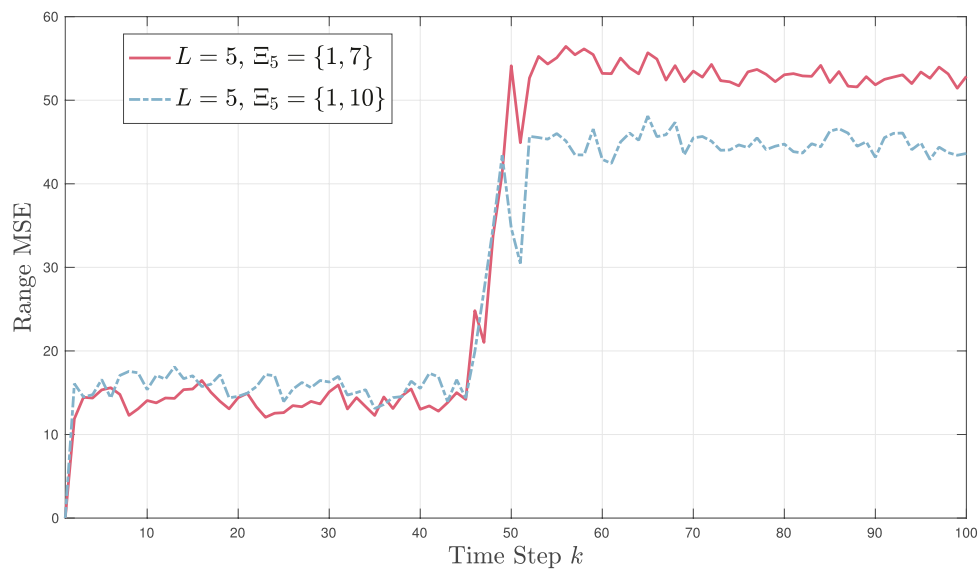


FIGURE 11 | TL-BNP tracking in Example 5: Range MSE performance with $L = 5$ learning sources with varying intensity values Ξ_5 .

Eq. 3. The initial NIWD hyperparameter set for DP was set to $\Psi_{\ell,0} = \{[0, 0], 3, \sigma_\ell^2 \mathcal{I}, 3\}$ where \mathcal{I} is the identity matrix. We used simulations to study the selection of the initial σ_ℓ^2 value, and we selected an exponential forgetting factor of 0.9 to ensure that the updated NIWD hyperparameters did not grow exponentially (Berntorp and Cairano, 2016). The noise intensity values used in different simulations, both for the primary source in set Ξ_p and the L learning sources in sets Ξ_L , are summarized in **Table 1**.

For tracking performance evaluation and comparison, we use the estimation mean-squared error (MSE) and root mean-squared error (RMSE) of the object's range. We use $L = 0$ to denote tracking without transfer learning. For this tracker, we generate the primary source noise intensity values from a uniform distribution, taking values from $\Xi_p = \{1, 10\}$ at each time step and Monte Carlo run.

3.2 Tracking With TL-GMM Approach

3.2.1 TL-GMM: Effect of Varying the Number of Learning Sources in Example 1

In the first simulation in Example 1, the primary tracking source noise intensity ξ_k varies within $\Xi_p = \{2, 8\}$, as shown in **Figure 1**. In particular, the intensity varies slowly from around $\xi_k \approx 7$ up to $k = 25$, before dropping to, and remaining at, around $\xi_k \approx 3$ for the remaining time steps. For performance comparison, we simulated a tracker that does not use transfer learning ($L = 0$) and four different trackers that use transfer learning using $L = 1, 2, 4, 10$ learning sources. The fixed known noise intensity value $\xi^{(\ell,L)}$ of the ℓ th learning source, for $\ell = 1, \dots, L$, is provided in **Table 1**. The RMSE of the estimated range is demonstrated as a function of the time step k in **Figure 2**. As expected, the tracking performance is worse when no prior information is transferred to the primary source. Also, the RMSE decreases as the number of learning resources L increases. For example, the RMSE performance is higher when $L = 2$ than when $L = 1$. Compared with the primary

source intensity values in **Figure 1** with the values used by the learning sources, although $\xi^{(1,1)} = 4.4$ for $L = 1$ and also $\xi^{(2,2)} = 5.8$ for $L = 2$ are not used by the primary source, the value $\xi^{(1,2)} = 8.2$ for $L = 2$ is close to the high values of ξ_k during the first 25 time steps. Note that, for all five trackers, the RMSE decreases when there is a large increase in the primary source SNR at $k = 25$. Also, as the SNR remains high after $k = 25$, the RMSE is lower during the last 75 time steps.

Figure 3 studies more closely the performance of the TL-GMM with $L = 4$ by providing the learned mixing weights $d_{\ell,k}$, for $k = 80$ and $\ell = 1, 2, 3, 4$. From **Figure 1**, the primary source intensity at $k = 80$ is 3.5, and the $L = 4$ learning source intensities, from **Table 1**, are $\xi^{(1,4)} = 1.5$, $\xi^{(2,4)} = 6.3$, $\xi^{(3,4)} = 4.2$, and $\xi^{(4,4)} = 9.4$. We then use $\Delta\xi^{(\ell)} = |\xi^{(\ell)} - 3.5|$, which is the absolute difference in intensity between the ℓ th learning source and the primary source at $k = 80$, to examine its relation to the ℓ th learned mixing weight $d_{\ell,80}$. We would expect that the learning source with the minimum $\Delta\xi^{(\ell)}$ is the best match to the primary source at $k = 80$ and thus have the mixing weight $d_{\ell,80}$. This is indeed the case, as shown in **Figure 3**: the largest weight is $d_{3,80}$ and $\Delta\xi^{(3)} = 0.7$ is the minimum difference. We also observe that $d_{4,80}$ is the smallest weight as $\Delta\xi^{(4)} = 5.9$ is the maximum difference, and $d_{1,80}$ and $d_{2,80}$ are about the same since $\Delta\xi^{(1)} = 2$ and $\Delta\xi^{(2)} = 2.8$ are close in value.

3.2.2. TL-GMM: Effect of Varying Learning Source Noise Intensity in Example 2

For this example, the primary source noise intensity ξ_k varies within $\Xi_p \in \{4, 10\}$ in **Figure 4**. Note that, as with Example 1, there is an abrupt change in intensity (at $k = 48$); however, before and after this change, the intensity undergoes higher variations than in the previous example in **Figure 1**. We consider five different cases using $L = 5, 10$ learning sources and vary the noise intensity for a fixed L . The learning source intensity values $\xi^{(\ell,L)}$, $\ell = 1, \dots,$

L , and corresponding Ξ_L set are provided in **Table 1**. The range of RMSE for the five cases is shown in **Figure 5**. We first note that the RMSE decreases when the number of learning sources increases from $L = 5$ to $L = 10$. Compared to the three cases with $L = 5$, the best performance is achieved when the values of noise intensity $\Xi_5 \in \{4, 10\}$ match those of the primary source $\Xi_p \in \{4, 10\}$. The longest interval, $\Xi_{10} \in \{1, 10\}$ results in the worst performance as the primary source does not have any values between 1 and 4. The overall best performance of the primary source is achieved using the highest L number, for which Ξ_L closely matches Ξ_p .

3.2.3. TL-GMM: Effect of Low Signal-To-Noise Ratio at the Primary Source in Example 3

In this example, we evaluate tracking under low SNR conditions for an object entering the scene at time step $k = 30$ and leaving the scene at time step $k = 70$. The primary source noise intensity ξ_k varies between the values of 12.5 and 16.5, with a sudden decrease at time step $k = 40$. We compare the performance of tracking without TL ($L = 0$) and with TL using $L = 3$ learning sources. The learning source noise intensities for $L = 3$ are provided in **Table 1**. The RMSE of the estimated range for both cases is shown in **Figure 6**. Note that the tracking performance improves with TL, as expected. Note that for both tracking methods, the RMSE is lower between time steps $k = 40$ and $k = 70$. This is because the SNR is higher during those time steps when compared to the first 10 time steps of the object entering the scene.

3.3 Tracking With the TL-BNP Approach

3.3.1 TL-BNP: Effect of Initial NIWD Hyperparameters on Noise Intensity Estimation in Example 4

When using the TL-BNP approach, we first demonstrate how the modeling of the initial NIWD prior hyperparameter σ_ℓ^2 in Ψ_ℓ affects the estimation of the noise intensity at the ℓ primary source. We consider $L = 2$ learning sources whose noise intensities are unknown. As shown in **Table 1**, their corresponding true intensity values are $\xi^{(1,2)} = 6$ and $\xi^{(2,2)} = 10$. Three different values of the variance hyperparameter are considered, $\sigma_\ell^2 = 8, 11, 15$. As shown in **Figure 7** (top), the noise intensity $\xi^{(2,2)} = 6$ for $\ell = 2$ was correctly estimated both when using $\sigma_2^2 = 11$ and $\sigma_2^2 = 15$. However, the unknown noise intensity was learned faster (within the first 10 steps) when $\sigma_2^2 = 11$ as this value better matched the actual noise intensity $\xi^{(2,2)} = 10$. Similarly, from **Figure 7** (bottom), the rate of learning $\xi^{(1,2)} = 6$ was faster with $\sigma_1^2 = 8$ than with $\sigma_1^2 = 15$.

3.3.2 TL-BNP: Effect of Varying Number of Learning Sources in Example 4

Figure 9 provides the estimation MSE performance comparison between tracking without TL ($L = 0$) and tracking using the TL-BNP approach with $L = 3$ and $L = 10$ learning sources for Example 4. Note that the TL-BNP is implemented using a particle filter (PF), as discussed in **Section 2.4.1**. The primary source time-varying noise intensity values ξ_k vary within $\Xi_p \in \{2, 8\}$. The variation with respect to time is as follows: the noise intensity was

$\xi_k \approx 2$ from $k = 1$ to $k = 30$, $\xi_k \approx 8$ from $k = 30$ to $k = 65$, and $\xi_k \approx 4$ from $k = 65$ to $k = 100$. As shown in **Figure 9**, the performance of the TL-BNP tracker is higher than that of the tracker without TL. It is also observed that the MSE performance using TL-BNP is higher for $L = 10$ than for $L = 3$. This is explained by considering the actual values of $\xi^{(\ell,3)}$ and $\xi^{(\ell,10)}$ in **Table 1**. Specifically, as the variation of ξ_k remains around values 2, 8, and 10, all three values are only in the set Ξ_L for $L = 10$ and not for $L = 3$.

For the same example, we also provide the range MSE in **Figure 8** for two additional numbers of preliminary sources, $L = 2$ and $L = 5$. It is interesting to note the similar MSE performance of the primary tracking source using $L = 5$ in **Figure 8** and $L = 10$ in **Figure 9**. This follows from the fact that the primary source noise intensity ξ_k takes only values 2, 8 and 4 throughout the $K = 100$ time steps, and both the $L = 5$ and $L = 10$ learning sources include all three values. Specifically, $\xi^{(1,5)} = \xi^{(5,10)} = 2$, $\xi^{(4,5)} = \xi^{(4,10)} = 8$, and $\xi^{(2,5)} = \xi^{(1,10)} = 4$.

3.3.3 TL-BNP: Algorithm Implementation in Example 4

Figure 8 also shows two additional MSE plots that correspond to a different implementation of the posterior PDF in Eq. 15. Specifically, the authors in (Caron et al., 2008) considered a tracking problem using DPMs to estimate measurement noise; their method did not include TL and also did not model the hyperparameter set Ψ_ℓ . They implemented their approach using a Kalman filter and a Rao-Blackwellized PF (RBPF). We incorporated their RBPF approach within our TL framework and hyperparameter modeling but with an extended Kalman filter as our measurement function is nonlinear. The performance comparison of the RBPF and our PF-based implementation in **Figure 8** showed a small improvement in performance for each L value when the PF is used. Note, however, that the RBPF is computationally more efficient than the PF.

3.3.4 TL-BNP: Effect of Initial NIWD Hyperparameters on Estimating Noise Intensity in Example 5

Similar to **Figure 7** in Example 4, we use **Figure 10** in Example 5 to study how the estimation accuracy of the learning source noise intensity $\xi^{(\ell,L)}$ is affected by the selection of the NIWD variance hyperparameter σ_ℓ^2 . In this example, we considered low intensity values for $\xi^{(\ell,L)}$ but high values for σ_ℓ^2 . Specifically, we used $L = 5$ learning sources with intensity values $\xi^{(2,5)} = 2$ and $\xi^{(3,5)} = 4$ from the set $\Xi_5 = \{1, 7\}$ (see **Table 1**) and we varied $\sigma_\ell^2 = 8, 11$. **Figure 10** (top) shows that, although both values of σ_3^2 resulted in learning $\xi^{(3,5)} = 4$, the learning process was faster when $\sigma_3^2 = 8$ was selected. Note that both $\sigma_2^2 = 8$ and $\sigma_2^2 = 11$ were slow to learn the mis-matched value of $5\xi^{(2,5)} = 2$.

3.3.5 TL-BNP: Effect of Varying Learning Source Intensity Values in Example 5

For the simulation in Example 5, we considered the noise intensity variation at the primary source to be was $\xi_k \approx 4$ from $k = 1$ to $k = 45$ and then $\xi_k \approx 10$ from $k = 45$ to $k = 100$. We compare the MSE performance of the TL-BNP tracker for $L = 5$ learning sources but with different noise intensity values, as listed in **Table 1**. In the first case, the learning source intensity set is $\Xi_5 = \{1, 7\}$ and, in the second case, it is $\Xi_5 = \{1, 10\}$. As

shown in **Figure 11**, both trackers perform about the same during the first 45 time steps. This is because $\xi^{(\ell,5)} = 4$ is included in both learning source cases. However, for the last 50 to 55 time steps, only the second tracker with $\Xi_5 = \{1, 10\}$ includes $\xi^{(\ell,5)} = 10$, matching the actual primary source noise intensity, and thus performs better than the first case with $\Xi_5 = \{1, 7\}$.

4 CONCLUSION

We proposed two methods for tracking a moving object under time-varying and unknown noise conditions at a primary source. Both methods use sequential Bayesian filtering with transfer learning, where multiple learning sources perform a similar tracking task as the primary source and provide it with prior information. The first method, the TL-GMM tracker, integrates transfer learning with parametric Gaussian mixture modeling to model the learning source measurement likelihood distributions. This method relies on the assumption that the noise intensity of each learning source is known and also that the learning source simultaneously track the same object as the primary source. As these assumptions limit the applicability of the TL-GMM in real tracking scenarios, we proposed a second method, the TL-BNP tracker, that integrates transfer learning with Bayesian nonparametric modeling. This method deals with the more realistic scenario where the learning sources do not track the same object and their measurement noise intensity is unknown and learned using Dirichlet process mixtures. The use of the Bayesian nonparametric learning method does not limit the number of modeling mixtures. Also, as the learning and primary sources do not need to track the same object, the learned models can be stored and accessed when needed. Using simulations, we demonstrated that the primary source tracking performance increases as the number of learning sources increases, provided that the learning source intensity values match the noise intensity variation at the primary source.

An important consideration in the proposed methods is the relevance of the learning sources selected by the primary

source. In particular, for the transfer to be successful, the noise intensity of most of the selected learning sources must match the range of possible noise intensity values of the primary source. As demonstrated by the simulations, the rate of learning the noise intensity was slow when there was a mismatch between the learning source intensity and the primary source noise variation. The methods would thus benefit from adapting the learning source selection process, for example, by using a probabilistic similarity measure as a selection criterion.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The authors confirm their contribution to the article as follows. OA developed and simulated the methods; AP-S supervised the work; and both authors reviewed the results and approved the final version of the manuscript.

FUNDING

This work was partially funded by AFOSR grant FA9550-20-1-0132.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frsip.2022.868638/full#supplementary-material>

REFERENCES

- Alotaibi, O., and Papandreou-Suppappola, A. (2021). "Bayesian Nonparametric Modeling and Transfer Learning for Tracking under Measurement Noise Uncertainty," in 2021 55th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, 31 Oct.-3 Nov. doi:10.1109/IEEECONF53345.2021.9723243
- Alotaibi, O., and Papandreou-Suppappola, A. (2020). "Transfer Learning with Bayesian Filtering for Object Tracking under Varying Conditions," in 2020 54th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 1-4 Nov. 2020, 1523-1527. doi:10.1109/IEEECONF51394.2020.9443276
- Antoniak, C. E. (1974). "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," in *The Annals of Statistics* (Beachwood, OH, United States: Institute of Mathematical Statistics), 1152-1174. doi:10.1214/aos/1176342871
- Arnold, A., Nallapati, R., and Cohen, W. W. (2007). "A Comparative Study of Methods for Transductive Transfer Learning," in Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), Omaha, NE, USA, 28-31 Oct. 2007, 77-82. doi:10.1109/icdmw.2007.109
- Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. (2002). A Tutorial on Particle Filters for Online Nonlinear/non-Gaussian Bayesian Tracking. *IEEE Trans. Signal Process.* 50, 174-188. doi:10.1109/78.978374
- Bar-Shalom, Y., and Fortmann, T. E. (1988). *Tracking and Data Association*. San Diego, CA, United States: Academic Press.
- Baxter, R. A. (2011). "Mixture Model," in *Encyclopedia of Machine Learning*. Editors C. Sammut and G. I. Webb (Boston, MA, United States: Springer), 680-682.
- Berntorp, K., and Cairano, S. D. (2016). "Process-noise Adaptive Particle Filtering with Dependent Process and Measurement Noise," in 2016 IEEE 55th Conference on Decision and Control (CDC), Las Vegas, NV, USA, 12-14 Dec. 2016, 5434-5439. doi:10.1109/cdc.2016.7799103
- Boers, Y., and Driessen, J. N. (2004). Multitarget Particle Filter Track before Detect Application. *IEEE Proc. Radar Sonar Navig.* 151, 351-357. doi:10.1049/ip-rsn:20040841
- Caron, F., Davy, M., Doucet, A., Duflos, E., and Vanheeghe, P. (2008). Bayesian Inference for Linear Dynamic Models with Dirichlet Process Mixtures. *IEEE Trans. Signal Process.* 56, 71-84. doi:10.1109/TSP.2007.900167
- Doucet, A., De Freitas, N., and Gordon, N. J. (2001). *Sequential Monte Carlo Methods in Practice*, 1. Boston, MA, United States: Springer.

- Ebenezer, S. P., and Papandreou-Suppappola, A. (2016). Generalized Recursive Track-Before-Detect with Proposal Partitioning for Tracking Varying Number of Multiple Targets in Low SNR. *IEEE Trans. Signal Process.* 64, 2819–2834. doi:10.1109/tsp.2016.2523455
- Escobar, M. D., and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *J. Am. Stat. Assoc.* 90, 577–588. doi:10.1080/01621459.1995.10476550
- Ferguson, T. S. (1973). “A Bayesian Analysis of Some Nonparametric Problems,” in *The Annals of Statistics* (Beachwood, OH, United States: Institute of Mathematical Statistics), 209–230. doi:10.1214/aos/1176342360
- Fox, E., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011). Bayesian Nonparametric Inference of Switching Dynamic Linear Models. *IEEE Trans. Signal Process.* 59, 1569–1585. doi:10.1109/tsp.2010.2102756
- Fräley, C., and Raftery, A. E. (2002). Model-based Clustering, Discriminant Analysis, and Density Estimation. *J. Am. Stat. Assoc.* 97, 611–631. doi:10.1198/016214502760047131
- Gómez-Villegas, M. A., Main, P., Navarro, H., and Susi, R. (2014). Sensitivity to Hyperprior Parameters in Gaussian Bayesian Networks. *J. Multivar. Analysis* 124, 214–225. doi:10.1016/j.jmva.2013.10.022
- Hastie, T., Tibshirani, R., and Friedman, J. (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2 edn. Boston, MA, United States: Springer.
- Hawkins, H. E., and La Plant, O. (1959). Radar Performance Degradation in Fog and Rain. *IRE Trans. Aeronaut. Navig. Electron.* ANE-6, 26–30. doi:10.1109/tane.1959.4201651
- Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge, England: Cambridge Univ. Press.
- Jaini, P., Chen, Z., Carbajal, P., Law, E., Middleton, L., Regan, K., et al. (2017). “Online Bayesian Transfer Learning for Sequential Data Modeling,” in *International Conference on Learning Representations*. Toulon, France: OpenReview.net
- Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Trans. ASME—Journal Basic Eng.* 82, 35–45. doi:10.1115/1.3662552
- Karbalayghareh, A., Qian, X., and Dougherty, E. R. (2018). Optimal Bayesian Transfer Learning. *IEEE Trans. Signal Process.* 66, 3724–3739. doi:10.1109/tsp.2018.2839583
- Kouw, W. M., and Loog, M. (2019). An Introduction to Domain Adaptation and Transfer Learning. *arXiv* 1812.11806.
- Lang, P., Fu, X., Martorella, M., Dong, J., Qin, R., and Meng, X. (2020). A Comprehensive Survey of Machine Learning Applied to Radar Signal Processing. *arXiv Eess*.2009.13702.
- Lindner, G., Shi, S., Vučetić, S., and Mišková, S. (2022). Transfer Learning for Radioactive Particle Tracking. *Chem. Eng. Sci.* 248, 1–16. doi:10.1016/j.ces.2021.117190
- Lindsay, B. G. (1995). “Mixture Models: Theory, Geometry, and Applications,” in *JSTOR, NSF-CBMS Regional Conference Series in Probability and Statistics*, 5 (Beachwood, OH, United States: Institute of Mathematical Statistics).
- Little, M. A. (2019). *Machine Learning for Signal Processing: Data Science, Algorithms, and Computational Statistics*. Oxford, England: Oxford University Press.
- Mitchell, T. M. (1997). *Machine Learning*. Oxford, England: McGraw-Hill.
- Moraffah, B., Brito, C., Venkatesh, B., and Papandreou-Suppappola, A. (2019). “Use of Hierarchical Dirichlet Processes to Integrate Dependent Observations from Multiple Disparate Sensors for Tracking,” in 2019 22th International Conference on Information Fusion (FUSION), Ottawa, ON, Canada, 2–5 July 2019, 1–7.
- Moraffah, B., and Papandreou-Suppappola, A. (2018). “Dependent Dirichlet Process Modeling and Identity Learning for Multiple Object Tracking,” in 2018 52nd Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 28–31 Oct. 2018, 1762–1766. doi:10.1109/acssc.2018.8645084
- Müller, P., and Mitra, R. (2013). Bayesian Nonparametric Inference - Why and How. *Bayesian Anal.* 8 (2), 269–302. doi:10.1214/13-BA811
- Munro, P., Toivonen, H., Webb, G. I., Buntine, W., Orbanz, P., Teh, Y. W., et al. (2011). “Bayesian Nonparametric Models,” in *Encyclopedia of Machine Learning*. Editors C. Sammut and G. I. Webb (Springer), 81–89. doi:10.1007/978-0-387-30164-8_66
- Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *J. Comput. Graph. Statistics* 9, 249–265. doi:10.1080/10618600.2000.10474879
- Pan, S. J., and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi:10.1109/tkde.2009.191
- Papež, M., and Quinn, A. (2019). “Robust Bayesian Transfer Learning between Kalman Filters,” in 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP), Pittsburgh, PA, USA, 13–16 Oct. 2019, 1–6. doi:10.1109/mlsp.2019.8918783
- Pereida, K., Helwa, M. K., and Schoellig, A. P. (2018). Data-efficient Multirobot, Multitask Transfer Learning for Trajectory Tracking. *IEEE Robot. Autom. Lett.* 3, 1260–1267. doi:10.1109/lra.2018.2795653
- Qiu, J., Wu, Q., Ding, G., Xu, Y., and Feng, S. (2016). A Survey of Machine Learning for Big Data Processing. *EURASIP J. Adv. Signal Process.* 2016. doi:10.1186/s13634-016-0355-x
- Rabaoui, A., Viandier, N., Duflos, E., Marais, J., and Vanheeghe, P. (2012). Dirichlet Process Mixtures for Density Estimation in Dynamic Nonlinear Modeling: Application to GPS Positioning in Urban Canyons. *IEEE Trans. Signal Process.* 60, 1638–1655. doi:10.1109/tsp.2011.2180901
- Reynolds, D. (2015). “Gaussian Mixture Models,” in *Encyclopedia of Biometrics* (Boston, MA, United States: Springer), 827–832. doi:10.1007/978-1-4899-7488-4_196
- Rojo-Álvarez, J. L., Martínez-Ramón, M., Muñoz-Marí, J., and Camps-Valls, G. (2018). *From Signal Processing to Machine Learning*. Hoboken, NJ, United States: Wiley-IEEE Press, 1–11. chap. 1. doi:10.1002/9781118705810.ch1
- Salmond, D. J., and Birch, H. (2001). “A Particle Filter for Track-Before-Detect,” in Proceedings of the 2001 American Control Conference. (Cat. No.01CH37148), Arlington, VA, USA, 25–27 June 2001, 3755–3760. doi:10.1109/acc.2001.946220
- Sethuraman, J. (1994). “A Constructive Definition of Dirichlet Priors,” in *Statistica Sinica* (Taipei, Taiwan: Institute of Statistical Science, Academia Sinica), 639–650.
- Theodoridis, S. (2020). *Machine Learning: A Bayesian and Optimization Perspective*. 2 edn. San Diego, CA, United States: Academic Press.
- Tonissen, S. M., and Bar-Shalom, Y. (1988). “Maximum Likelihood Track-Before-Detect with Fluctuating Target Amplitude,” in *IEEE Transactions on Aerospace and Electronic Systems* (Piscataway, NJ, United States: IEEE), 34, 796–809.
- Torrey, L., and Shavlik, J. (2010). “Transfer Learning,” in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. Editors E. S. Olivas, J. D. M. Guerrero, M. M. Sober, J. R. M. Bedito, and A. J. S. Lopez (Hershey, PA, United States: Information Science Reference), 242–264. chap. 11. doi:10.4018/978-1-60566-766-9.ch011
- Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A Survey of Transfer Learning. *J. Big Data* 3, 9. doi:10.1186/s40537-016-0043-6
- West, M. (1992). *Hyperparameter Estimation in Dirichlet Process Mixture Models*. NC: Duke University. Tech. rep.
- Xuan, J., Lu, J., and Zhang, G. (2019). “A Survey on Bayesian Nonparametric Learning,” in *ACM Computing Surveys*, 52 (New York, NY, United States: Association for Computing Machinery).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Alotaibi and Papandreou-Suppappola. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



OPEN ACCESS

EDITED BY
Frederic Dufaux,
CentraleSupélec, France

REVIEWED BY
Abdullah Bülbül,
Ankara Yıldırım Beyazıt University,
Turkey
Mehmet Turkan,
İzmir University of Economics, Turkey

*CORRESPONDENCE
Sana Alamgeer,
sanaalamgeer@gmail.com
Mylène C. Q. Farias,
mylene@ieee.org

SPECIALTY SECTION
This article was submitted to Image
Processing,
a section of the journal
Frontiers in Signal Processing

RECEIVED 15 November 2021

ACCEPTED 29 July 2022

PUBLISHED 26 August 2022

CITATION
Alamgeer S and Farias MCQ (2022),
Blind visual quality assessment of light
field images based on distortion maps.
Front. Sig. Proc. 2:815058.
doi: 10.3389/frsip.2022.815058

COPYRIGHT
© 2022 Alamgeer and Farias. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which does
not comply with these terms.

Blind visual quality assessment of light field images based on distortion maps

Sana Alamgeer* and Mylène C. Q. Farias*

¹Department of Electrical Engineering, University of Brasília, Brasília, Brazil

Light Field (LF) cameras capture spatial and angular information of a scene, generating a high-dimensional data that brings several challenges to compression, transmission, and reconstruction algorithms. One research area that has been attracting a lot of attention is the design of Light Field images quality assessment (LF-IQA) methods. In this paper, we propose a No-Reference (NR) LF-IQA method that is based on reference-free distortion maps. With this goal, we first generate a synthetically distorted dataset of 2D images. Then, we compute SSIM distortion maps of these images and use these maps as ground error maps. We train a GAN architecture using these SSIM distortion maps as quality labels. This trained model is used to generate reference-free distortion maps of sub-aperture images of LF contents. Finally, the quality prediction is obtained performing the following steps: 1) perform a non-linear dimensionality reduction with a isometric mapping of the generated distortion maps to obtain the LFI feature vectors and 2) perform a regression using a Random Forest Regressor (RFR) algorithm to obtain the LF quality estimates. Results show that the proposed method is robust and accurate, outperforming several state-of-the-art LF-IQA methods.

KEYWORDS

image quality assessment, epipolar planes, canny edge detector, two-stream convolution neural network, 4D light field images

1 Introduction

Unlike conventional cameras, Light Field (LF) camera captures spatial and angular information of scene, which is represented by a scalar function $L(u, v, s, t)$, where (u, v) and (s, t) depict the angular and spatial domains, respectively. The 4D light field can be described as a 2D projection of sub-aperture images (SAIs). Figure 1 illustrates a grid of 10×10 sub-aperture images of LFI (ArtGallery3) from the MPI dataset (Adhikarla et al., 2017). SAIs are generated from micro-lens images, with an operation known as raw data decoding. LF images (LFI) carry rich information that is widely used for refocusing (Hahne et al., 2018) and 3-Dimensional (3D) reconstruction. However, the high-dimensionality of LFIs creates several challenges to the area of communications, requiring the development of specific compression (Hou et al., 2019), transmission, and reconstruction techniques. Unfortunately, these techniques inevitably distort the perceived quality of LFIs (Paudyal et al., 2017). In order to monitor the visual quality and

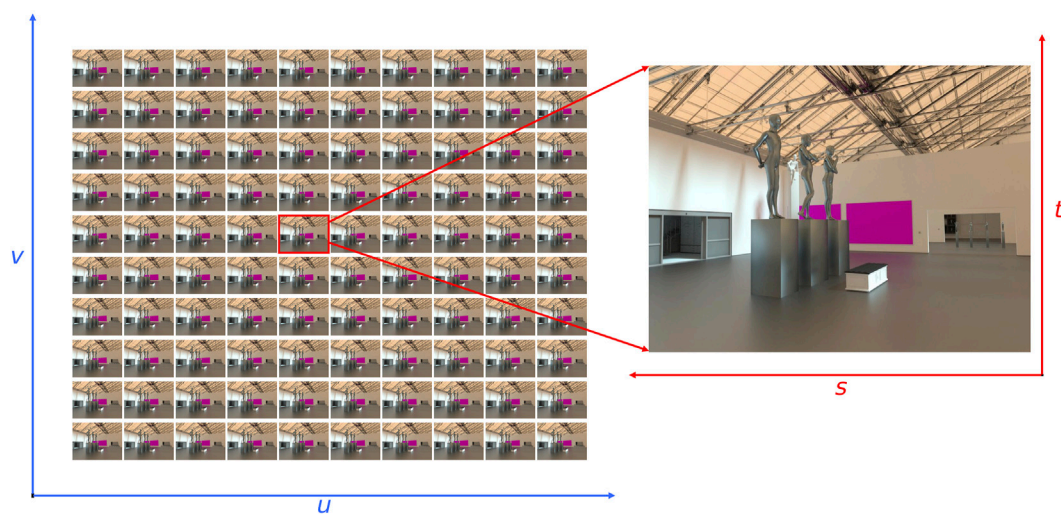


FIGURE 1

A grid of 10 × 10 sub-aperture images of a Light Field image (ArtGallery3) from MPI dataset (Adhikarla et al., 2017).

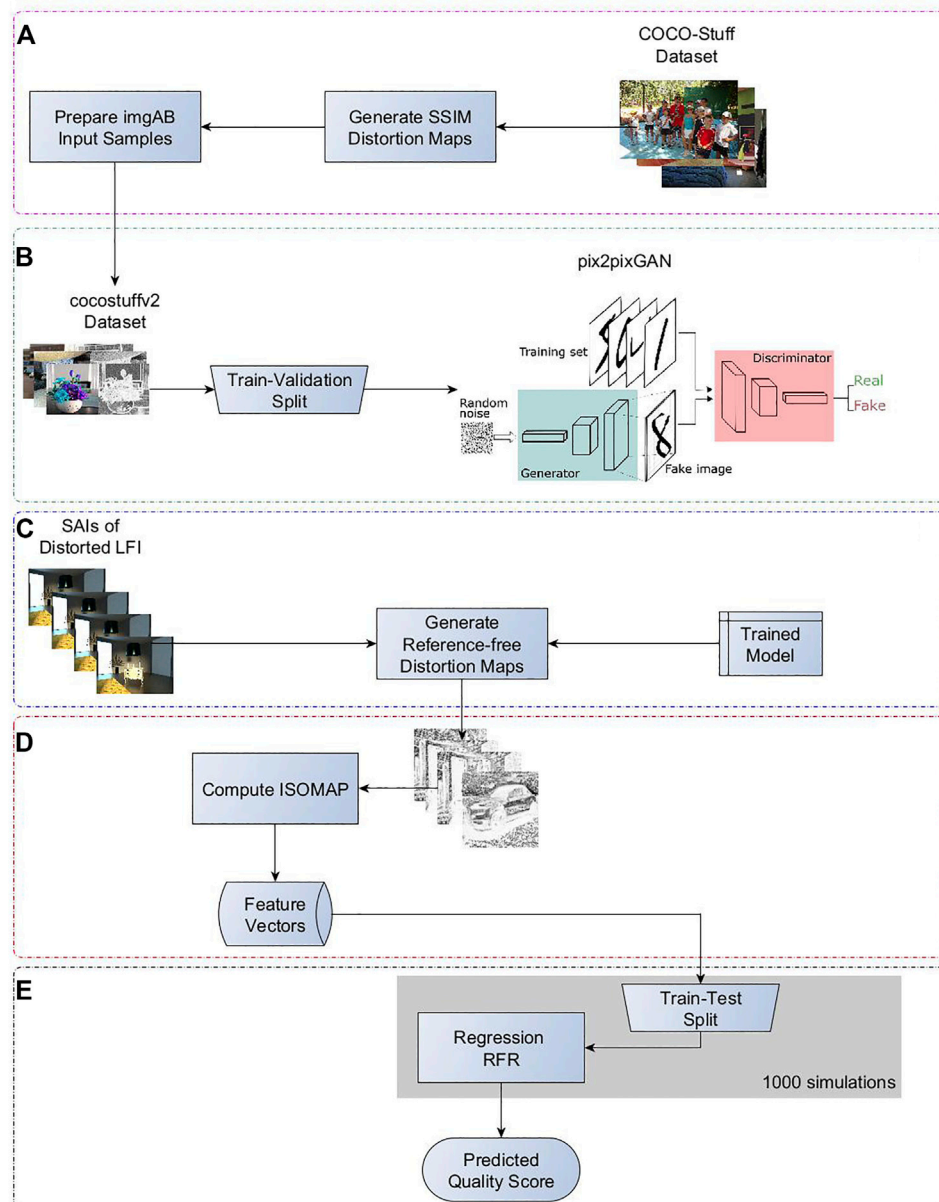
quantify the amount of visual distortion in LF contents, we must develop efficient LFI Quality Assessment (LF-IQA) methods.

The visual quality of images can be assessed (IQA) subjectively or objectively. Subjective quality assessment methods are experimental methodologies in which human observers are asked to estimate one or more features of the stimuli (e.g., the overall quality of a video). The data collected in these experiments are statistically analyzed, often generating Mean Opinion Scores (MOS) for each of the test stimuli. Although subjective quality assessment methods are considered as the most reliable ways of estimating quality, these methods are time-consuming and cannot be implemented in real-time systems. Objective quality assessment methods (also known as quality metrics) are algorithms that automatically assess the quality of a content by measuring the physical signal. Based on the available reference information, objective quality assessment methods are divided into the full-reference (FR), reduced-reference (RR) and blind/no-reference (NR) methods. To estimate the performance of objective quality methods, we compare their output scores with MOS values obtained using subjective methods.

As mentioned before, LFIs contain not only spatial information, but also angular information. Therefore, classical 2D image quality assessment methods cannot be directly used for LFI quality assessment. In the past few years, efforts have been devoted to develop LF-IQA methods. For example, Tian et al. (Tian et al., 2018) proposed a FR LF-IQA metric that uses a multi-order derivative feature-based method (MDFM), which extracts detailed features with different degrees with a discrete derivative filter. Paudyal et al. (Paudyal et al., 2019) proposed a RR LF-IQA that uses two IQA methods - SSIM (Wang et al., 2004) and PSNR - to process the depth maps. Fang et al. (Fang

et al., 2018) presented a FR LF-IQA method that uses local and global features to predict quality. The local features are extracted from reference and test LFIs using a gradient magnitude operator, while the global features are extracted from the epipolar-plane images (EPIs) of reference and test LFIs using the same operator. Tian et al. (Tian et al., 2020b) proposed a FR LF-IQA method that uses symmetry information, which are extracted using Gabor filters (Field, 1987), and depth features computed from EPIs of the reference and test LFIs. Meng et al. (Meng et al., 2020) presented a FR LF-IQA method that computes the spatial LFI quality using the structural similarity index metric (SSIM) of the Difference of Gaussian (DoG) features of central SAIs and computes the angular LFI quality using the SSIM of refocused images.

Tian et al. (Tian et al., 2020a) presented a FR LF-IQA method in which the salient features are extracted from reference and test EPIs and SAIs using single-scale and multi-scale log-Gabor operators. The NR LFQA method proposed by Shi et al. (Shi et al., 2019a) predicts quality using EPI information and natural statistics. The NR LF-IQA method proposed by Luo et al. (Luo et al., 2019) employs the spatial information from SAIs and the angular information from the micro-lens images. Jiang et al. (Jiang et al., 2018) proposed a FR LF-IQA method that uses the entropy information and the gradient magnitude features to extract spacial features from the SAIs. To extract the angular features from SAIs, dense distortion curves are generated and the best fitting features are chosen. Wei Zhou et al. (Shi et al., 2019b; Zhou et al., 2019) proposed NR LF-IQA methods (BELIF and Tensor-NLFQ, respectively) that are based on tensor theory, employing SAIs view stacks along horizontal, vertical, left diagonal, and right diagonal orientations. The local spatial quality features are extracted using local frequency

**FIGURE 2**

Block Diagram of the proposed no-reference light field image quality assessment method. **(A)** computation of SSIM distortion maps corresponding to the original and test images, **(B)** training the GAN network using the SSIM distortion maps as labels, **(C)** testing the trained GAN network to generate reference-free distortion maps of sub-aperture images of test LFIs, **(D)** computation of ISOMAP to generate feature vectors for distortion maps generated in **(C)**, and **(E)** training Random Forest Regressor with 1000 simulations for quality predictions.

distribution, while the global spatial quality features are extracted using the naturalness distribution of individual color channels. Ak *et al.* (Ak *et al.*, 2020) proposed a NR LF-IQA method based on the structural representations of EPIs, by training a convolutional sparse coding codebook and a Bag Of World dictionary on EPIs. Shan *et al.* (Shan *et al.*, 2019) designed a NR LF-IQA method that is based on 2D (from SAIs) and 3D (from EPIs) LFI features. Xiang *et al.* (Xiang *et al.*, 2020)

presented a NR LF-IQA method (VBLFI) based on the mean difference image and on curvelet-transform characteristics of LFIs.

Despite of the work mentioned above, there is a lot of room for improvement in terms of prediction accuracy, robustness, computational complexities, and generality of NR LF-IQA methods. In this paper, we propose a NR LF-IQA method that is based on reference-free distortion maps. Considering



FIGURE 3
Sample images taken from COCO-Stuff dataset (Caesar et al., 2018).

that pixel distortions are affected by neighboring pixels, we have focused on designing a blind deep learning quality model using pixel-by-pixel distortion maps. In summary, we present two main contributions: 1) generation of reference-free distortion maps, and 2) NR LF-IQA method that is derived from the generated distortion maps.

To generate LF reference-free distortion maps, we have used a deep-learning architecture called Generative Adversarial Network (GAN) network (Goodfellow et al., 2014) that can learn from synthetically generated distorted images and their corresponding ground truth distortion maps. Since ground truth distortion maps are not available in any of the existing IQA datasets (Sheikh et al., 2006), we use distortion maps generated by SSIM (Wang et al., 2004) as ground-truth distortion maps to train the GAN. Specifically, first we generate a synthetically distorted dataset of 2D images (because the sub-aperture images are 2D representation of LFIs) and, then, we compute SSIM distortion maps corresponding to the original and test images, as shown in Figure 2A. Then, we train the GAN network using the SSIM distortion maps as labels, as shown in Figure 2B. The trained model is used to generate reference-free distortion maps of sub-aperture images of test LFIs, as shown in Figure 2C. The generated distortion maps (GDMs) are used as measurement maps for describing the test LFIs. Results show that the proposed method outperforms other state-of-the-art LF-IQA methods.

The rest of the paper is organized as follows. Section 2 describes the proposed LF-IQA method. Section 3 describes

the experimental results. Finally, Section 5 presents our conclusion.

2 Proposed methodology

Generative Adversarial Networks (GAN) consists of a pair of competing network structures called generator (G) and discriminator (D) respectively, which can learn deep features with sufficient labeled training data. In this work, to learn the features of distorted images, we use the Pix2PixGAN architecture (Isola et al., 2017) for the GAN architecture because of its strong fitting capability. The Pix2PixGAN is composed of promising approach for many image-to-image translation tasks, especially those involving highly structured graphical outputs. Most importantly, the Pix2PixGAN is general-purpose, i.e., it learns a loss adapted to the task and data at hand, which makes it feasible in a wide variety of settings.

Since a GAN architecture requires a large number of training samples and LF-IQA datasets do not have a large number of samples, we use the COCO-Stuff dataset (Caesar et al., 2018) to generate a synthetically distorted dataset. The COCO-Stuff dataset is derived from the COCO dataset (Lin et al., 2014). This dataset has 1.2 million images captured from diverse scenes, with a total of 182 semantic classes. Sample images of COCO-Stuff dataset are shown in Figure 3.

To generate synthetic distorted versions of 1.2 million images, we used the Albumentation library (Buslaev et al., 2020). In total,

TABLE 1 Distortions used from Albumentation Library.

Number	Function	Description
1	VerticalFlip	Flip the input vertically around the x -axis
2	HorizontalFlip	Flip the input horizontally around the y -axis
3	IAAPerspective	Apply random four point perspective transformations to images
4	RandomRotate90	Randomly rotate the input by 90°
5	Transpose	Transpose the input by swapping rows and columns
6	ShiftScaleRotate	Randomly apply affine transforms: translate, scale and rotate the input
7	Blur	Blur the input image using a random-sized kernel
8	OpticalDistortion	Image magnification decreases with distance from the optical axis. Straight lines appear to bend outwards from the center of the image
9	GridDistortion	Grid-distortion is an image warping technique which is driven by the mapping between equivalent families of curves, arranged in a grid structure Arad (1998)
10	HueSaturationValue	Randomly change hue, saturation and value of the input image
11	IAAAdditiveGaussianNoise	Apply additive gaussian noise to the input image
12	GaussNoise	Apply gaussian noise to the input image
13	MotionBlur	Apply motion blur to the input image using a random-sized kernel
14	MedianBlur	Blur the input image using a median filter with a random aperture linear size
15	IAAPiecewiseAffine	Place a regular grid of points on each image and then randomly move each point around by 1–5 percent with respect to the image height and width
16	IAASharpener	Sharpen the input image and overlays the result with the original image
17	IAAEmboss	Emboss the input image and overlays the result with the original image
18	RandomContrast	Adjust the contrast of an image or images by a random factor
19	RandomBrightness	Randomly change brightness of the input image
20	Flip	Flip the input vertically around the x -axis
21	strong_aug_oneOfs	Custom function combined of distortions IAAAdditiveGaussianNoise, GaussNoise, MotionBlur, Blur, OpticalDistortion, GridDistortion, IAAPiecewiseAffine, CLAHE, IAASharpener, IAAEmboss, RandomContrast, RandomBrightness and HueSaturationValue
22	augment_flips_color	Custom function combined of distortions CLAHE, RandomRotate90, Transpose, ShiftScaleRotate, Blur, OpticalDistortion, GridDistortion and HueSaturationValue
23	RGBShift	Randomly shift values for each channel of the input RGB image
24	JpegCompression	Decrease Jpeg compression of an image
25	ToGray	Convert the input RGB image to grayscale. If the mean pixel value for the resulting image is greater than 127, invert the resulting grayscale image
26	RandomGamma	Draw samples from a Gamma distribution
27	InvertImg	Invert the input image by subtracting pixel values from 255
28	ChannelShuffle	Randomly rearrange channels of the input RGB image
29	CLAHE	Apply Contrast Limited Adaptive Histogram Equalization to the input image

we used 29 augmentation functions with pre-set parameters, which are depicted in [Table 1](#). In total, we generated a dataset with 3.57 million synthetically distorted images, which we named cocostuffv2. Then, we computed SSIM distortion maps between each reference and test images in the cocostuffv2 dataset. For training the Pix2PixGAN architecture, we prepared an input tuple *imgAB* of size 512×256 , which consists of a concatenation of the test image *A* and the SSIM distortion map *B*. [Figure 4](#) shows some examples of input tuples *imgAB*.

For training the Pix2PixGAN network, we divided the cocostuffv2 dataset into two content-independent training and validation subsets, i.e. distorted images generated from one reference in the test subset are not present in the training

subset and vice-versa. We define a group of scenes as a group containing the reference LFI and its corresponding test versions. Then, 80% of the groups were randomly selected for training and the remaining 20% were used for validation. It is worth mentioning that we trained the Pix2PixGAN network from scratch (instead of using pre-trained model) with 50 epochs.

Next, the trained Pix2PixGAN network is used to generate reference-free distortion maps of the sub-aperture images of corresponding test LFIs. [Figure 5](#) illustrates examples of generated distortion maps of central LF SAs taken from the MPI dataset ([Adhikarla et al., 2017](#)). Even though we have not used any of LFIs in the training process, the Pix2PixGAN network is able to localize distortions in test SAs. As

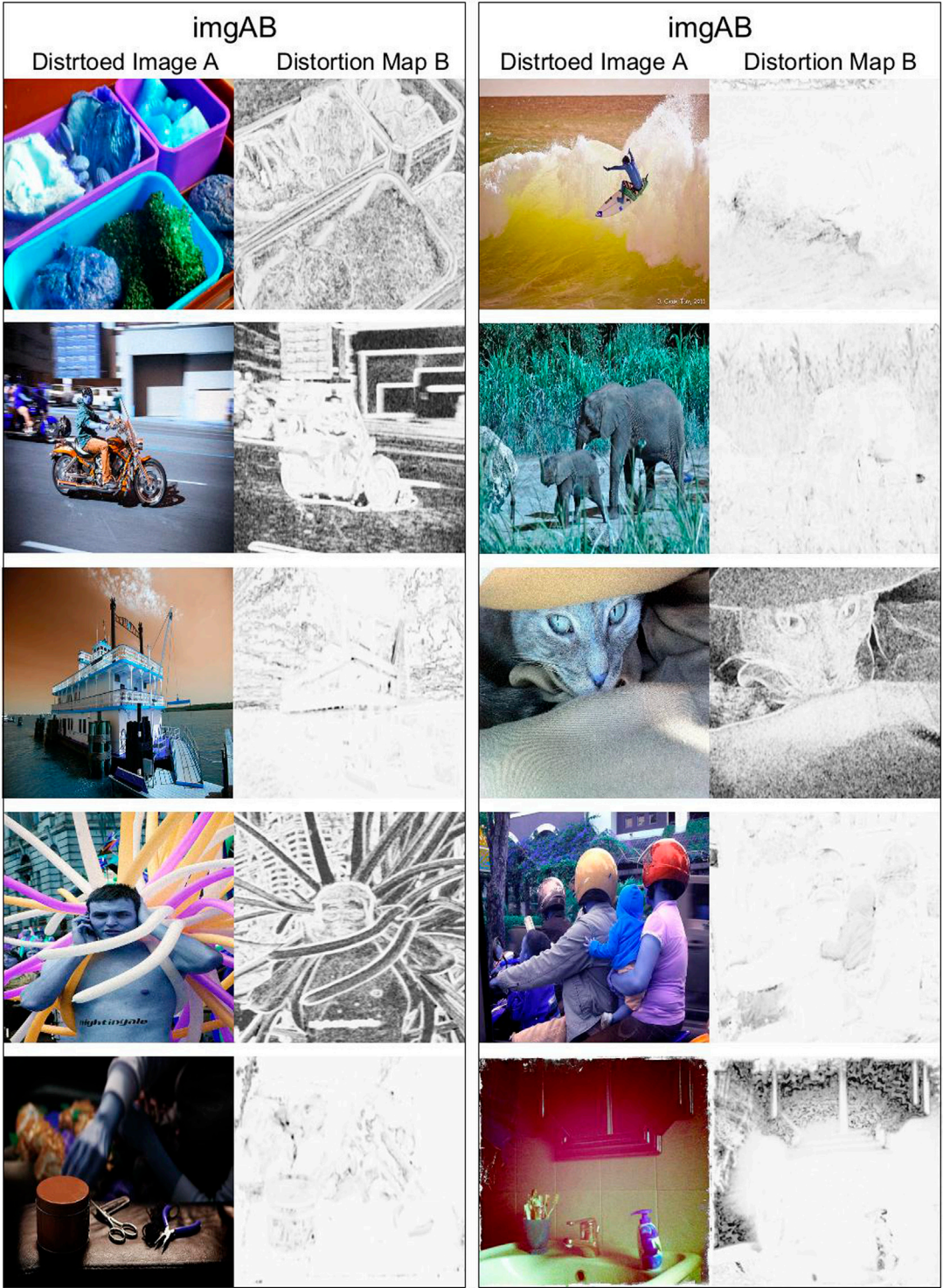


FIGURE 4
Random input samples from cocstuffv2 dataset. Distorted image A is obtained by Albumentation library of augmentations, where Distortion Map B is obtained by SSIM index method.

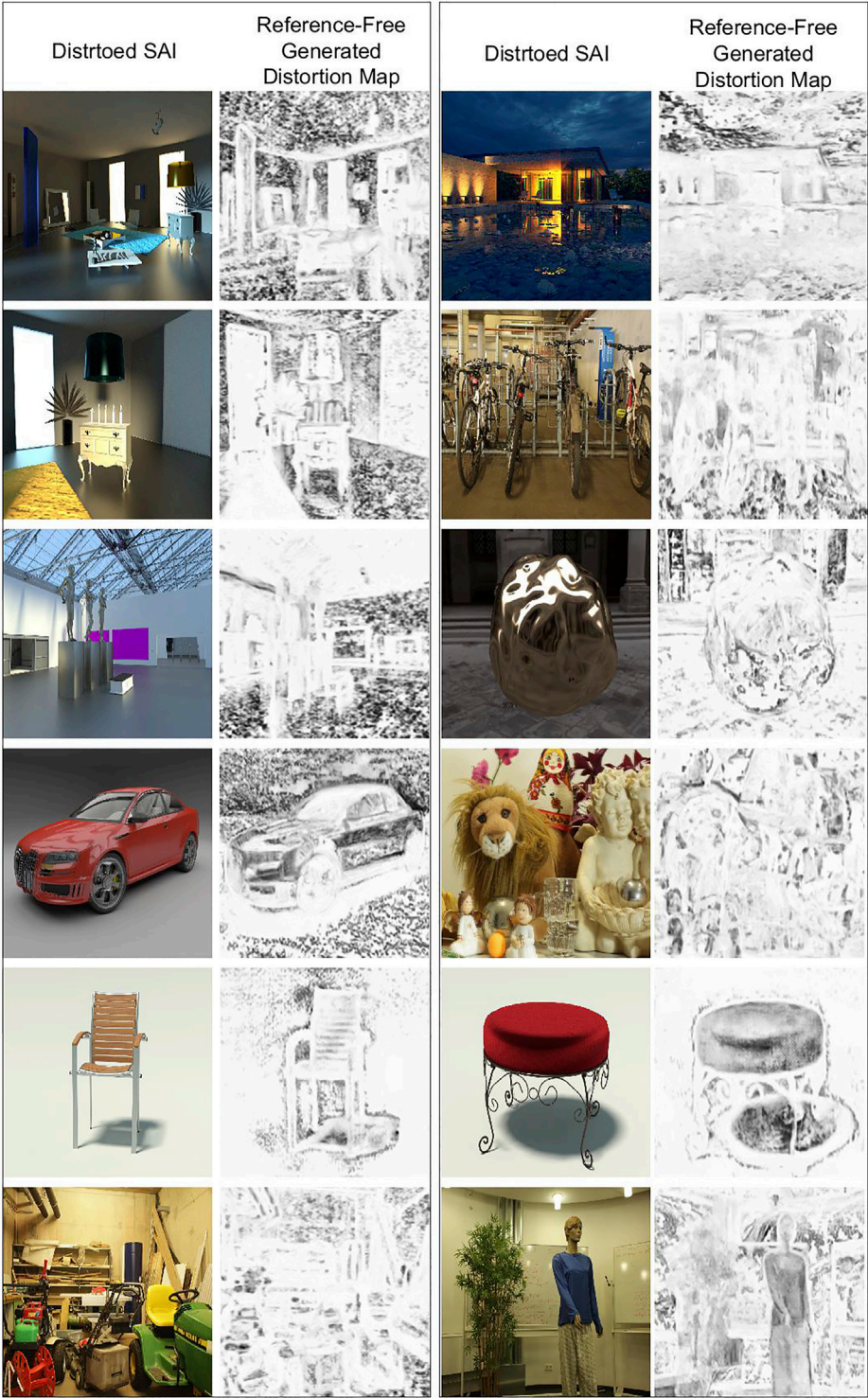


FIGURE 5
Examples of generated distortion maps of central SAIs of different test LFI from MPI dataset (Adhikarla et al., 2017)

illustrated in Figure 2D, to prepare the feature vectors for test LFIs, we perform a non-linear dimensionality reduction using an Isometric Mapping (Tenenbaum et al., 2000) (ISOMAP) on the generated distortion maps. The ISOMAP algorithm contains three stages. First, it computes k nearest neighbors. Then, it searches for and establishes the shortest path graph. Finally, the Eigen-vectors are computed over the largest Eigen values. The algorithm outputs feature vectors in a multi-dimensional Euclidean space that best represent the intrinsic geometry of the data. As illustrated in Figure 2E, the prepared features vectors of LFIs are fed to a Random Forest Regressor, which performs a regression to predict the LF quality. We chose the RFR because in previous studies it has shown a robust performance (Fernández-Delgado et al., 2014; Freitas et al., 2018), when compared to other machinelearning algorithms (e.g. neural networks, support vector machines, generalized linear models, etc.).

3 Experimental setup

To train and test the proposed method, we have used the following four LF image quality datasets. We have chosen these datasets because of their content diversity, the types of distortions, and the availability of the corresponding subjective quality scores.

- The MPI Light Field image quality dataset (Adhikarla et al., 2017) contains 13 different scenes with references, 336 test LFIs, and the corresponding subjective quality scores (Mean Observer Scores - MOS). This dataset has typical light field distortions that are specific to transmission, reconstruction, and display.
- The VALID Light Field image quality dataset (Viola and Ebrahimi, 2018) contains five contents, taken from EPFL (Rerabek and Ebrahimi, 2016) light field image dataset, and 140 test LFIs that are compressed using state-of-the-art compression algorithms.

The dataset contains both subjective (MOS) and objective quality scores (PSNR and SSIM).

- The SMART Light Field image quality dataset (Paudyal et al., 2016, 2017) has 16 original LFIs representing both indoor and outdoor scenes. The image content corresponds not only to the scenes with different levels of colorfulness, spatial information, and texture, but also LF specific characteristics such as reflection, transparency, and depth of field. The dataset also contains 256 distorted sequences obtained using four compression algorithms, with their corresponding MOS values.
- The Win5-LID Light Field image quality dataset (Shi et al., 2018) contains six real scenes (captured by a Lytro Illum) and four synthetic scenes, with a total of 220 test LFIs. The

TABLE 2 Mean SROCC and PLCC values for VALID, SMART, MPI, and Win5-LID datasets obtained by 1,000 simulations of RFR.

Dataset	Distortion	Proposed	
		SROCC	PLCC
MPI	QD	0.9290	0.9866
	Gaussian	0.9886	0.9698
	HEVC	0.9581	0.9876
	OPT	0.9347	0.9394
	Linear	0.9499	0.9960
	NN	0.9753	0.9955
	ALL	0.9743	0.9878
VALID	10bit_HEVC	0.9275	0.9871
	10bit_P3	0.9864	0.9931
	10bit_P5	0.9866	0.9843
	10bit_VP9	0.9258	0.9797
	8bit_HEVC	0.9758	0.9678
	8bit_VP9	0.9380	0.9781
	ALL	0.9650	0.9388
SMART	HEVC	0.9101	0.9463
	JPEG	0.9069	0.9501
	JPEG2000	0.9529	0.8947
	SSDC	0.9050	0.9713
	ALL	0.9307	0.9529
Win5-LID	HEVC	0.9690	0.9398
	JPEG2000	0.9367	0.9752
	LN	0.9550	0.9148
	NN	0.9324	0.9286
	EPICNN	0.9059	0.9383
	ALL	0.9441	0.9535

selected contents carry abundant semantic features, such as people, nature, and objects. The LFIs have an identical angular resolution of $9, \times, 9$. The real scenes have spatial resolution equal to 434×625 , while the synthetic scenes have a spatial resolution equal to 512×512 . The distortions are obtained with compression and interpolation algorithms.

As performance evaluation methods, we used only the Spearman's Rank-Order Correlation Coefficient (SROCC) and the Pearson's Linear Correlation Coefficient (PLCC) for simplicity. We compared the proposed NR LF-IQA method with the following state-of-art LF-IQA methods: MDFM (Tian et al., 2018), LFIQM (Paudyal et al., 2019), Fang et al. (Fang et al., 2018), SDFM (Tian et al., 2020b), Meng et al. (Meng et al., 2020), LGF-LFC (Tian et al., 2020a), NR-LFQA (Shi et al., 2019a), LF-QMLI (Luo et al., 2019), Jiang et al. (Jiang et al., 2018), BELIF Shi

TABLE 3 SROCC and PLCC values obtained for state-of-the-art LF-IQA methods tested on VALID, SMART, MPI, and Win5-LID datasets.

Category	Type	Methods	Year	MPI		VALID		SMART		Win5-LID	
				SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
Pre-defined Functions	FR	UQI	2002	0.7400	0.8460	0.9310	0.9550	0.6480	0.7980	0.8252	0.8764
	FR	SSIM	2004	0.9120	0.9320	0.9500	0.9640	0.7550	0.8010	0.6812	0.7880
	FR	VIF	2006	0.8600	0.8960	0.9620	0.9790	0.7260	0.8370	0.9347	0.9555
	FR	NICE	2009	0.5821	0.5122	0.6211	0.6544	0.5214	0.5426	0.4892	0.5002
	FR	STMAD	2011	0.8650	0.8940	0.7940	0.8020	0.6640	0.8010	0.8489	0.9074
	FR	IW-SSIM	2011	0.9320	0.9440	0.9650	0.9780	0.8060	0.8850	0.8212	0.8736
	FR	IW-PSNR	2011	0.9300	0.9160	0.9470	0.9670	0.7840	0.8520	0.8842	0.9022
	FR	MJ3DFR	2013	0.8720	0.9300	0.9560	0.9700	0.8160	0.8480	0.8836	0.8998
	FR	GMSD	2014	0.7358	0.7410	0.6821	0.6948	0.7264	0.8000	0.4352	0.5041
	FR	FI-PSNR	2014	0.8740	0.8510	0.7060	0.7060	0.7730	0.8320	0.6951	0.7419
	FR	PSNR-YUV	2014	0.9342	0.9452	0.9230	0.9310	0.9102	0.9211	0.9007	0.9215
	FR	MW-PSNR	2016	0.7251	0.7698	0.6869	0.6904	0.5281	0.5869	0.7582	0.7758
	FR	MDFM Tian et al. (2018)	2018	0.8346	0.8123	0.7120	0.7198	0.7535	0.7683	0.8157	0.8591
	FR	Fang et al. (2018)	2018	0.8065	0.7942	—	—	—	—	—	—
	RR	LFIQM Paudyal et al. (2019)	2019	0.6815	0.7013	0.3934	0.5001	0.4503	0.4763	0.4503	0.4763
	FR	SDFM Tian et al. (2020b)	2020	0.8435	0.8423	0.824	0.8542	0.7514	0.7941	0.6742	0.7142
	FR	Meng et al. (2020)	2020	—	—	0.9579	0.9762	—	—	—	—
	FR	LGF-LFC Tian et al. (2020a)	2020	0.8543	0.8476	—	—	0.8246	0.8276	—	—
DL	FR	Jiang et al. (2018)	2018	—	0.8954	—	—	—	—	—	—
	NR	BELIF Shi et al. (2019a)	2019	0.8854	0.9096	0.8863	0.8950	0.8367	0.8833	0.8719	0.8910
	NR	NR-LFQA Shi et al. (2019b)	2019	0.9119	0.9155	0.9257	0.9658	0.8803	0.9105	0.9032	0.9206
	NR	LF-QMLI Luo et al. (2019)	2019	—	—	0.9286	0.9683	—	—	0.8802	0.9038
	NR	Shan et al. (2019)	2019	—	—	—	—	0.8917	0.9106	—	—
	NR	Tensor-NLFQ Zhou et al. (2019)	2019	0.9101	0.9225	0.9326	0.9746	0.8702	0.9028	0.9101	0.9217
	NR	Ak et al. (2020)	2020	0.8942	0.9005	—	—	—	—	—	—
	NR	VBLIF Xiang et al. (2020)	2020	0.9015	0.9158	—	—	—	—	0.9009	0.9232
DL + ML	NR	Proposed	2020	0.9743	0.9878	0.9650	0.9781	0.9307	0.9529	0.9441	0.9535

et al. (2019b), Tensor-NLFQ (Zhou et al., 2019), Ak et al. (Ak et al., 2020), Shan et al. (Shan et al., 2019) and VBLIF (Xiang et al., 2020). We also compared the proposed method with the following 2D image/video quality metrics: PSNR-YUV (Sze et al., 2014), IW-PSNR (Wang and Li, 2011), FI-PSNR (Lin and Wu, 2014), MW-PSNR (Sandić-Stanković et al., 2016), SSIM (Wang et al., 2004), IW-SSIM (Wang and Li, 2011), UQI (Zhou and Bovik, 2002), VIF (Sheikh and Bovik, 2006), MJ3DFR (Chen et al., 2013), GMSD (Xue et al., 2014), NICE (Rouse and Hemami, 2009) and STMAD (Vu et al., 2011).

For training and testing the RFR method, we divided each dataset into two content-independent training and testing subsets, i.e., distorted images generated from one reference in the testing subset are not present in the training subset and vice-

versa. We define a group of scenes as a group containing the reference LFI and its corresponding distorted versions. Then, 80% of the groups were randomly selected for training and the remaining 20% were used for testing. The partition was repeated 1,000 times to eliminate the bias caused by data division. We reported the mean correlation values for the test set over 1,000 simulations.

4 Experimental results

Table 2 shows the correlation values obtained for the VALID, SMART, MPI, and Win-LID LFI quality datasets. The rows in this table show the results for each dataset and for each

distortion, with the “All” row corresponding to the results obtained for the complete datasets. The proposed method performs very well for the MPI dataset obtaining SROCC of 0.97 and PLCC of 0.98, and for the VALID dataset obtaining SROCC of 0.96 and PLCC of 0.93. For the SMART dataset, the method obtains SROCC of 0.93 and PLCC of 0.95, while for the Win5-LID dataset, the method achieves SROCC of 0.94 and PLCC of 0.95. Across the different distortions, the proposed method also performed very well, with only a few distortions showing slightly lower values (e.g. SROCC values of JPEG and SSDC in SMART dataset, and EPICNN in Win5-LID dataset).

Table 3 illustrates the comparison of the results with other state-of-the-art LFI-IQA methods. In this table, the NR and FR LF-IQA methods are classified into three categories, taking into consideration the models used to map the pooled features into quality estimates. The categories include methods that use 1) a pre-defined function, 2) a machine-learning (ML) algorithm, or 3) a deep-learning (DL) approach to obtain the predicted quality score. Notice that, for simplicity, only the overall performance (“ALL”) correlation values are reported for each dataset. Also, since the authors of these LF-IQA methods did not publish their results for all four datasets, our matrix is incomplete. For VALID dataset, NR-LFQA (Shi et al., 2019a) and Tensor-NLFQ (Zhou et al., 2019) methods have reported correlations separately for 8bit and 10bit compressed LF images. For comparison of results in Table 3, we have shown averaged SROCC and PLCC obtained by these methods for complete VALID dataset. Notice that the proposed method has achieved the highest correlation values among all LF-IQA methods for all of the four datasets. It is also worth pointing out that the pooling and mapping strategies in the proposed NR LF-IQA method has achieved significant improvement in quality predictions and shown higher SROCC and PLCC than the original SSIM method.

5 Conclusion

In this paper, we have proposed a blind LF-IQA method that is based on reference-free distortion maps. To generate reference-free distortion maps from test LFIs, we have used a GAN deep-learning architecture, the Pix2PixGAN, which learns from synthetically generated distorted images and their corresponding ground truth distortion maps. Since the ground truth distortion maps are not available in any of the existing LF or image quality datasets, we use distortion maps generated by SSIM as the ground truth distortion map. Next, we train the Pix2PixGAN using the synthetically generated dataset. The proposed LF-IQA method has following five stages: 1) Generation of a synthetically distorted dataset of 2D images, 2) Pix2PixGAN training to generate 2D distortion maps, using SSIM distortion maps as ground truth, 3) generation of distortion

maps of sub-aperture images using the trained Pix2PixGAN, 4) non-linear reduction of dimensionality through Isometric Mapping on the generated distortion maps to obtain the LFI feature vectors, and 5) perform regression using RFR algorithm to predict LFI quality. The correlation values of the proposed method computed on four different datasets are higher than what is obtained by other state-of-the-art LF-IQA methods. As future work, we plan to explore using different state-of-the-art FR-IQA metrics to generate the ground truth distortion maps, and train the Pix2PixGAN architecture. It is worth pointing out that the proposed method can work as a framework to train for other types of no-reference LF-IQA methods.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

SA performed all the coding and tests. MF and PG worked on the idea, discussed the work. SA and MF wrote the paper.

Funding

This work was supported by the Fundação de Apoio a Pesquisa do Distrito Federal (FAP-DF), by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and by the University of Brasília (UnB).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Adhikarla, V. K., Vinkler, M., Sumin, D., Mantiuk, R., Myszkowski, K., Seidel, H.-P., et al. (2017). "Towards a quality metric for dense light fields," in Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).
- Ak, A., Ling, S., and Le Callet, P. (2020). "No-reference quality evaluation of light field content based on structural representation of the epipolar plane image," in The 1st ICME Workshop on Hyper-Realistic Multimedia for Enhanced Quality of Experience, London, UK.
- Arad, N. (1998). Grid-distortion on nonrectangular grids. *Comput. Aided Geom. Des.* 15, 475–493. doi:10.1016/S0167-8396(98)00003-X
- Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., and Kalinin, A. A. (2020). Albumentations: Fast and flexible image augmentations. *Information* 11. doi:10.3390/info11020125
- Caesar, H., Uijlings, J., and Ferrari, V. (2018). "Coco-stuff: Thing and stuff classes in context," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1209–1218. doi:10.1109/CVPR.2018.00132
- Chen, M.-J., Su, C.-C., Kwon, D.-K., Cormack, L. K., and Bovik, A. C. (2013). Full-reference quality assessment of stereopairs accounting for rivalry. *Signal Process. Image Commun.* 28, 1143–1155. doi:10.1016/j.image.2013.05.006
- Fang, Y., Wei, K., Hou, J., Wen, W., and Imamoglu, N. (2018). "Light filed image quality assessment by local and global features of epipolar plane image," in 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), 1–6. doi:10.1109/BigMM.2018.8499086
- Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems. *J. Mach. Learn. Res.* 15, 3133–3181.
- Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A, Opt. image Sci.* 4 (12), 2379–2394.
- Freitas, P. G., Alamgeer, S., Akamine, W. Y. L., and Farias, M. C. Q. (2018). "Blind image quality assessment based on multiscale salient local binary patterns," in MMSys '18: Proceedings of the 9th ACM Multimedia Systems Conference (New York, NY, USA: Association for Computing Machinery), 52–63. doi:10.1145/3204949.3204960
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (Cambridge, MA, USA: MIT Press), 2672–2680.
- Hahne, C., Lumsdaine, A., Aggoun, A., and Velisavljevic, V. (2018). Real-time refocusing using an fpga-based standard plenoptic camera. *IEEE Trans. Industrial Electron.* 65, 9757–9766. doi:10.1109/TIE.2018.2818644
- Hou, J., Chen, J., and Chau, L. (2019). Light field image compression based on bi-level wave compensation with rate-distortion optimization. *IEEE Trans. Circuits Syst. Video Technol.* 29, 517–530. doi:10.1109/TCSVT.2018.2802943
- Isola, P., Zhu, J., Zhou, T., and Efros, A. A. (2017). "Image-to-image translation with conditional adversarial networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5967–5976. doi:10.1109/CVPR.2017.632
- Jiang, G., Huang, Z., Yu, M., Xu, H., Song, Y., and Jiang, H. (2018). New quality assessment approach for dense light fields. *Proc. Volume 10817, Optoelectron. Imaging Multimedia Technol.* V 44, 1081717. doi:10.1117/12.2502277
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). "Microsoft coco: Common objects in context," in *Computer vision – eccv 2014*. Editors D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Cham: Springer International Publishing), 740–755.
- Lin, Y., and Wu, J. (2014). Quality assessment of stereoscopic 3d image compression by binocular integration behaviors. *IEEE Trans. Image Process.* 23, 1527–1542.
- Luo, Z., Zhou, W. R., Shi, L., and Chen, Z. (2019). No-reference light field image quality assessment based on micro-lens image. *ArXiv abs/1908.10087*.
- Meng, C., An, P., Huang, X., Yang, C., and Liu, D. (2020). Full reference light field image quality evaluation based on angular-spatial characteristic. *IEEE Signal Process. Lett.* 27, 525–529.
- Paudyal, P., Battisti, F., and Carli, M. (2019). Reduced reference quality assessment of light field images. *IEEE Trans. Broadcast.* 65, 152–165.
- Paudyal, P., Battisti, F., Sjöström, M., Olsson, R., and Carli, M. (2017). Toward the perceptual quality evaluation of compressed light field images. *IEEE Trans. Broadcast.* 63, 1–16. doi:10.1109/TBC.2017.2704430
- Paudyal, P., Olsson, R., Sjöström, M., Battisti, F., and Carli, M. (2016). "Smart: A light field image quality dataset," in Procs. of the ACM Multimedia Systems 2016 Conference, (MMSYS).
- Rerabek, M., and Ebrahimi, T. (2016). *New light field image dataset*.
- Rouse, D. M., and Hemami, S. S. (2009). "Natural image utility assessment using image contours," in 2009 16th IEEE International Conference on Image Processing (ICIP), 2217–2220. doi:10.1109/ICIP.2009.5413882
- Sandić-Stanković, D., Kukulj, D., and Le Callet, P. (2016). Dibr-synthesized image quality assessment based on morphological multi-scale approach. *EURASIP J. Image Video Process.* 2017, 4. doi:10.1186/s13640-016-0124-7
- Shan, L., An, P., Meng, C., Huang, X., Yang, C., and Shen, L. (2019). A no-reference image quality assessment metric by multiple characteristics of light field images. *IEEE Access* 7, 127217–127229.
- Sheikh, H. R., and Bovik, A. C. (2006). Image information and visual quality. *IEEE Trans. Image Process.* 15, 430–444. doi:10.1109/TIP.2005.859378
- Sheikh, H. R., Sabir, M. F., and Bovik, A. C. (2006). A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.* 15, 3440–3451. doi:10.1109/TIP.2006.881959
- Shi, L., Zhao, S., and Chen, Z. (2019a). "Belif: Blind quality evaluator of light field image with tensor structure variation index," in 2019 IEEE International Conference on Image Processing (ICIP), 3781–3785.
- Shi, L., Zhao, S., Zhou, W., and Chen, Z. (2018). "Perceptual evaluation of light field image," in 2018 25th IEEE International Conference on Image Processing (ICIP), 41–45. doi:10.1109/ICIP.2018.8451077
- Shi, L., Zhou, W. R., and Chen, Z. (2019b). No-reference light field image quality assessment based on spatial-angular measurement. *ArXiv abs/1908.06280*.
- V. Sze, M. Budagavi, and G. Sullivan (Editors) (2014). *High efficiency video coding (HEVC): Algorithms and architectures* (Cham: Springer). doi:10.1007/978-3-319-06895-4
- Tenenbaum, J. B., Silva, V. d., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323. doi:10.1126/science.290.5500.2319
- Tian, Y., Zeng, H., Hou, J., Chen, J., and Ma, K. (2020a). Light field image quality assessment via the light field coherence. *IEEE Trans. Image Process.* 29, 7945–7956.
- Tian, Y., Zeng, H., Hou, J., Chen, J., Zhu, J., and Ma, K. (2020b). A light field image quality assessment model based on symmetry and depth features. *IEEE Trans. Circuits Syst. Video Technol.* 31, 2046–2050. doi:10.1109/TCSVT.2020.2971256
- Tian, Y., Zeng, H., Xing, L., Chen, J., Zhu, J., and Ma, K.-K. (2018). A multi-order derivative feature-based quality assessment model for light field image. *J. Vis. Commun. Image Represent.* 57, 212–217.
- Viola, I., and Ebrahimi, T. (2018). "Valid: Visual quality assessment for light field images dataset," in 10th International Conference on Quality of Multimedia Experience (QoMEX), Sardinia, Italy, 3. doi:10.1109/QoMEX.2018.8463388
- Vu, P. V., Vu, C. T., and Chandler, D. M. (2011). "A spatiotemporal most-apparent-distortion model for video quality assessment," in 2011 18th IEEE International Conference on Image Processing, 2505–2508.
- Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Trans. image Process.* 13, 600–612.
- Wang, Z., and Li, Q. (2011). Information content weighting for perceptual image quality assessment. *IEEE Trans. Image Process.* 20, 1185–1198.
- Xiang, J., Yu, M., Chen, H., Xu, H., Song, Y., and Jiang, G. (2020). "Vblfi: Visualization-based blind light field image quality assessment," in 2020 IEEE International Conference on Multimedia and Expo (ICME), 1–6.
- Xue, W., Zhang, L., Mou, X., and Bovik, A. C. (2014). Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Trans. Image Process.* 23, 684–695. doi:10.1109/TIP.2013.2293423
- Zhou, W., and Bovik, A. C. (2002). A universal image quality index. *IEEE Signal Process. Lett.* 9, 81–84. doi:10.1109/97.995823
- [Dataset] Zhou, W., Likun, S., and Chen, Z. (2019). Tensor oriented no-reference light field image quality assessment. *IEEE Trans. Image Process.* 29, 4070. doi:10.1109/TIP.2020.2969777

Frontiers in Signal Processing

Explores the science of one-dimensional and multi-dimensional signals

An exciting journal which explores the detection, prediction, classification, and understanding of signal processing - from acoustics to biomedical signals and RADAR.

Discover the latest Research Topics

[See more →](#)

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact

