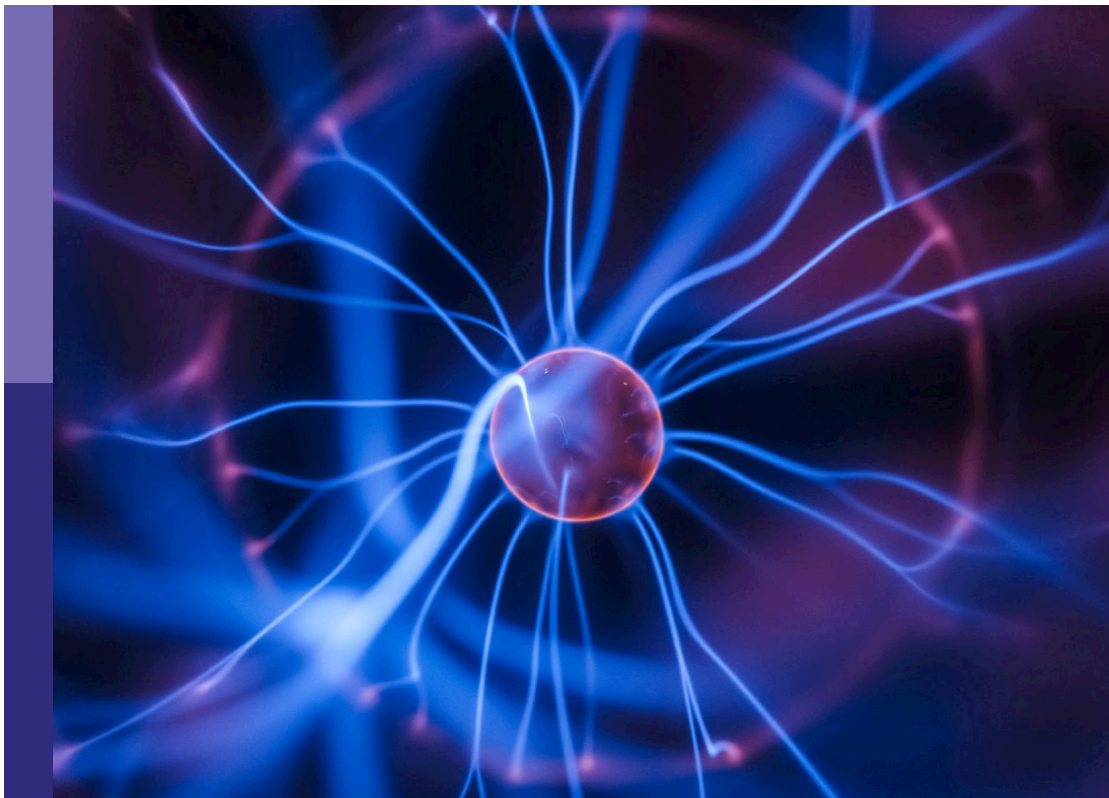
frontiers | Research Topics

Acquisition and application of multimodal sensing information

Edited by
Xukun Yin, Changhui Jiang, Huadan Zheng, Kaijie Xu and Angelo Sampaolo

Published in
Frontiers in Physics
Frontiers in Environmental Science
Frontiers in Ecology and Evolution



ISSN 1664-8714
ISBN 978-2-8325-3577-6
DOI 10.3389/978-2-8325-3577-6

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

Acquisition and application of multimodal sensing information

Topic editors

Xukun Yin — Xidian University, China
Changhui Jiang — Finnish Geospatial Research Institute, Finland
Huadan Zheng — Jinan University, China
Kaijie Xu — University of Alberta, Canada
Angelo Sampaolo — Politecnico di Bari, Italy

Citation

Yin, X., Jiang, C., Zheng, H., Xu, K., Sampaolo, A., eds. (2023). *Acquisition and application of multimodal sensing information*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3577-6

Table of contents

frontiers | Frontiers in Physics

TYPE Editorial
PUBLISHED 13 September 2023
DOI 10.3389/fphy.2023.1284176

OPEN ACCESS

EDITED AND REVIEWED BY
Lorenzo Pavesi,
University of Trento, Italy

*CORRESPONDENCE
Kaijie Xu,
kjxu@xidian.edu.cn

RECEIVED 28 August 2023
ACCEPTED 30 August 2023
PUBLISHED 13 September 2023

CITATION
Yin X, Jiang C, Zheng H, Sampaolo A and Xu K (2023), Editorial: Acquisition and application of multimodal sensing information.
Front. Phys. 11:1284176.
doi: 10.3389/fphy.2023.1284176

Editorial: Acquisition and application of multimodal sensing information

Xukun Yin[1], Changhui Jiang[2], Huadan Zheng[3], Angelo Sampaolo[4] and Kaijie Xu[5]*

[1]School of Optoelectronic Engineering, Xidian University, Xi'an, China, [2]Department of Photogrammetry and Remote Sensing, Finnish Geospatial Research Institute, Kirkkonummi, Finland, [3]Key Laboratory of Optoelectronic Information and Sensing Technologies of Guangdong Higher Education Institutes, Department of Optoelectronic Engineering, Jinan University, Guangzhou, China, [4]PolySense Lab, Dipartimento Interateneo di Fisica, University and Politecnico of Bari, Bari, Italy, [5]School of Electronic Engineering, Xidian University, Xi'an, China

KEYWORDS

optical sensor, laser spectroscopy, signal processing, LiDAR point cloud processing, 3D geospatial sensing

Editorial on the Research Topic
Acquisition and application of multimodal sensing information

With the development of advanced sensors in modern science and technology, sensors have already penetrated into such fields as industrial production, space development, ocean exploration, environmental protection, medical diagnosis, biological engineering, and even cultural relic protection [1–3]. Sensors are the basis of all data acquisition. Almost every modern project, from the exploration of the vast Universe and ocean, to the complex engineering systems, is inseparable from a variety of sensors. Sensors convert physical parameters (e.g., temperature, pressure, humidity, speed, etc.) into signals that can be measured electrically. A high-performance sensor with the advantages of high sensitivity, high resolutions, and real-time observation is desirable. In addition, the application of computers and advanced signal processing signal-processing methods makes sensor management possible ([4]; Najmeh et al. 2019). The use of sensing data from multiple sensors has been proven to be an efficient way to improve service experiences in different fields, e.g., intelligent transportation, remote sensing, smart city, and UAVs.

Recently, various optical sensors develop rapidly with the help of the advanced light source. For example, the laser spectroscopy-based trace gas sensor exhibits excellent sensitivity and selectivity compare compared with the electrochemistry gas sensor [5–7]. The laser radar or scanner are widely used in target detecting, tracking, aiming, and imaging recognition since the extremely high resolution in frequency domain, space domain and time domain. Therefore, it is meaningful to set up a Research Topic for the acquisition and application of multimodal sensing information. The Research Topic includes seven original research works, which are summarized below:

In the first article, Xiao et al. introduced an innovative approach to boost the Direction of Arrival (DOA) detection performance in the realm of array signal processing. Focusing on the optimization of the array's steering matrix, this method plays a pivotal role in accurately estimating the angles of incoming signals. The application of this enhanced steering matrix exhibits tremendous potential to

significantly elevate the accuracy and dependability of DOA detection, which bears paramount importance in various fields like radar systems, wireless communications, and sonar applications.

In the second study, Ruan et al. gained access to a valuable open-source resource that effectively extracts essential vegetation phenology information from satellite remote sensing data, through the innovative phenoC++ tool. The tool's sophisticated algorithms and cutting-edge techniques enable the processing of vast amounts of satellite data to derive meaningful insights into plant life cycle events, encompassing budding, flowering, leaf senescence, and more. This transformative tool empowers scientists to study ecological changes, understand climate impact on vegetation dynamics, and monitor ecosystem health across different regions and over extended periods.

In the third article, Guo et al. employed a rigorous quantitative assessment methodology, this study delves into the intricate relationship between human activities and nature within Chinese national parks, focusing intently on the renowned Three-River-Source National Park. The assessment encompasses multifaceted factors, including human impacts, ecological interactions, and conservation efforts, culminating in a comprehensive measure of the delicate balance and harmonious coexistence between human presence and natural ecosystems. By illuminating the level of harmony achieved, this research provides valuable guidance for policymakers, ecologists, and park management authorities, facilitating sustainable preservation and responsible utilization of natural resources.

In the fourth work in this Research Topic, Li et al. embarked on an empirical journey to analyze the performance of an innovative shipboard gravity gradient measurement system. Crucial in diverse fields of geophysics, navigation, and geodesy, the measurement of gravity gradients from a moving platform poses significant challenges. Leveraging the advances in instrumentation and data processing techniques, the paper meticulously details the design, implementation, and evaluation of the novel system, effectively demonstrating its prowess in delivering accurate and reliable gravity gradient measurements, thereby unlocking new possibilities for precise navigation and geophysical exploration on maritime ventures.

In the fifth contribution, Zhou et al. delved deep into the intricate web of factors influencing long-range single-site lightning location systems. By meticulously accounting for the propagation time delay of ground waves, the research undertakes a rigorous evaluation of existing algorithms, leading to their refinement and optimization. The revised lightning location methodology exhibits heightened precision in pinpointing lightning strikes, offering a substantial boost to lightning monitoring and early warning systems, thereby bolstering public safety and safeguarding critical infrastructure.

In the sixth contribution, Lv et al. presented a novel method that synergistically addresses the joint detection of both Direction of Departure (DOD) and Direction of Arrival (DOA) for radar targets. Leveraging signal subspace reconstruction and matching techniques, the proposed approach transcends the limitations of conventional radar systems, ensuring accurate detection and precise localization of targets. The integration of DOD and DOA detection engenders a paradigm shift in radar technology, bolstering applications such as target tracking, radar imaging, and airborne surveillance.

In the seventh contribution, Li et al. introduced an ingenious method for recognizing and combating discrete side-lobe clutter that plagues radar systems. Operating in the unforgiving environment of space, radar systems encounter a barrage of clutter signals that jeopardize accurate target detection. However, by ingeniously harnessing the power of sliding filter response loss, the proposed approach excels in discriminating and suppressing clutter, thereby elevating the radar's efficacy in space-based applications, such as satellite imaging and space object tracking.

In the eighth contribution, Zhang confronted the challenge of ensuring robust predictability in discrete event systems amid sensor attacks. Discrete event systems, governed by a sequence of events, confront disruptions when sensor data becomes compromised due to adversarial activities. In response, this research investigates resilient methods that uphold predictability and system stability, empowering industries ranging from manufacturing to transportation with fortified cyber defenses and resilient operational capacities.

In the ninth contribution, Zhang et al. introduced a dual pulse heterodyne system featuring a semiconductor optical amplifier (SOA)-based fiber ring laser. Capitalizing on the phenomenon of distributed acoustic sensing, this system continuously monitors and analyzes acoustic signals propagating along optical fibers, ushering in transformative capabilities in seismic monitoring, structural health assessment, and intrusion detection. The dual pulse heterodyne technique amplifies sensitivity, elevating the spatial resolution of acoustic measurements to unprecedented levels, thereby revolutionizing safety and security applications across diverse industries.

In the 10th contribution to this Research Topic, Li et al. reviewed unfolds the myriad applications and principles that define this cutting-edge discipline. Beyond traditional imaging techniques, polarization 3D imaging captures the spatial information alongside the intricate polarization state of light, unlocking new vistas in remote sensing, computer vision, and medical imaging domains. Unveiling its wide-ranging potential, the review serves as an authoritative guide to researchers and practitioners alike, underscoring the transformative impact of polarization 3D imaging across diverse scientific and technological frontiers.

Finally, in the 11th Original Research article, Fan et al. pioneered a novel approach that marries local neighborhood information granularity with the principles of rough set theory. In the realm of sensor networks, where the deluge of data poses computational challenges, the proposed method crafts an ingenious balance, retaining essential information while eliminating redundancy and noise. By virtue of its streamlined data representation, the approach bestows remarkable advantages in data transmission efficiency, storage optimization, and real-time processing in multifaceted applications like the Internet of Things (IoT) devices, environmental monitoring networks, and smart grids.

*In summary, this Research Topic of research papers represents a comprehensive exploration of the frontiers in various scientific fields, showcasing groundbreaking innovations and advancements. These papers contribute valuable insights and applications, propelling technology, environmental protection, and human wellbeing forward. Their publication will undoubtedly have a profound and lasting impact on academia and industry, inspiring further research endeavors in science and engineering.

Author contributions

XY: Writing–original draft. CJ: Writing–original draft. HZ: Writing–review and editing. AS: Writing–review and editing. KX: Writing–original draft, Writing–review and editing.

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

1. Mehrotra P, Chatterjee B, Sen S. EM-wave biosensors: A review of rf, microwave, mm-wave and optical sensing. *mm-Wave Opt Sensing Sensors* (2019) 19(5):1013. doi:10.3390/s19051013

2. Yin X, Wu H, Dong L, Li B, Ma W, Zhang L, et al. Ppb-level SO_2 photoacoustic sensors with a suppressed Absorption–Desorption effect by using a 7.41 μm. *External-Cavity Quan Cascade Laser Acs sensors* (2020) 5:549–56. doi:10.1021/acssensors.9b02448

3. Wu H, Dong L, Zheng H, Yu Y, Ma W, Zhang L, et al. Beat frequency quartz-enhanced photoacoustic spectroscopy for fast and calibration-free continuous trace-gas monitoring. *Nat Commun* (2017) 8:15331. doi:10.1038/ncomms15331

4. Yang R, Zhang W, Tiwari N, Yan H, Li T, Cheng H. Multimodal sensors with decoupled sensing mechanisms. *Adv Sci* (2022) 9:2202470. doi:10.1002/advs.202202470

5. Hu P, Huang H, Chen Y, Qi J, Li W, Jiang C, et al. Analyzing the angle effect of leaf reflectance measured by indoor hyperspectral light detection and ranging (LiDAR). *Remote Sensing* (2020) 12(6):919. doi:10.3390/rs12060919

6. Chen B, Li H, Zhao X, Gao M, Cheng K, Shao X, et al. Trace photoacoustic SO_2 gas sensor in SF_6 utilizing a 266 nm UV laser and an acousto-optic power stabilizer. *Opt Express* (2023) 31(4):6974–81. doi:10.1364/OE.483240

7. Yin X, Su Y, Xi T, Chen B, Zhang L, Zhang X, et al. Research progress on photoacoustic SF_6 decomposition gas sensor in gas-insulated switchgear. *J Appl Phys* (2022) 131:130701. doi:10.1063/5.0089426

8. Samadiani N, Huang G, Cai B, Luo Chi Xiang , et al. A review on automatic facial expression recognition systems assisted by multimodal sensor data. *Sensors* (2019) 19(8):1863. doi:10.3390/s19081863

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 30 November 2022
DOI 10.3389/fphy.2022.1094638

OPEN ACCESS

EDITED BY
Huadan Zheng,
Jinan University, China

REVIEWED BY
Ye Cui,
University of Alberta, Canada
Xiaoan Tang,
Hefei University of Technology, China
Hengrong Ju,
Nantong University, China

*CORRESPONDENCE
Guisheng Liao,
liaogs@xidian.edu.cn

SPECIALTY SECTION
This article was submitted to Optics and Photonics, a section of the journal Frontiers in Physics

RECEIVED 10 November 2022
ACCEPTED 21 November 2022
PUBLISHED 30 November 2022

CITATION
Xiao G and Liao G (2022), Enhancement of the DOA detection performance through optimization of the steering matrix of the array.
Front. Phys. 10:1094638.
doi: 10.3389/fphy.2022.1094638

Enhancement of the DOA detection performance through optimization of the steering matrix of the array

Guoyao Xiao[1] and Guisheng Liao[1,2]*

[1]School of Electronic Engineering, Xidian University, Xi'an, China, [2]Key Laboratory for Radar Signal Processing, Xidian University, Xi'an, China

Direction Of Arrival (DOA) of signals detection technology is an important vehicle in the field of in remote sensing, radar, wireless communication. In this study, we elaborate on an enhanced method to detect the DOA. In the developed scheme, we mainly focus on solving the steering matrix of the array which contains all the information of the signals. The iterative relation between the steering matrix and the signal vector is first established on the basis of the equation of the array output. Then, to get a more accurate of steering matrix, we construct a cost function that aims to minimize some signal subspace error. In the optimization process of the developed scheme, we also set a constraint for the steering matrix which can effectively eliminate convergence on local optimum and also reduce the number of iterations. Subsequently, the steering matrix of the array can be recovered faithfully. Finally, the DOA can be solved from the estimated steering matrix. Explicit analysis and derivation of the proposed scheme are presented.

KEYWORDS

signal processing, direction of arrival (DOA), partial noise subspace, multiple signal classification (MUSIC), sensors

Introduction

Array signal processing is an indispensable technique in signal processing with ubiquitous applications [1, 2]. The Direction Of Arrival (DOA) detection technology is a very popular topic in array signal processing [3, 4] in the field of in remote sensing, radar, wireless communication, *etc.* High-resolution subspace-based DOA methods have attracted considerable attention concerning the accurate detection of the DOA from observations of array output. The most representative high-resolution subspace-based approaches are the MUltiple SIgnal Classification (MUSIC) [5] and the Estimation Signal Parameter *via* Rotational Invariance Techniques (ESPRIT) [6]. The MUSIC method detects the DOA based on the orthogonality between the signal subspace and noise subspace [7, 8], and the ESPRIT algorithm builds on the rotational invariance of signal subspaces [9, 10]. The detection performance of this type of methods mainly depends on

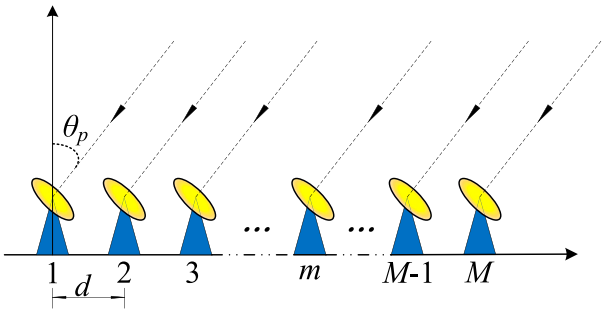


FIGURE 1
Signal model of the ULA.

the accuracy of the signal subspace. Thus, how to capture a high-precision signal subspace has always been the pursuit of these approaches [11, 12].

In this study, we design an enhanced scheme for the DOA detection. During the design process, a cost function by minimizing some signal subspace error is established to optimize the steering matrix of the array [13, 14]. In the optimization, a constraint is set to converge rapidly and eliminate converging on local optimum. Ultimately, the DOA can be solved from the obtained steering matrix of the array. We provide a series of simulations to demonstrate the superiority of the proposed method. To the best of our knowledge, the idea in this paper has not been considered in previous studies.

The organization of the paper reflects the key phases of the design process. The array signal model is first presented to formulate the problem. Next, we develop an enhanced DOA detection scheme through optimization of the steering matrix of the array. This is followed by the experimental results. Conclusions are covered in the last section.

Problem formulation

Without loss of generality, in this letter, we use a Uniform Linear Array (ULA) to illustrate the array signal model for DOA detection. We consider P narrow band noncoherent far field signals $\{s_p(t)\}_{p=1}^{P}$ [15, 16] with different DOAs impinging on the ULA which is composed of M antenna elements. Based on the above conditions, the array output is generally written in the following manner

$$\boldsymbol{X}(t) = \boldsymbol{A}\boldsymbol{s}(t) + \boldsymbol{n}(t) \tag{1}$$

where $\boldsymbol{n}(t)$ is the noise vector, and $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_p, \cdots, \boldsymbol{a}_P] \in C^{M\times P}$ contains the DOA information that is the so-called steering matrix of the array. For a given ULA, the steering vector in $\mathbf{A}$ is usually written as

$$\boldsymbol{a}_p = \exp\left[0, \cdots, j\frac{2\pi d}{\lambda}(m-1)\sin\theta_p, \cdots, j\frac{2\pi d}{\lambda}(M-1)\sin\theta_P\right]^T$$
$$m = 1, 2, \cdots, M \tag{2}$$

where T stands for the transpose operation, d denotes the spacing between adjacent antenna elements, λ and θ_p are the wavelength and the pth DOA of the signals, respectively. The array signal model is shown in Figure 1.

The high-resolution subspace-based approaches detect the DOA based on the accurate signal and the noise subspaces. Normally, the signal and the noise subspaces can be achieved through the Eigen decomposition of the array output covariance matrix [17]. Theoretically, the Eigen decomposition of the array output covariance matrix is computed in the following manner

$$\boldsymbol{R}_X = E\{\boldsymbol{X}\boldsymbol{X}^H\} = \boldsymbol{A}\boldsymbol{R}_s\boldsymbol{A}^H + \boldsymbol{\sigma^2}\mathbf{I} \tag{3}$$

where H stands for the complex conjugate transpose, R_s is the correlation matrix of the signal vector, and σ^2 means the noise power. The eigenvalue decomposition of the array output covariance matrix is expressed as

$$\boldsymbol{R}_X = \boldsymbol{U}_s\boldsymbol{A}_s\boldsymbol{U}_s^H + \sigma^2\boldsymbol{U}_n\boldsymbol{U}_s^H \tag{4}$$

where $\boldsymbol{A}_s$ is a diagonal matrix composed of P signal eigenvalues, $\boldsymbol{U}_s$ and $\boldsymbol{U}_n$ are respectively the signal and noise subspaces determined by the distribution of eigenvalues. Then, the DOA of the signals can be solved with the high-resolution subspace-based approaches.

Most of the existing subspace-based methods enhance the DOA detection performance through solving or optimizing an accurate signal subspace, which has always been a hot topic for scholars [18].

Optimization of the signal subspace

Based on the above array signal model, in this section, we develop a novel optimization scheme of the signal subspace. Mathematically, the detection of the DOA can be considered as the solution of the steering matrix of the array, and the corresponding problem is formulated as

$$\hat{\boldsymbol{A}} = \arg\min_{\boldsymbol{A}} \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{s}\|_2^2 \tag{5}$$

where $\|\bullet\|_2^2$ denotes the 2-norm. Normally, if we fix one of the variables, the other one can be solved through the method of least squares (by minimizing the standard squared error), which is expressed in the following manner

$$\hat{\boldsymbol{A}} = \boldsymbol{X}\hat{\boldsymbol{s}}\left[\hat{\boldsymbol{s}}\hat{\boldsymbol{s}}^T\right]^{-1} \quad (a)$$
$$\hat{\boldsymbol{s}} = \left[\hat{\boldsymbol{A}}^T\hat{\boldsymbol{A}}\right]^{-1}\hat{\boldsymbol{A}}\boldsymbol{X} \quad (b) \tag{6}$$

It seems that the steering matrix of the array can be obtained in the above way (iteratively update the steering matrix and the

signal vector). However, the array output contains not only signals but also noises, minimizing the standard squared error of (5) to produce the steering matrix is probably not desirable. To capture an accurate steering matrix so as to solve the DOA of the signals, we carry out the following design.

Assume that the steering matrix of the array computed by minimizing some cost function during the iteration is $\tilde{\mathbf{A}}_t$, where t denotes the index of the successive iteration. Then, we build up a signal subspace in the following form

$$\tilde{\boldsymbol{U}}_t = \tilde{\boldsymbol{A}}_t \left[\tilde{\boldsymbol{A}}_t^H \tilde{\boldsymbol{A}}_t \right]^{-\frac{1}{2}} \tag{7}$$

and the projection matrix [19, 20] of the signal subspace is defined as

$$\boldsymbol{Q}_t = \tilde{\boldsymbol{A}}_t \left[\tilde{\boldsymbol{A}}_t^H \tilde{\boldsymbol{A}}_t \right]^{-1} \tilde{\boldsymbol{A}}_t^H \tag{8}$$

Ideally, this reconstructed signal subspace and the estimated signal subspace through the covariance matrix of the array output should be equal. Thus, from this point of view, we establish such a cost function

$$J = \tilde{\boldsymbol{A}} \left[\tilde{\boldsymbol{A}}^H \tilde{\boldsymbol{A}} \right]^{-1} \tilde{\boldsymbol{A}}^H - \hat{\boldsymbol{U}}_s \hat{\boldsymbol{U}}_s^H \tag{9}$$

and combine it with (5) to optimize the signal subspace to determine the DOA.

Proceeding with more details, the developed scheme starts by computing the covariance matrix of the array output. Then, a set of initial DOAs is estimated using some classical approaches (say, MUSIC, ESPRIT, *etc.*) to form an initial steering matrix of the array $\tilde{\mathbf{A}}_0$ to promote the implementation of the algorithm. Subsequently, the steering matrix of the array and the signal vector update according to (6), and then minimize the constructed cost function. The entire process is repeated until there are no significant changes to the entries of the cost function reported in the two successive iterations of the method. Finally, the DOA can be solved from the resulting steering matrix of the array.

In order to avoid the algorithm falling into a local optimum, we set a constraint for the steering matrix of the array. Let $U(\theta_{p0}, \delta)$ denote the δ-neighborhood of θ_{p0} (the pth initial DOA), which is expressed in the following form

$$U\left(\theta_{p0}, \delta\right) = \left\{ \chi \,\middle|\, \theta_{p0} - \delta < \chi < \theta_{p0} + \delta \right\} \tag{10}$$

That is, during the iteration process, we limit the steering matrix of the array to a certain range by keeping the DOA to be detected to a certain range, which can effectively eliminate convergence on local optimum and also reduce the number of iterations of the algorithm. Obviously, this strategy can not only ensure the detection accuracy of DOA, but also accelerate the convergence speed of the method.

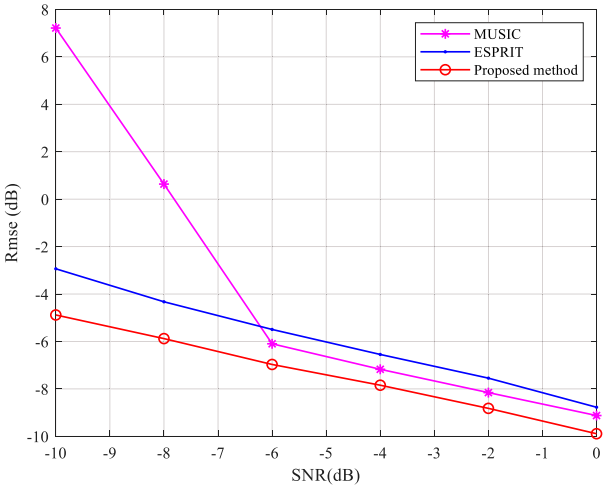


FIGURE 2
RMSE *versus* SNR.

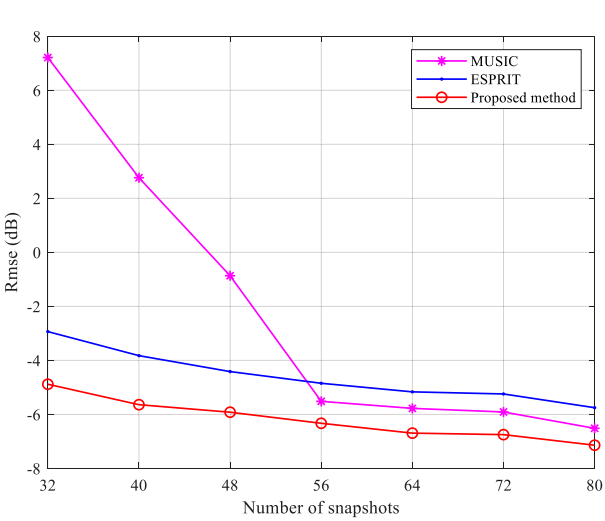


FIGURE 3
RMSE *versus* number of snapshots.

Experimental studies

We offer a series of simulations to demonstrate the Root-Mean-Square Error (RMSE) [20] performance of the approach compared with the MUSIC and the ESPRIT methods. In all simulations, a 15 elements ULA with a relative interelement spacing of $d = \lambda/2$ is used, and four narrowband signals with the DOAs [5°, 10°, 15°, 30°] impinge on the array. In this letter, the RMSE is defined as [21, 22].

$$10\log 10\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left\{\frac{1}{P}\sum_{p=1}^{P}\left[\hat{\theta}_p(n)-\theta_p\right]^2\right\}}\quad(\mathrm{dB})\qquad(11)$$

where N denotes the independent trials, and in the following simulations we set it as 200; $\hat{\theta}_p$ is the pth estimated DOA of the θ_p.

First, we test the RMSE performance of the methods *versus* the SNR, where the number of snapshots is fixed at 32, and the SNR varies from −10 to 0 with two intervals. The means of the simulation results are plotted in Figure 2. It is apparent that the performance of DOA detection is enhanced compared with the MUSIC and the ESPRIT methods with the developed method, and the developed method is also not very sensitive to the low SNR.

After that, we test the RMSE performance of the methods *versus* the number of snapshots. In the simulation, the SNR is fixed as −10 dB, and the number of snapshots varies from 32 to 80 with eight intervals. Figure 3 shows the simulation results. It is noticeable that the proposed method outperforms the MUSIC method and becomes insensitive to the changes of the number of snapshots. As previously mentioned in this letter, the detection performance of these subspace-based methods mainly depends on the accuracy of the signal subspace. The developed scheme optimizes the signal subspace through constructing a cost function and determining an optimal solution of the steering matrix of the array so as to solve the DOA. During this process, the signal subspace is optimized and the performance of the DOA detection becomes enhanced.

Conclusion

A scheme for DOA detection is put forward in this paper. The proposed scheme mainly involves the construction of the cost function of the steering matrix and the design of the steering matrix optimization. A constraint for the steering matrix is also set to make the method converge fast and eliminate the convergence on local optimum. The DOA is solved from the resulting steering matrix of the array. The simulation results indicate that the developed scheme achieves much better estimation performance than the traditional algorithms.

Hence, this paper proposes a fresh way to detect the DOA and also poses a problem of reducing the complexity of the method, as the developed scheme includes a series of iterations. Furthermore, hardware design and consideration of a real noise environment would also be interesting topics for research.

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

All the authors made significant contributions to the work. The idea was proposed by GL; GX simulated the algorithm, analysed the data designed the experiments and polish the English, and wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported the National Natural Science Foundation of China under Grant 61971349.

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

1. Ahmad Z. Fundamentals of narrowband array signal processing. In: W Cao Q Zhang, editors. *Adaptive filtering-recent advances and practical implementation IntechOpen*. Germany: Researchgate (2021). doi:10.5772/intechopen.98702

2. Wang M, Gao F, Jin S, Lin H. An overview of enhanced massive MIMO with array signal processing techniques. *IEEE J Sel Top Signal Process* (2019) 13(5): 886–901. doi:10.1109/jstsp.2019.2934931

3. Liu H, Lu H, Lin J, Han F, Spencer BF, Cui J, et al. Penetration properties of ground penetrating radar waves through rebar grids. *IEEE Geosci Remote Sensing Lett* (2021) 18(7):1199–203. doi:10.1109/lgrs.2020.2995670

4. Duplouy J, Morlaas C, Aubert H, Potier P, Pouliguen P, Djoma C. Wideband and reconfigurable vector antenna using radiation pattern diversity for 3-D direction-of-arrival estimation. *IEEE Trans Antennas Propagat* (2019) 67(6): 3586–96. doi:10.1109/tap.2019.2905729

5. Schmidt RO. Multiple emitter location and signal parameter estimation. *IEEE Trans Antennas Propagat* (1986) 34(3):276–80. doi:10.1109/tap.1986.1143830

6. Roy R, Kailath T. ESPRIT-estimation of signal parameters via rotational invariance techniques. *IEEE Trans Acoust Speech, Signal Process* (1989) 37(7): 984–95. doi:10.1109/29.32276

7. Wang Z, Yang Z, Wu S, Li H, Tian S, Chen X. An improved multiple signal classification for nonuniform sampling in blade tip timing. *IEEE Trans Instrum Meas* (2020) 69(10):7941–52. doi:10.1109/tim.2020.2980912

8. Yang X, Wang K, Zhou P, Xu L, Liu J, Sun P, et al. Ameliorated-multiple signal classification (Am-MUSIC) for damage imaging using a sparse sensor network. *Mech Syst Signal Process* (2022) 163:108154–16. doi:10.1016/j.ymssp.2021.108154

9. Liu M, Cao H, Wu Y. Improved subspace-based method for 2-D DOA estimation with L-shaped array. *Electron Lett* (2020) 56(8):402–5. doi:10.1049/el.2019.4235

10. Xu KJ, Quan YH, Bie BW, Xing MD, Nie WK, Hanyu E. Fast direction of arrival estimation for uniform circular arrays with a virtual signal subspace. *IEEE Trans Aerosp Electron Syst* (2021) 57(3):1731–41. doi:10.1109/taes.2021.3050667

11. Xu KJ, Pedrycz W, Li ZW, Nie WK. High-accuracy signal subspace separation algorithm based on Gaussian kernel soft partition. *IEEE Trans Ind Electron* (2019) 66(1):491–9. doi:10.1109/tie.2018.2823666

12. Castanheira D, Gameiro A. Low Complexity and high-resolution line spectral estimation using cyclic minimization. *IEEE Trans Signal Process* (2019) 67(24):6285–300. doi:10.1109/tsp.2019.2953582

13. Wang X, Meng D, Huang M, Wan L. Reweighted regularized sparse recovery for DOA estimation with unknown mutual coupling. *IEEE Commun Lett* (2019) 23(2):290–3. doi:10.1109/lcomm.2018.2884457

14. Hu W, Wang Q. DOA estimation for UCA in the presence of mutual coupling via error model equivalence. *IEEE Wireless Commun Lett* (2020) 9(1):121–4. doi:10.1109/lwc.2019.2944816

15. Qin L, Wu S, Zhang C, Li X. Narrowband and full-angle refractive index sensor based on a planar multilayer structure. *IEEE Sensors J* (2019) 19(8):2924–30. doi:10.1109/jsen.2019.2890863

16. Ioushua SS, Yair O, Cohen D, Eldar YC. CaSCADE: Compressed carrier and DOA estimation. *IEEE Trans Signal Process* (2017) 65(10):2645–58. doi:10.1109/tsp.2017.2664054

17. Xu KJ, Nie WK, Feng DZ, Chen XJ, Fang DY. A multi-direction virtual array transformation algorithm for 2D DOA estimation. *Signal Process.* (2016) 125:122–33. doi:10.1016/j.sigpro.2016.01.011

18. Xu KJ, Xing M, Zhang R, Hanyu E, Sha MH, Nie WK, et al. High-accuracy DOA estimation algorithm at low SNR through exploiting a supervised index. *IEEE Trans Aerosp Electron Syst* (2022) 58(4):3658–65. doi:10.1109/taes.2022.3144121

19. Al-Sadoon MAG, Al-Nedawe BM, Bin-Melha M, Abd-Alhammed RA. The selected samples effect on the projection matrix to estimate the direction of arrival," in Proceedings of the 2019 UK/China Emerging Technologies. Glasgow, UK, 21-22 August 2019, IEEE (2019). p. 1–5.

20. Ferreol A, Larzabal P, Viberg M. Performance prediction of maximum-likelihood direction-of-arrival estimation in the presence of modeling errors. *IEEE Trans Signal Process* (2008) 56(10):4785–93. doi:10.1109/tsp.2008.921794

21. Nie WK, Xu KJ, Feng DZ, Wu CQ, Hou AQ, Tin XY. A fast algorithm for 2D DOA estimation using an omnidirectional sensor array. *Sensors* (2017) 17(3):515–29. doi:10.3390/s17030515

22. Varanasi V, Agarwal A, Hegde RM. Near-field acoustic source localization using spherical harmonic features. *Ieee/acm Trans Audio Speech Lang Process* (2019) 27(12):2054–66. doi:10.1109/taslp.2019.2939782

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 12 January 2023
DOI 10.3389/fphy.2022.1121633

OPEN ACCESS

EDITED BY
Huadan Zheng,
Jinan University, China

REVIEWED BY
Cai Tijing,
Southeast University, China
Ruihang Yu,
National University of Defense Technology, China

*CORRESPONDENCE
Jiahong Zhang,
agrs_zhang@163.com
Guangjing Xu,
bdxgj@163.com

SPECIALTY SECTION
This article was submitted to Optics and Photonics, a section of the journal Frontiers in Physics

RECEIVED 12 December 2022
ACCEPTED 29 December 2022
PUBLISHED 12 January 2023

CITATION
Li R, Li D, Shu Q, Fan Z, Lu K, Zhou J, Zhang J and Xu G (2023), Experimental analysis of the performance of a new shipboard gravity gradient measurement system.
Front. Phys. 10:1121633.
doi: 10.3389/fphy.2022.1121633

Experimental analysis of the performance of a new shipboard gravity gradient measurement system

Rui Li[1], Da Li[2], Qing Shu[1], Zhenyu Fan[1], Kai Lu[3], Jianxin Zhou[1], Jiahong Zhang[1]* and Guangjing Xu[1]*

[1]China Aero Geophysical Survey and Remote Sensing Center for Natural Resources, Beijing, China, [2]Tianjin Navigation Instrument Research Institute, Tianjin, China, [3]Qingdao Institute of Marine Geology, Qingdao, China

The gravity gradient tensor, which has a higher resolution than gravity, is used in a variety of fields, including the discovery of energy resources, auxiliary navigation, and national defense building. Our team has achieved significant advancements in various essential technologies, such as high-resolution accelerometers, and has constructed China's first self-controllable shipboard gravity gradient measurement system. In the laboratory, accuracy is determined using the mass gravitation technique, static test accuracy of T_{uv} and T_{xy} is 7.22 E and 3.58 E, while dynamic test accuracy of T_{uv} and T_{xy} is 9.09 E and 4.16 E. For outfield shipborne test measurement, the internal accord accuracy of T_{uv} and T_{xy} of the repeat line is 28.2E@750m and 28.8E@750m, and that of the intersection point is 28.2E@750m and 26.8E@750m. The performance of the system is completely validated by dynamic and static testing, laying the groundwork for the practical implementation of gravity gradient technology.

KEYWORDS

shipborne, quartz flexible rotational accelerometer, gravity gradiometer, gravity gradient measurement, internal accord accuracy

1 Introduction

The gravity gradient is the second-order derivative of the gravity potential and indicates the rate of change of the gravity field vector in three-dimensional space [1], which can better convey the comprehensive information of its field source [2]. The history of gravity gradient measurement dates back to 1886, when Baron von Eötvös devised the gravity gradiometer based on the principle of torsional scale balancing [3], and gravity gradient measurement became the earliest means of oil exploration. Gravity gradient technology is extensively employed in auxiliary navigation, geology and geophysics, geodesy, and other fundamental research domains [4–10] due to technological advancements.

The gravity gradient system based on quartz flexible rotational accelerometer technology is presently the only commercially viable gravity gradient measuring system, such as the Air-FTG and Marine-FTG developed by Bell Geospace (now Lockheed Martin, United States), the Falcon system by BHP (Australia), and the FTGeX by ARKEX (UK). All are rotating accelerometers [11, 12], with an accuracy of around 10 E and varying resolutions depending on the mounting platform (such as airships, aircraft, submarines, etc.). For more than 3 decades, gravity gradiometry has been acquired offshore using marine vessels and onshore using fixed wing aircraft. The data acquired on a slow moving, large ship would be higher quality than data acquired at higher speed on a fixed wing airplane [13, 14].

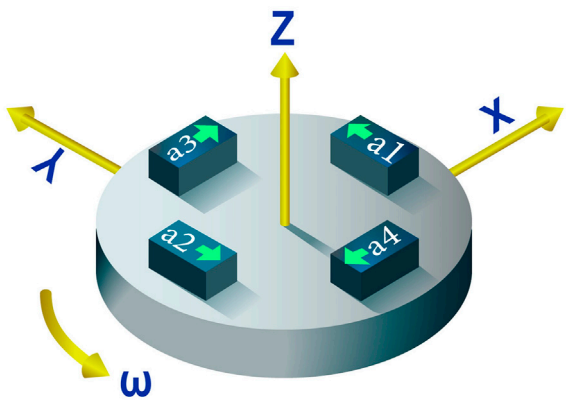


FIGURE.1
Gravity gradient horizontal tensor measurement schematic diagram.

Some gravity gradiometers with a higher degree of precision, such as the electrostatic levitation gravity gradiometer [15, 16], the superconducting gravity gradiometer [17], and the gravity gradient system based on the principle of atomic interference [18], have also made significant strides and are approaching practicality. For instance, the cold atomic gravity gradiometer developed by the University of Birmingham research team in the United Kingdom has been used to detect underground cavities [19]. After measuring for more than 10 min, the sensitivity can reach 20 E (1 E = 10^{-9}/s^2). The units of the gravity gradient are usually in Eotvos E), where 1 E corresponds to a gravitational difference of 10^{-10} g between two points separated by 1 m.

The China Aero Geophysical Survey and Remote Sensing Center for Natural Resources, and Tianjin Institute of Nautical Instruments have made significant breakthroughs in various key technologies since the beginning of the 11th Five-Year Plan in year 2006. The resolution of the quartz flexible accelerometer has been increased from 1×10^{-5} g to 1×10^{-9} g [20, 21], and China's first self-controllable gravity gradient measurement system (GGMS) for a shipborne mobile platform has been developed and shipborne tested. Here we briefly present this new GGMS, and the results of laboratory and shipboard tests conducted to assess system performance.

1.1 Principle of gravity gradient measurements

Gravity gradiometers can measure the spatial rate of change in gravitational acceleration. The gravitational field vector is the gradient of the gravitational potential U. In the Cartesian coordinate system, the gravitational gradient tensor T can be defined as follows:

$$T = \nabla\nabla U = \begin{bmatrix} \frac{\partial^2 U}{\partial x^2} & \frac{\partial^2 U}{\partial x \partial y} & \frac{\partial^2 U}{\partial x \partial z} \\ \frac{\partial^2 U}{\partial y \partial x} & \frac{\partial^2 U}{\partial y^2} & \frac{\partial^2 U}{\partial y \partial z} \\ \frac{\partial^2 U}{\partial z \partial x} & \frac{\partial^2 U}{\partial z \partial y} & \frac{\partial^2 U}{\partial z^2} \end{bmatrix} = \begin{bmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{yx} & T_{yy} & T_{yz} \\ T_{zx} & T_{zy} & T_{zz} \end{bmatrix} \quad (1)$$

Only five of the nine components of the gravity gradient are independent, as the external gravity field of the earth satisfies the Laplace equation and the gradient tensor is symmetric, whereby $T_{xy} = T_{yx}$, $T_{xz} = T_{zx}$, $T_{yz} = T_{zy}$, and $T_{xx} + T_{yy} + T_{zz} = 0$.

The main gravity gradiometers that are currently available for mobile platform measurements can be divided into full-tensor gravity gradiometers and partial-tensor gravity gradiometers based on their measurement elements. The core measurement component of a gravity gradiometer is the gravity gradient instrument (GGI). The Air-FTG, which has three sets of GGIs mounted on an inertially stabilized platform, is a typical instrument for full-tensor measurements. The partial-tensor gravity gradiometer, on the other hand, measures either a partial component or a combination of the partial components of the gravity gradient, with the Falcon system being a typical partial-tensor instrument. The principle of our gravity gradiometer is similar to that of the Falcon system, which measures the horizontal components of the gravity gradient tensor (i.e., T_{yy} - T_{xx} and T_{xy}). The x-, y-, and z-axes of the gravity gradient measurement coordinate system are defined to correspond to the local geographic coordinate system's E (eastward), N (northward), and U (skyward) coordinates, respectively, whereby T_{yy}, T_{xx}, an T_{xy} ($T_{xy} = T_{yx}$) correspond to T_{NN}, T_{EE}, and T_{NE}, respectively.

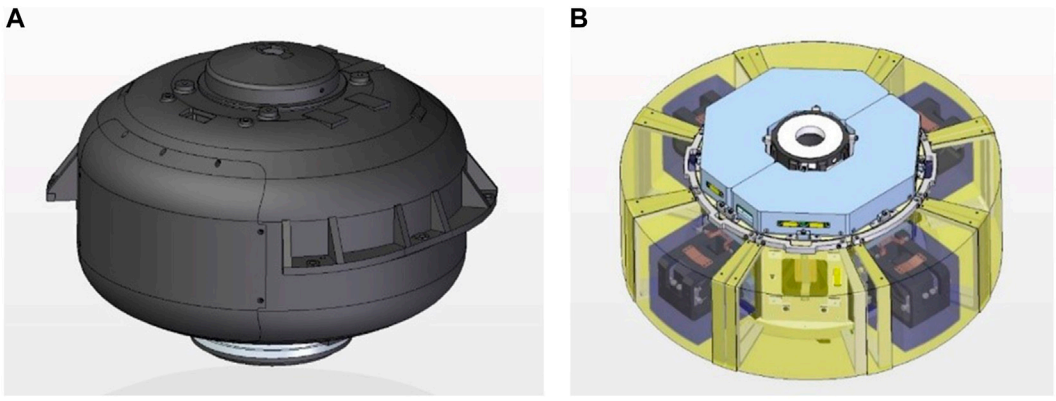


FIGURE 2
GGI **(A)** and rotating body parts **(B)**.

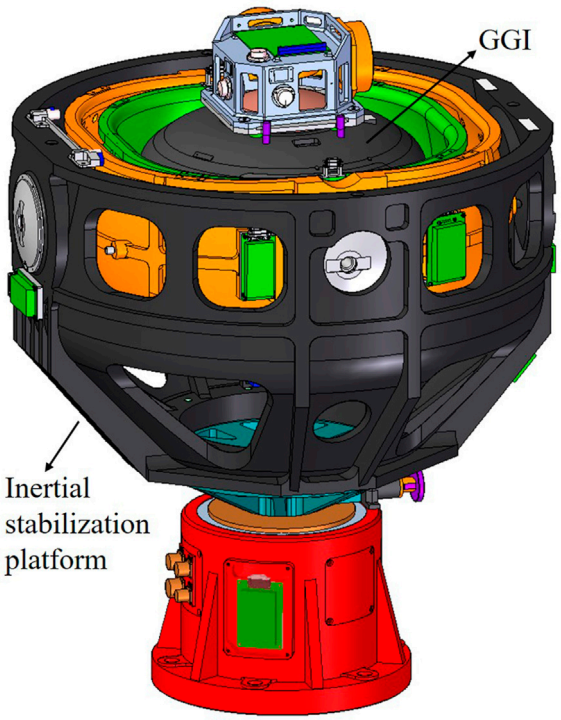


FIGURE 3
Installation structure diagram of inertial stabilization platform and GGI.

The principle of a GGI measurement is illustrated in Figure 1. Four accelerometers, a1–a4, are evenly distributed on a round plate of radius R that is centered at point O and rotates at a specified angular rate. The input axis of the accelerometer is along the tangential direction of the rotation plane and perpendicular to the rotation axis.

The output of the accelerometers is related to the gravitational gradient as follows [22]:

$$(a_1 + a_2) - (a_3 + a_4) = 2R\left(T_{yy} - T_{xx}\right)\sin 2\omega t + 4RT_{xy}\cos 2\omega t \quad (2)$$

where ω is the rotation frequency and t is the measurement time. The sum and difference of the accelerometer output signals expressed in Eq. 2 are demodulated using twice the rotation frequency (i.e., 2ω) signal to extract the gravity gradient signals ($T_{yy} - T_{xx}$ and T_{xy}).

2 Gravity gradiometer system and laboratory performance tests

2.1 System components

The new shipborne GGMS includes a GGI, inertial stabilization platform, buffer damping unit, temperature control unit, and power supply unit.

The shipboard GGI is the core component of the gravity gradiometer and consists of three major parts: the rotating body, rotating shaft system, and rotating support frame. The high-resolution accelerometer, accelerometer servo circuit board, gravity gradient measurement circuit board, and accelerometer temperature control board are all installed on the rotating body. The shape of the gravity gradient sensor is shown in Figure 2. The rotating body is the mounting base for the accelerometer. The four accelerometers are evenly and symmetrically arranged along the circumference of the rotating body, and centered on the rotation axis.

The inertial stabilization platform (Figure 3) is one of the key components of the system, and its main functions are to house the GGI, isolate the carrier angular motion and track the local geographic coordinate system, and provide a dynamic environment for the gravity gradient sensor to meet the gravity gradiometer requirements. The principal scheme of the stabilized platform adopts the mechanical arrangement scheme of the classical three-loop semi-analytic inertial navigation system. Here the output information of the platform accelerometer, in combination with the initial velocity and position binding information of the carrier, are integrated to solve for the carrier velocity and position, and further solve for the corrected

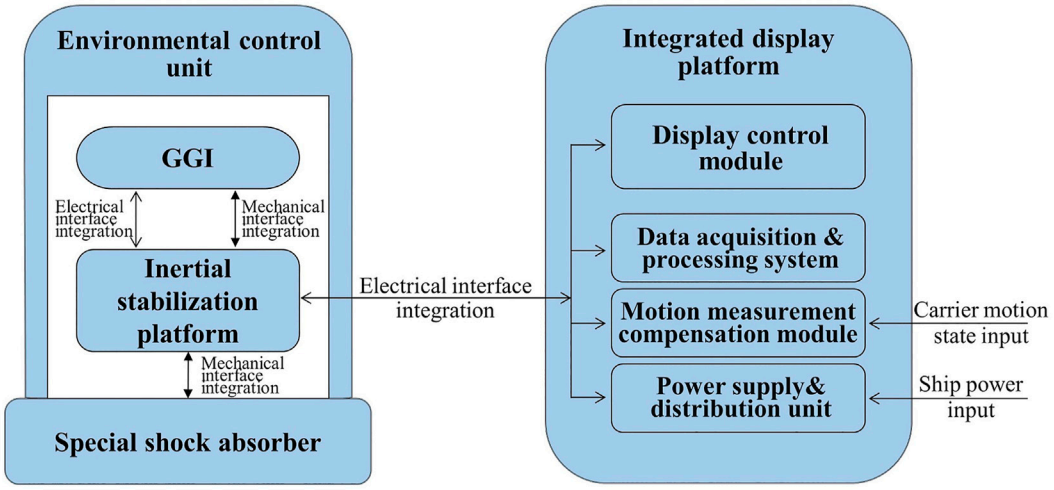


FIGURE 4
Integration scheme of shipboard GGMS.

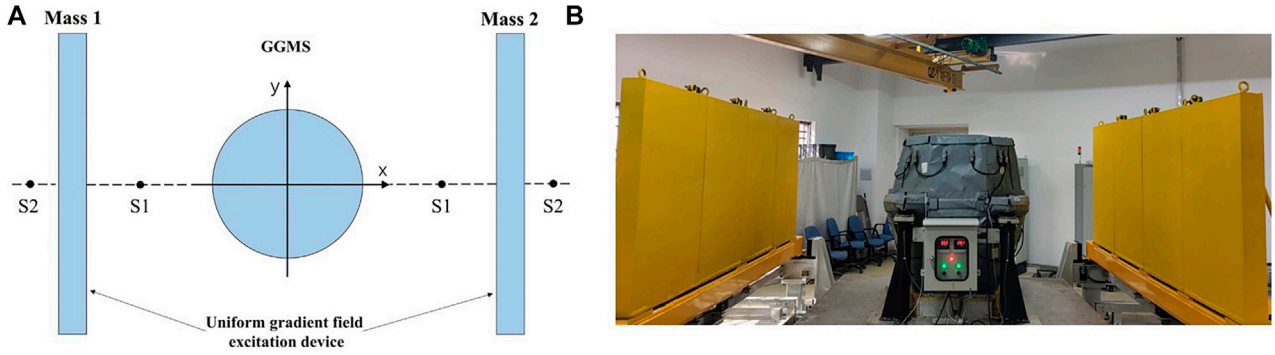


FIGURE 5
Schematic diagram of static accuracy test of GGMS **(A)** and test scene **(B)**.

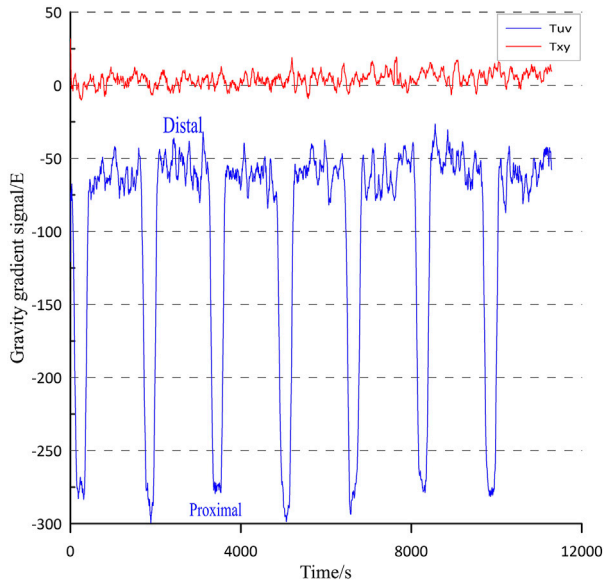


FIGURE 6
Static accuracy test curve of GGMS.

angular velocity, which is required for the stabilized platform to track the geographic coordinate system. This corrected angular velocity is then used to correct the output data of the corresponding fiber optic gyroscope on the platform, and is also applied to the platform through the stabilization loop to make the platform coordinate system track the geographic coordinate system, without accounting for the stabilization loop error, such that the local horizontal north-pointing inertial platform is then realized.

The GGI, inertial stabilization platform, buffer shock absorption unit, temperature control unit and power distribution unit, gravity gradient data acquisition and processing system, and motion compensation module with mechanical and electrical interfaces are all connected in the new shipboard GGMS to ensure effective communication and connectivity among the system components, and realize the systematic gravity gradient measurement operation function (Figure 4).

2.2 Laboratory-based static performance testing

A laboratory-based static accuracy test was the first performance test conducted after assembling the shipboard GGMS. The static

TABLE 1 Statistics of static accuracy test data of GGMS(T_{uv}).

	Proximal	Distal	Measurement difference	Measurement error
1	−274.40	−60.68	213.72	−10.28
2	−286.20	−53.29	232.91	8.91
3	−274.84	−59.08	215.76	−8.24
4	−293.03	−61.37	231.66	7.66
5	−277.78	−55.33	222.45	−1.55
6	−274.51	−53.77	220.74	−3.26
7	−278.29	−60.66	217.63	−6.37
Mean	−279.86	−57.74	222.12	—
Standard deviation	7.12	3.50	7.53	—
RMS	—	—	—	7.22

TABLE 2 Statistics of static accuracy test data of GGMS(T_{xy}).

	Proximal	Distal	Measurement difference	Measurement error
1	−6.87	−2.18	4.69	4.69
2	−1.42	4.08	5.50	5.50
3	0.11	1.47	1.36	1.36
4	1.83	3.61	1.78	1.78
5	5.49	6.28	0.79	0.79
6	10.21	5.84	−4.37	−4.37
7	8.68	5.07	−3.61	−3.61
Mean	2.58	3.45	0.88	—
Standard deviation	6.00	2.95	3.76	—
RMS	—	—	—	3.58

FIGURE 7
Dynamic accuracy test on swaying platform.

TABLE 3 Dynamic test sway spectrum of GGMS.

	Roll	Pitch	Yaw
Amplitude (°)	6	1.3	0.8
Cycle time (s)	8	5	5

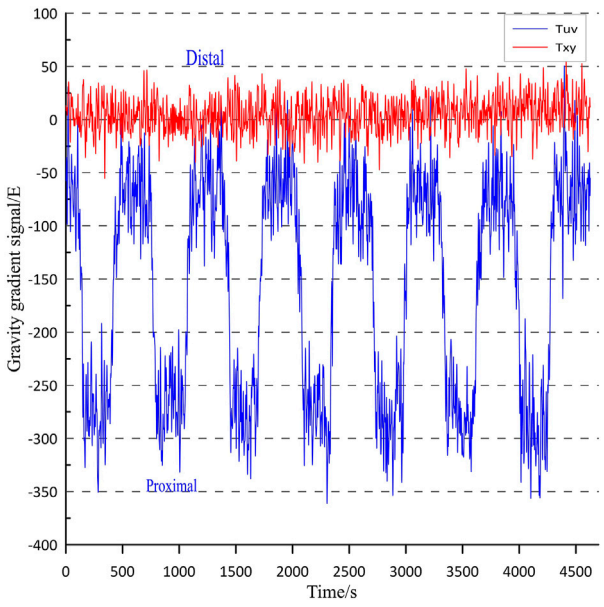


FIGURE 8
Dynamic accuracy test curve of GGMS.

accuracy test employs the mass gravitational excitation method. A mass gravitational uniform gradient field excitation device is used to conduct the test. The mass walls on both sides of the gravity gradiometer are high-density bodies that can generate a uniform gravitational gradient field in the space where the gradiometer is located.

A schematic diagram of the laboratory-based static test is shown in Figure 5. The static working GGMS is in the center, with the two high density mass walls that generate the uniform gravitational gradient excitation positioned on the left and right sides of the GGMS. The two mass walls move along the dotted line to either the center or both sides simultaneously during the test, such that different gravitational gradients can be excited at different positions relative to the spatial position of the GGMS. The static measurement accuracy of the GGMS can be determined by comparing the difference between the theoretical gravitational gradient excitation at two positions and the measured gradient difference of the GGMS.

The gravitational gradient excitation on the gradiometer, which is produced by two rectangular mass walls at a distance S2 from the distal end and a distance S1 from the proximal end of the gradiometer, can be calculated theoretically [23] by defining $T_{uv} = (T_{yy}-T_{xx})/2$, such that the variation of the T_{uv} signal is 224 E and the variation of the T_{xy} signal is 0 E. The mass walls were moved repeatedly (seven times) between the distal and proximal ends during the test experiment. The output signal of the GGMS was recorded continuously for 3 and 20 min at the proximal and distal ends, respectively. The average value within

TABLE 4 Statistics of dynamic accuracy test data of GGMS(T_{uv}).

	Proximal	Distal	Measurement difference	Measurement error
1	−282.20	−65.68	216.52	−7.48
2	−259.10	−53.28	205.82	−18.18
3	−285.92	−61.03	224.89	0.89
4	−274.51	−57.66	216.85	−7.15
5	−294.73	−68.50	226.23	2.23
6	−284.60	−72.17	212.43	−11.57
7	−293.07	−68.09	224.98	0.98
Mean	−282.02	−63.77	218.25	—
Standard deviation	12.16	6.71	7.60	—
RMS	—	—	—	9.09

TABLE 5 Statistics of dynamic accuracy test data of GGMS(T_{xy}).

	Proximal	Distal	Measurement difference	Measurement error
1	3.43	1.79	−1.64	−1.64
2	3.74	−0.75	−4.49	−4.49
3	1.14	7.43	6.29	6.29
4	0.32	2.29	1.97	1.97
5	3.22	5.31	2.09	2.09
6	7.51	10.98	3.47	3.47
7	8.63	14.81	6.18	6.18
Mean	4.00	5.98	1.98	—
Standard deviation	3.07	5.51	3.95	—
RMS	—	—	—	4.16

100 s of system stabilization was taken as the signal output value, and the repeatability of the output signal and accuracy of the difference between the output signals at two positions were counted for each position multiple times to determine the static measurement accuracy of the GGMS. Figure 6 shows the static accuracy test curve of the GGMS, and Tables 1, 2 list the static accuracy test data statistics. The T_{uv} and T_{xy} measurement accuracies are 7.22 E and 3.58 E, respectively. The results indicate that the static measurement accuracy of the system is better than 7.22 E.

2.3 Laboratory-based dynamic accuracy test

The laboratory-based dynamic accuracy test of the GGMS was similar to the static test method. However, the GGMS was not stationary during the dynamic accuracy test, but was instead placed on a motion simulation platform to simulate the swaying motion of a ship. The test site is shown in Figure 7.

The moving process and resting position of the mass body were the same as those in the laboratory-based static test, and the stationary measurement times at the distal and proximal ends were both 6 min. A motion simulation table simulated the swaying motion for both large and medium-sized ships under four sea-state levels during the test. The swaying spectrum is shown in Table 3. Figure 8 shows the dynamic accuracy test curve of the GGMS, and Tables 4, 5 list the data statistics of the dynamic accuracy test. The measurement accuracy of the GGMS is better than 9.09 E for external gradient excitation changes.

3 Shipborne test

After laboratory tests were completed, we then placed the shipboard GGMS into a removable container to form a mobile laboratory for easy transportation and installation. The temperature and humidity in the mobile laboratory were controlled within a certain range to provide a stable external environment for the gravity gradiometer. We fixed the mobile laboratory on the deck of the

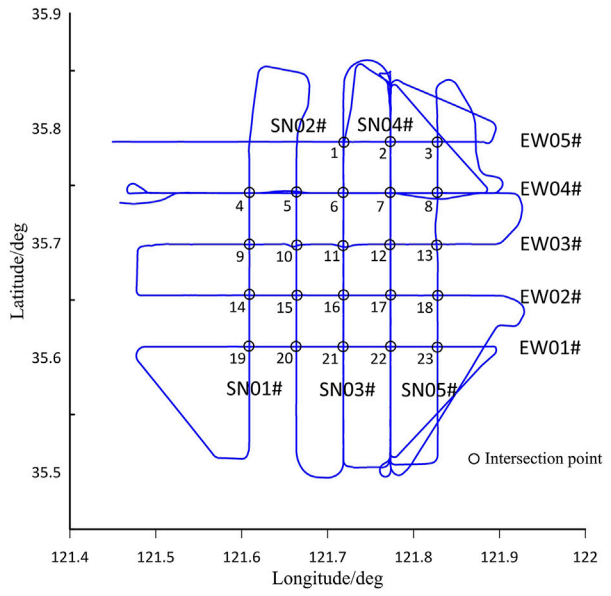


FIGURE 9
Location map of sea trial survey network and intersection points.

ship, and completed the first shipboard sea test in the offshore waters of the Yellow Sea of China.

The survey network consisted of a 5 line × 5 line grid, with one north–south line (SN04#) selected for three round-trip repeat measurements, one north–south line (SN03#) for one round-trip repeat measurement, and one east–west line (EW04#) for one round-trip repeat measurement. The line spacing was 5 km, with the north–south and east–west lines being 30 and 40 km in length, respectively. The ship speed was 10 knots (1 knot ≈1.852 km/h) during the test. The locations of the test survey network and intersection points are shown in Figure 9.

3.1 Data processing

The obtained test data included the original output data of the gravity gradient sensor, which were acquired at a continuous sampling rate of 800-Hz; the output motion information data of the inertial measurement unit on the stable platform, which were acquired at a continuous sampling rate of 800-Hz; the attitude data of the ship, which were acquired at a continuous sampling rate of 200-Hz using external fiber-optic rosettes; and GPS position data, which were acquired at a continuous sampling rate of 100-Hz. The walk-away measurement mode was adopted during the shipborne test, with the part of the navigation path that possessed a uniform speed and straight heading selected as an effective measurement line.

The output motion information data of the inertial measurement component on the stabilized platform were used during the data processing to compensate the original output data of the GGI for vertical motion, demodulation of the gravity gradient signal, horizontal angular velocity compensation, and self-gradient compensation. The gravity gradient measurement data results were then obtained after lowpass filtering the compensated gravity gradient output data with a cutoff frequency of 1/300 Hz (equivalent spectral resolution of 750 m [24]). The data processing flow chart is shown in Figure 10.

3.2 Accuracy evaluation

3.2.1 Repeat measurement accuracy evaluation

The gravity gradient data from the repeated line measurements in the survey network were compared for internal conformity, and the corresponding accuracy index was calculated based on the existing gravity measurement repetition line evaluation method [25]. The root mean square (RMS) error of the measurement line intersection discrepancy value was taken as the main evaluation criterion to measure the accuracy of the gravity gradiometer. To ensure that the order in which the figures appear is consistent with

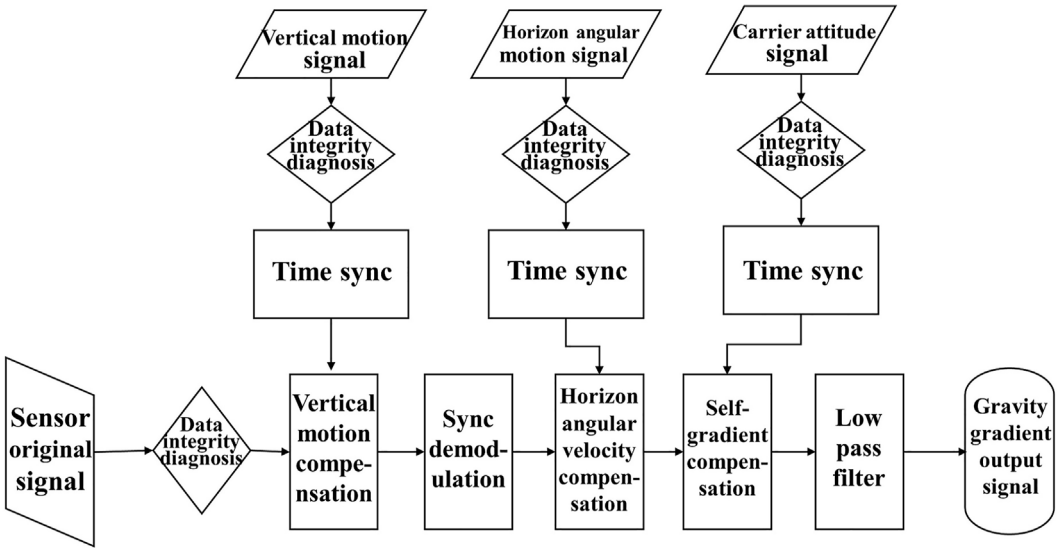


FIGURE 10
Data processing flowchart of GGMS.

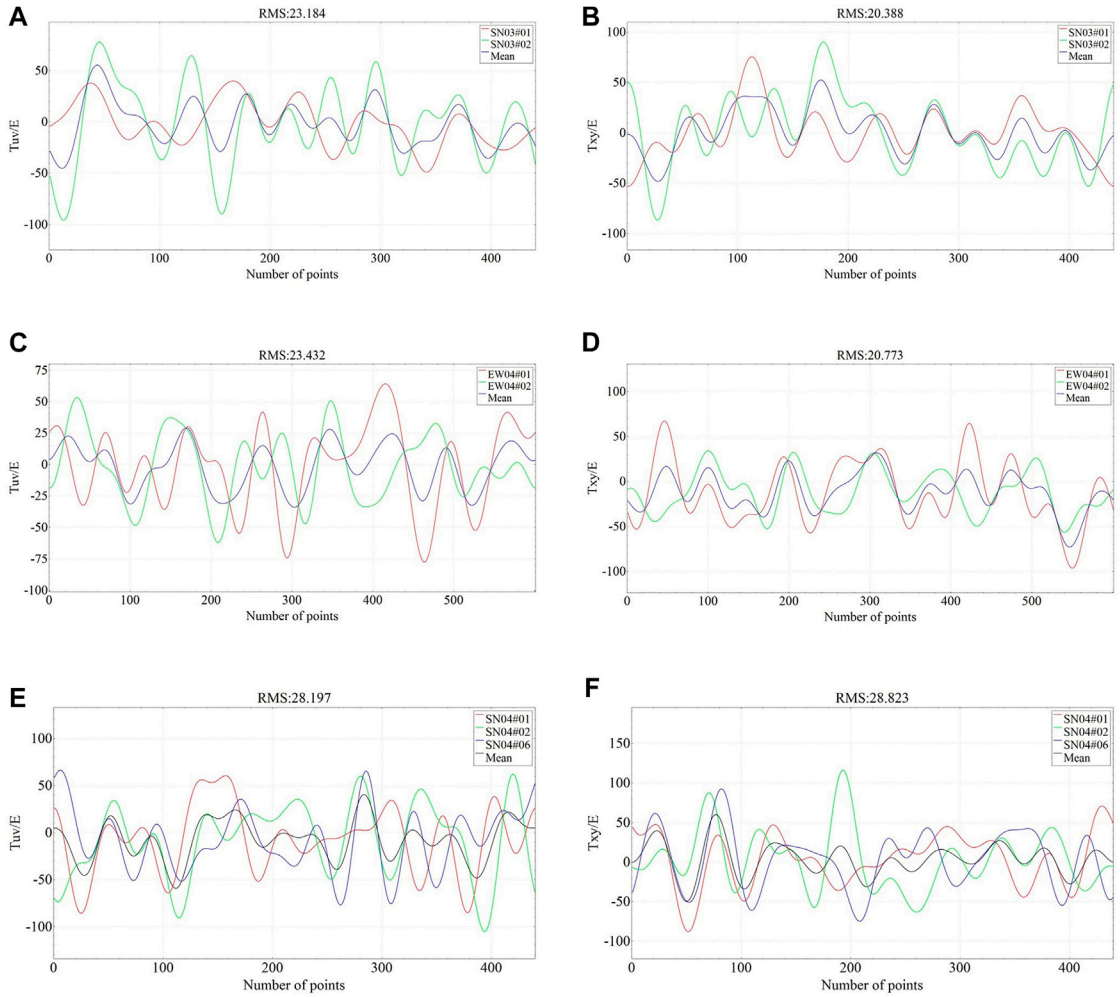


FIGURE 11
(A–F) T_{uv} and T_{xy} component of Repeat line internal accord accuracy chart.

the text, the highlighted text is modified as follows, The internal accord accuracy for the T_{uv} and T_{xy} components of the repetition line of SN03 were 23.2 E@750m and 20.4 E@750m, respectively (Figures 11 A, B). The internal accord accuracy for the T_{uv} and T_{xy} components of the repetition line of EW04 were 23.4 E@750m and 20.8 E@750m, respectively (Figures 11 C, D). Three of six repetitive measurements were performed along the SN04# survey line in this sea trial, as shown in Figure 9. Because of high data noise of three repetition lines of SN04, the internal accord accuracy for the T_{uv} and T_{xy} components of the other three repetition line were 28.2E@750m and 28.8E@750m, respectively (Figures 11 E, F).

As shown in the Figure 11, in general, the internal accord accuracy of the T_{xy} is higher than that of T_{uv}. Although the characteristics and change trends of some measurement curve lines are not very similar especially in Figures 11C, E, we believe this is a signal distortion caused by noise in the data, that is, there is still a lot of work to be done in improving the signal-to-noise ratio of the data, including updating data processing methods, especially in the optimization of dynamic compensation algorithm.

3.2.2 Internal conformity accuracy evaluation of the intersection points

The grid measurement accuracy was evaluated based on the mean squared difference of the residuals at the intersections of the survey and tie lines:

$$\sigma = \sqrt{\frac{1}{2N}\sum_{i=1}^{N}\delta_i^{\,2}} \tag{3}$$

where δ_i is the difference between the ith survey line and the tie line at the intersection of the measurement line, and N is the number of intersection points in the calculation.

There were 23 valid intersection points according to the survey grid shown in Figure 9. Survey lines SN01# and SN02# were curved before the start of EW05#, resulting in a reduced accuracy during the lowpass filtering process; these intersection points were therefore excluded from the evaluation. The accuracies of the T_{uv} and T_{xy} measurements are 28.2E@750m and 26.8@750m, respectively. A statistical chart of the gravity gradient difference values at the intersection points is shown in Figure 12.

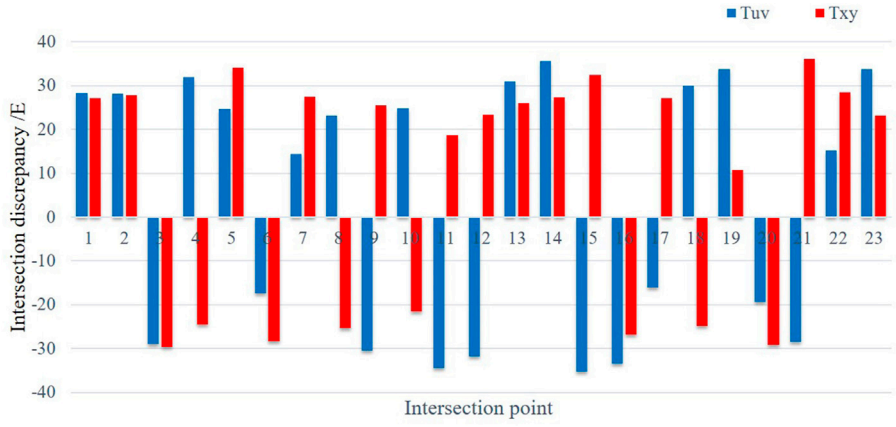


FIGURE 12
Statistics of discrepancy values at the intersection points.

4 Conclusion

The working principle of a new shipboard GGMS was fully verified after a series of static and dynamic tests, and shipboard sea trials. The T_{uv} accuracies of the laboratory-based static test, laboratory-based dynamic test on a shaking table, and shipborne sea trial under dynamic conditions were better than 7.22 E, 9.09 E, and 28.2E@750m, respectively, and the T_{xy} were better than 3.58 E, 4.16 E and 28.8E@750m, respectively. The precision index of our shipboard GGMS reached the same order of magnitude as those of existing commercial instruments.

Next, we will continue to improve the hardware level of this system to enhance the accelerometer resolution, and make the system more compact and light-weight. We will also improve the signal-to-noise ratio and measurement accuracy of measurement data by upgrading and optimizing the motion compensation algorithms. We note that we only use the internal accord accuracy as the evaluation index in this paper since there are no other gravity data in the measurement area. The next step is to conduct an evaluation of external accord accuracy, in combination with the existing gravity data, and develop a measurement calibration method for gravity gradient measurements, which will improve the ability to verify the accuracy index of the instrument.

Data availability statement

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

Author contributions

JZ and QS contributed to conception of the study. All authors have made a substantial contribution to the study and approved the submitted version.

Funding

The authors thank the China Geological Survey Project of China Geological Survey (DD20191004) and Cooperation Research and Demonstration Application of Monitoring Technologies for the Snow, Glaciers and Geohazards in High Mountain Asia and Arctic (21YFE0116800).

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

1. Zeng HL. *Gravity field\and gravity exploration*. Beijing: Geological Press (2005). p. 273.

2. Li X. Vertical resolution: Gravity versus vertical gravity gradient. *The Leading Edge* (2001) 20(8):901–4. doi:10.1190/1.1487304

3. Bell RE. Gravity gradiometry. *Scientific Am* (1998) 278(6):74–9. doi:10.1038/scientificamerican0698-74

4. Evstifeev MI. The state of the art in the development of onboard gravity gradiometers. *Gyroscopy and Navigation* (2017) 8(1):68–79. doi:10.1134/S2075108717010047

5. Difrancesco D, Meyer T, Christensen A, Fitzgerald D. *Gravity gradiometry–today and tomorrow*. 11th SAGA Biennial Technical Meeting and Exhibition (2009) doi:10.3997/2214-4609-pdb.241

6. Ji B, Liu M, Lv L, Bian SF. Application of gradiometer to underwater safe navigation. *Hydrographic Surv charting* (2010) 30(4):23. doi:10.3969/j.issn.1671-3044.2010.04.007

7. Beiki M, Pedersen LB. Eigenvector analysis of gravity gradient tensor to locate geologic bodies. *Geophysics* (2010) 75(6):I37–I49. doi:10.1190/1.3484098

8. Yang J, Jekeli C, Liu L. Seafloor topography estimation from gravity gradients using simulated annealing. *J Geophys Res Solid Earth* (2018) 123:6958–75. doi:10.1029/2018JB015883

9. Fan D, Li S, Li X, Yang J, Wan X. Seafloor topography estimation from gravity anomaly and vertical gravity gradient using nonlinear iterative least square method. *Remote Sensing* (2020) 13(1):64. doi:10.3390/rs13010064

10. Dransfield MH. Airborne gravity gradiometry in the search for mineral deposits. In: B Mikereit, editor. *Proceedings of exploration 07: Fifth decennial international conference on mineral exploration* (2007). p. 341.

11. Shu Q, Zhou JX, Yin H. Present research situation and development trend of aviation gravity gradiometer. *Geophys Geochemical Exploration* (2007) 31(12):485.

12. Murphy CA, Mumaw GR. 3D full tensor gradiometry: A high resolution gravity measuring instrument resolving ambiguous geological interpretations. *ASEG Extended Abstr* (2004) 1:1–4. doi:10.1071/aseg2004ab104

13. Mims J, Selman D, Dickinson J, Murphy C, Mataragio J. *Comparison study between airborne and ship-borne full tensor gravity gradiometry (FTG) data*. SEG Expanded Abstracts (2009). doi:10.1190/1.3255906

14. Roth M. *Marine full tensor gravity gradiometry data analysis and euler deconvolution*. Stuttgart: University of Stuttgart (2009). p. 50.

15. Rummel R, Yi W, Stummer C. GOCE gravitational gradiometry. *J Geodesy* (2011) 85(11):777–90. doi:10.1007/s00190-011-0500-0

16. Wu Q, Teng YT, Zhang B, Zhang T. The research situation of the gravity gradiometer in the world. *Geophys Geochemical Exploration* (2013) 22(5):761. doi:10.11720/j.issn.1000-8918.2013.5.01

17. Aravanis T, Chen J, Fuechsle M, Grujic M, Johnston P, Kok Y, et al. VK1™ — a next-generation airborne gravity gradiometer. *ASEG Extended Abstr* (2016) 1:1–5. doi:10.1071/ASEG2016ab318

18. Yang GD, Weng KX, Wu B, Cheng B, Lin Q. Research progress of quantum gravity gradiometer. *Navigation Positioning and Timing* (2021) 8(2):18. doi:10.19306/j.cnki.2095-8110.2021.02.003

19. Stray B, Lamb A, Kaushik A, Vovrosh J, Rodgers A, Winch J, et al. Quantum sensing for gravity cartography. *Nature* (2022) 602(7898):590–4. doi:10.1038/s41586-021-04315-3

20. Shu Q. *Research on airborne gravity gradient measurement technology*. Jilin: Jilin University (2018). p. 163.

21. Meng Z, Yang Y, Li Z. Development of airborne gravity gradiometer based on a quartz flexible accelerometer. *Acta Geologica Sinica - English Edition* (2019) 93(S1):352–64. doi:10.1111/1755-6724.14133

22. Yang Y, Li D, Gao W. Output signal demodulation and filter for rotating accelerometer gravity gradiometer[J]. *J Chin Inertial Technol* (2016) 24(6):701. doi:10.13695/j.cnki.12-1222/o3.2016.06.001

23. Luo Y, Yao CL. Theoretical study of rectangular magnetic fields and their gradient-free analytic singularity expressions. *Oil Geophys Prospecting* (2007)(06) 714.

24. Sun ZM, Xia ZR. Design of fir lowpass differentiator and its applications in airborne gravimetry. *Chin J Geophys* (2000) 43(6):897–903. doi:10.1002/cjg2.106

25. Guo ZH, Xiong SQ, Zhou JX, Zhou XH. A research on data quality evaluation method of repeat lines in airborne gravity survey. *Chin J Geophys* (2008) 51(5):1093–9. doi:10.1002/cjg2.1303

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 26 January 2023
DOI 10.3389/fphy.2023.1134160

OPEN ACCESS

EDITED BY
Huadan Zheng,
Jinan University, China

REVIEWED BY
Yinghua Shen,
Chongqing University, China
Xiaoan Tang,
Hefei University of Technology, China
Hengrong Ju,
Nantong University, China

*CORRESPONDENCE
Weiwei Mao,
maoweiweiojiang@163.com

SPECIALTY SECTION
This article was submitted to Optics and Photonics, a section of the journal Frontiers in Physics

RECEIVED 30 December 2022
ACCEPTED 09 January 2023
PUBLISHED 26 January 2023

CITATION
Lv Y, Mao W and Cui Y (2023), Joint DOD and DOA detection for MIMO radar based on signal subspace reconstruction and matching.
Front. Phys. 11:1134160.
doi: 10.3389/fphy.2023.1134160

Joint DOD and DOA detection for MIMO radar based on signal subspace reconstruction and matching

Yan Lv[1], Weiwei Mao[2]* and Ye Cui[3]

[1]School of Optoelectronic Engineering, Xidian University, Xi'an, China, [2]Beijing Jianzhu Technology Co., LTD, Xi'an, China, [3]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada

In this study, the Direction Of Departure (DOD) and Direction Of Arrival (DOA) of signals detection for Multi-Input Multi-Output (MIMO) radar is discussed. A novel signal subspace reconstruction model to match the signal subspace obtained based on the covariance matrix of the array output is developed to enhance the performance of the DOD and DOA detection. In the developed scheme, the technology of beamforming is first introduced to define an objective space in mathematics for the targets to be detected. By considering the orthogonality between the signal subspace and the noise subspace and defining a reconstruction index of the signal subspace, a multi-dimensional objective function of the DOD and DOA is established. Therefore, the problem of DOD and DOA detection is transformed into an optimization of the multi-dimensional objective function. Subsequently, the Quantum-Behaved Particle Swarm Optimization (QPSO) is employed to optimize the multi-dimensional objective function and to determine an optimal signal subspace. At the same time the DOD and DOA can be fast captured. A series of simulations demonstrate that the proposed method provides significant accuracy improvements in DOD and DOA detection, especially for low signal-to-noise ratio thresholds and small snapshots.

KEYWORDS

multi-input multi-output (MIMO) radar, direction of departure (DOD), direction of arrival (DOA), quantum-behaved particle swarm optimization (QPSO), signal-to-noise ratio

Introduction

Multi-Input Multi-Output (MIMO) radar [1–3] is a new type of radar in recent years. Compared with traditional antenna arrays, the MIMO radars have potential advantages. The reason is that the MIMO radars make use of multiple antenna elements to transmit diverse waveforms and receive echo signals simultaneously in similar ways [4–6]. The detection of Direction Of Departure (DOD) and Direction Of Arrival (DOA) for MIMO radars has been actively studied and widely applied in many technical fields [7–10]. Subspace-based algorithms are an important kind of methods for DOD and DOA detection, which has the advantages of strong resolution and high detection accuracy [11, 12].

Among the subspace-based methods the most representative ones are the MUltiple SIgnal Classification (MUSIC) [13] and the Estimation Signal Parameter *via* Rotational Invariance Techniques (ESPRIT) [14]. The MUSIC approach is well-known for its high-resolution capability, and can be used for arrays with any form of geometry to detect DOA and DOA of each target [15, 16]. However, the massive spectral peak search makes it difficult to popularize this algorithm in the engineering field. Without spectral peak search (search-free detection), the

ESPRIT algorithm can detect the DOA and DOA in a closed form solution [8, 17–19]. However, this type of methods can only be applied to the Uniform Linear antenna Array (ULA), or the antenna arrays themselves possess a same structure of (shift invariant) subarrays. This limitation makes it difficult to extend such algorithms to practical applications [20–22].

Thus, fast detection of the DOD and DOA in the MIMO radar system with an arbitrary antenna array has always been the pursuit for researchers. In [18], an ambiguity function-based algorithm is designed to detect the DOD and DOA in a MIMO Radar system. By constructing a spatial time-frequency distribution matrix, this method uses ESPRIT and Root-MUSIC to realize the joint DOD and DOA fast detection. This process does not involve multiple-dimensional spectrum peak searching and the parameters can also be paired automatically. In [20], a robust method for joint DOD and DOA detection in a non-Gaussian noise environment is proposed. This method uses a robust M-estimator to form an estimate of the covariance matrix of the array output and then utilizes the random matrix theory and polynomial rooting method to capture the DOD and DOA in a large scale MIMO radar system. By exploiting the banded complex symmetric Toeplitz structure of the mutual coupling matrices [21], a robust sparse Bayesian learning algorithm for DOD and DOA detection is also developed, and this method is demonstrated to work well and have better detection performance in unknown non-uniform noise and mutual coupling.

In this study, a joint DOD and DOA detection scheme through signal subspace optimal reconstruction is developed. In the developed scheme, we first use the beamforming [23–25] to set an objective space in mathematics for the targets to be detected. Then, we randomly select a collection of DODs and DOAs from the objective space to establish a potential signal subspace with the steering matrix of the MIMO array. Considering of the orthogonality between the signal subspace and the noise subspace and a reconstruction error of the signal subspace, we build a multi-dimensional objective function, which contains all information of DODs and DOAs. Subsequently, the Quantum-Behaved Particle Swarm Optimization (QPSO) [26, 27] is employed to optimize the multi-dimensional objective function so as to obtain the optimal reconstruction of the signal subspace. Finally, the DOD and DOA can be fast detected with a high accuracy. A detailed derivation and comprehensive analysis of the proposed method are presented. The simulation results verify that the proposed algorithm outperforms the other methods, especially under low signal to noise ratio (SNR) and small snapshots. To the best of our knowledge, the proposed scheme has not been considered in previous studies.

This paper is structured as follows. The signal model for a MIMO array is formulated in Section 2. A fast DOD and DOA detection method is discussed in detail in Section 3. Section 4 includes experimental setup and analysis of simulation results. Finally, Section 5 concludes the paper.

Signal model of the MIMO array

Consider a bistatic MIMO radar array composed of M_t transmitter sensors and M_r receiver sensors, in the transmit array and receive array, respectively. Both of these arrays are ULAs, and the inter-element spacings of adjacent sensor elements are d_t and d_r (not larger than half wavelength of the signals), respectively. Assume that there are P far-field uncorrelated targets with the DODs and DOAs (φ_p, θ_p), $p = 1, 2, \mathrm{L}, P$ in the target space. Figure 1 shows the signal model of the bistatic MIMO array. The output of the matched filters at the receiver can be expressed in the following manner:

$$\begin{aligned}\boldsymbol{X}(t) &= \left[\boldsymbol{a}_r(\varphi_1) \otimes \boldsymbol{a}_t(\theta_1), \boldsymbol{a}_r(\varphi_2) \otimes \boldsymbol{a}_t(\theta_2), \mathrm{L}, \boldsymbol{a}_r(\varphi_P) \otimes \boldsymbol{a}_t(\theta_P)\right]\boldsymbol{b}(t) + \boldsymbol{n}(t) \\ &= \boldsymbol{A}\boldsymbol{b}(t) + \boldsymbol{n}(t)\end{aligned} \tag{1}$$

where $\boldsymbol{A}$ is the direction matrix (steering matrix), $\boldsymbol{b}(t) = [b_1(t), b_2(t), \mathrm{L}, b_P(t)]^T$ is a vector containing the reflection coefficients and Doppler phase shifts of the targets, and $\boldsymbol{n}(t)$ is the complex additive white gaussian noise vector. $\boldsymbol{a}_r$ and $\boldsymbol{a}_t$ are receive and transmit steering vectors which have the following structures:

$$\begin{aligned}\boldsymbol{a}_r(\varphi_p) &= \left[1, e^{j\frac{\pi d_r \sin\varphi_p}{\lambda}}, \mathrm{L}, e^{j\frac{\pi(M_r-1)d_r \sin\varphi_p}{\lambda}}\right]^T \\ \boldsymbol{a}_t(\theta_p) &= \left[1, e^{j\frac{\pi d_t \sin\theta_p}{\lambda}}, \mathrm{L}, e^{j\frac{\pi(M_t-1)d_r \sin\theta_p}{\lambda}}\right]^T\end{aligned} \tag{2}$$

where T stands for the transpose operation, and λ is the wavelength of the signals. The covariance matrix of the array output is computed by

$$\boldsymbol{R}_X = E\left[\boldsymbol{X}(t)\boldsymbol{X}^H(t)\right] \tag{3}$$

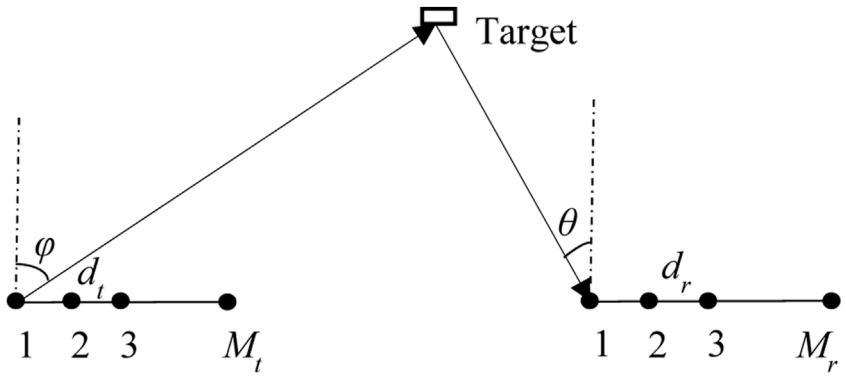


FIGURE 1
Signal model of the bistatic MIMO array.

where H represents the complex conjugate transpose. The signal and noise subspaces are commonly obtained through the eigenvalue decomposition of the covariance matrix of the array output. That is, the eigenvectors corresponding to the P largest eigenvalues span the signal subspace $\boldsymbol{U}_s$, and the remainder eigenvectors form the noise subspace $\boldsymbol{U}_n$. The MUSIC method detects the DOD and DOA by constructing and optimizing the following spatial spectrum function

$$f_{MUSIC} = \frac{1}{\left[\boldsymbol{a}_r(\varphi_i) \otimes \boldsymbol{a}_t(\theta_i)\right]^H \boldsymbol{U}_n \boldsymbol{U}_n^H \left[\boldsymbol{a}_r(\varphi_i) \otimes \boldsymbol{a}_t(\theta_i)\right]} \tag{4}$$

The spatial spectrum function is expected to show a large positive value if φ and θ are a true DOD and DOA, which is due to the orthogonality between the signal subspace and noise subspace. Obviously, the peak search implies a huge amount of computation. How to reduce the amount of computation generated in the search process has always been the goal of researchers.

Joint DOD and DOA detection scheme through signal subspace optimal reconstruction and matching

In this section, a fast joint DOD and DOA detection scheme through signal subspace optimal reconstruction and matching will be developed in detail. The developed scheme, as we will demonstrate, has excellent detection performance under the scenarios of low SNR and small snapshot.

To reduce the search scope, we firstly use the beamforming to capture the potential regions (including the DODs and DOAs) of the targets. Let Ω_k ($k = 1, 2, \ldots, K$) be the K spectrum peaks determined by the beamforming. We take the neighbourhoods of the K spectrum peaks in mathematics as the potential regions of the targets, and the following searching will be carried out on these regions (observation space).

Then, we randomly select P (P can obtained by the Minimum Descriptive Length, MDL) criterion directions to form a direction vector $\tilde{\boldsymbol{\Omega}}$ to construct an initial steering matrix $\tilde{\boldsymbol{A}}$ according to Eq. 2, with which an initial signal subspace can be reconstructed in the following way:

$$\tilde{\boldsymbol{U}}_s = \tilde{\boldsymbol{A}}\left(\tilde{\boldsymbol{\Omega}}\right)\left[\tilde{\boldsymbol{A}}^H\left(\tilde{\boldsymbol{\Omega}}\right)\tilde{\boldsymbol{A}}\left(\tilde{\boldsymbol{\Omega}}\right)\right]^{-\frac{1}{2}} \tag{5}$$

Ideally, we consider that this signal subspace and the estimated signal subspace with the eigenvalue decomposition of the covariance matrix of the array output should be equal, i.e.,

$$\left\|\tilde{\boldsymbol{U}}_s\tilde{\boldsymbol{U}}_s^H - \boldsymbol{U}_s\boldsymbol{U}_s^H\right\|_2 \to 0 \tag{6}$$

where $\|g\|_2$ denotes l_2 norm. On the other hand, consider the orthogonality between the signal and noise subspaces, the following equation should also hold:

$$\left\|\tilde{\boldsymbol{U}}_s^H\boldsymbol{U}_n\right\|_2 = 0 \tag{7}$$

In order to detect more accurate DODs and DOAs, the above two equations should hold at the same time. Thus, we build a multi-dimensional objective function in the following form:

$$J = \left\|\tilde{\boldsymbol{U}}_s\tilde{\boldsymbol{U}}_s^H - \boldsymbol{U}_s\boldsymbol{U}_s^H\right\|_2 + \left\|\tilde{\boldsymbol{U}}_s^H\boldsymbol{U}_n\right\|_2 \tag{8}$$

The DODs and DOAs can be obtained through optimizing the objective function in a $2P$ dimensional space. In this study, we use the Quantum-Behaved Particle Swarm Optimization (QPSO) as an optimization method to minizine the multi-dimensional objective function and to solve the DODs and DOAs of the signals. Suppose there are N particles, and the optimization process can be described as:

1) Initialize the particle position vector $\tilde{\boldsymbol{\Omega}}^{(n)}$ ($n = 1, 2, \ldots, N$) and the best previous position $\tilde{\boldsymbol{\Omega}}_{best}^{(n)}$ of each particle.
2) Capture the mean of the best position $\boldsymbol{mbest}(t)$ according to

$$\boldsymbol{mbest} = \frac{1}{N}\sum_{n=1}^{N}\tilde{\boldsymbol{\Omega}}_n^{pbest}(t) \tag{9}$$

3) Substitute the particle position vector $\tilde{\boldsymbol{\Omega}}^{(n)}$ into the objective function 8) that is being optimized and at the same time compare it with the particle's previous best value so as to determine the current fitness value.
4) Determine the current global best ($g$best) position through comparing it with the previous one and determine whether to update it or not.
5) Modify the positions of particles to make all of them fall within the definition domain.
6) Repeat steps **(2)** to **(5)** until the termination condition or the number of executions has been saturated.

In the implementation of the QPSO, the approach is run for $n$max = 300 iterations with 30 particles by considering the dimension of the target space. However, we allow the method to be terminated if no changes in $g$best is 20%$n$max consecutive iterations. The particle position vector generated at algorithm termination is the DODs and DOAs we want to detect.

Experimental studies:

In order to evaluate the performance of the developed scheme, in this section we present two groups of simulations to assess the DOD and DOA detection performance of our algorithm in comparison with the other two commonly methods (MUSIC and ESPRIT). In the comparison, the Root-Mean-Square Error (RMSE) criterion [28, 29] is used as an evaluation indicator which is define as

$$\sqrt{\frac{1}{N_s}\sum_{n=1}^{N_s}\left\{\frac{1}{P}\sum_{p=1}^{P}\left[\hat{\alpha}_p(n) - \alpha\right]^2\right\}} \tag{10}$$

where N_s denotes the number of the independent simulations, α_p is the pth DOD or DOA of the P impinging sources, and $\hat{\alpha}_p$ is the estimate value of α_p captured in the nth simulation.

In the simulation, we normally adopt the bistatic MIMO radar system with $?_t = 8$ and $M_r = 8$, and assume that there are three non-coherent signals located at $(\varphi_1, \theta_1) = (5^\circ, 15^\circ)$, $(\varphi_2, \theta_2) = (10^\circ, 30^\circ)$, and $(\varphi_3, \theta_3) = (30^\circ, 10^\circ)$, respectively.

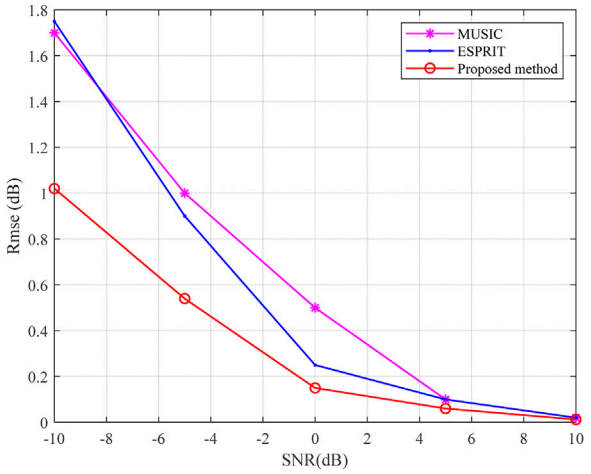


FIGURE 2
DOD estimation performance with different SNR.

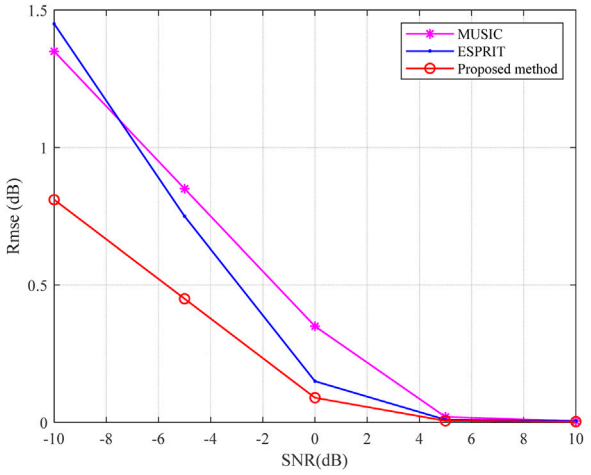


FIGURE 3
DOA estimation performance with different SNR.

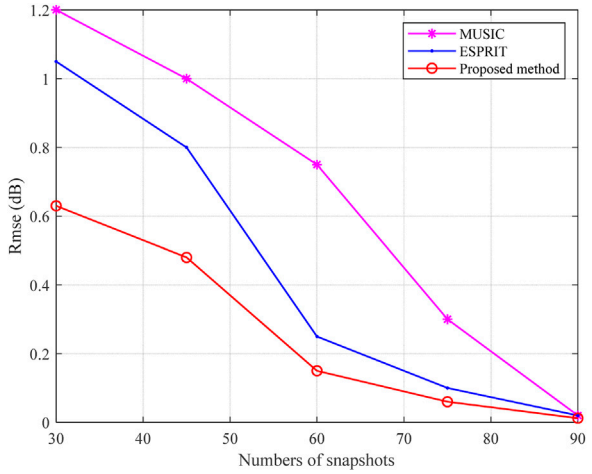


FIGURE 4
DOD estimation performance with different snapshots.

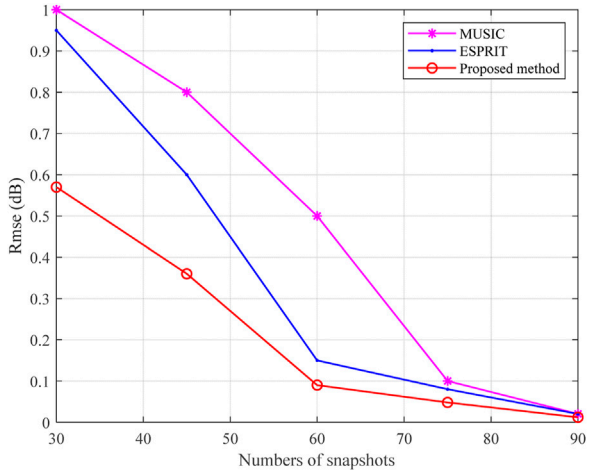


FIGURE 5
DOA estimation performance with different snapshots.

Simulation 1

This simulation quantifies the performance of the RMSE *versus* SNR. In the simulation, the number of snapshots is fixed as 50, and 100 Monte Carlo experiments are completed. Figures 2, 3 visualize the experimental results of the DOD and DOA detection. It can be seen that these methods can all achieve DOD and DOA detection when the SNR is high (say >0 dB). The RMSE drops with the SNR decreasing, and the proposed method outperforms the other methods under most situations.

Simulation 2

In the second simulation, we test the RMSE of different methods *versus* the numbers of snapshots. In this simulation, the SNR is fixed as -5dB, and 100 Monte Carlo experiments are carried out. Figures 4, 5 show the plot of the performance of RMSE of the DOD and DOA detection *versus* the numbers of snapshots. It is apparent that the developed scheme outperforms the other approaches and becomes insensitive to the changes of the numbers of snapshots.

The principles of MUSIC and ESPRIT imply that their detection performance is mainly dependent on the covariance matrix of the array output and the signal and noise subspaces. At the situation of low SNR and small number of snapshots, their performance deteriorates significantly or even fails largely due to the inaccuracy of the covariance matrix of the array output and signal and noise subspaces. However, the proposed method detects the DOD and DOA through exploiting a signal subspace reconstruction model to match the signal subspace obtained based on the covariance matrix of the array output, which eliminates the dependence on the covariance matrix to a certain extent.

Conclusion

In this paper, we put forward a novel signal subspace reconstruction model to match the signal subspace obtained based on the covariance matrix of the array output to enhance the performance of the DOD and DOA detection. In the design progress, the problem of DOD and DOA detection is transformed into a signal subspace reconstruction and matching problem. A multi-dimensional objective function is built and optimized to solve the DOD and DOA. Simulation results prove the performance of the developed algorithm compared with the other subspace-based methods.

At the current stage, we have completed a theoretical analysis and offered a comprehensive suite of experiments. Some practical experiments (including hardware) would be an interesting avenue to explore in future studies.

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

All the authors made significant contributions to the work. The idea was proposed by WM; YL and YC simulated the algorithm, analysed the data designed the experiments; WM, YC and YL polished the English and wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the Shaanxi Provincial Fund for Distinguished Young Scholars under Grant 2021JC-23.

Conflict of interest

WM was employed by Beijing Jianzhu Technology Co., LTD

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.

References

1. Sempere Chaves C, Geschke RH, Shargorodskyy M, Herschel R, Kose S, Leuchs S, et al. Multisensor polarimetric MIMO radar network for disaster scenario detection of persons. *IEEE Microwave Wireless Components Lett* (2022) 32(3):238–40. doi:10.1109/lmwc.2021.3132788

2. Han K, Hong S. High-resolution phased-subarray MIMO radar with grating lobe cancellation technique. *IEEE Trans Microwave Theor Tech* (2022) 70(5):2775–85. doi:10.1109/tmtt.2022.3151633

3. Niu S, Zhu D, Jin G, Cheng Y, Wang Y. A novel transmitter-interpulse phase coding MIMO-radar for range ambiguity separation. *IEEE Trans Geosci Remote Sensing* (2022) 60:1–16. doi:10.1109/tgrs.2022.3221406

4. Norrdine A, Cetinkaya H, Herschel R. Radar wave based positioning of spatially distributed MIMO radar antenna systems for near-field nondestructive testing. *IEEE Sensors Lett* (2020) 4(5):1–4. doi:10.1109/lsens.2020.2989546

5. Dontamsetti SG, Kumar RVR. A distributed MIMO radar with joint optimal transmit and receive signal combining. *IEEE Trans Aerospace Electron Syst* (2021) 57(1):623–35. doi:10.1109/taes.2020.3027103

6. Liu Y, Xu X, Xu G. MIMO radar calibration and imagery for near-field scattering diagnosis. *IEEE Trans Aerospace Electron Syst* (2018) 54(1):442–52. doi:10.1109/taes.2017.2760758

7. Schüßler C, Hoffmann M, Bräunig J, Ullmann I, Ebelt R, Vossiek M. A realistic radar ray tracing simulator for large MIMO-arrays in automotive environments. *IEEE J Microwaves* (2021) 1(4):962–74. doi:10.1109/jmw.2021.3104722

8. Liao Y, Zhao R, Gao L. Joint DOD and DOA estimation in bistatic MIMO radar with distributed nested arrays. *IEEE Access* (2019) 7:50954–61. doi:10.1109/access.2019.2904613

9. Xu F, Vorobyov SA, Yang X. Joint DOD and DOA estimation in slow-time MIMO radar via PARAFAC decomposition. *IEEE Signal Process. Lett* (2020) 27:1495–9. doi:10.1109/lsp.2020.3018904

10. Xu J, Wang WQ, Gui R. Computational efficient DOA, DOD, and Doppler estimation algorithm for MIMO radar. *IEEE Signal Process. Lett* (2019) 26(1):44–8. doi:10.1109/lsp.2018.2879546

11. Zhang Y, Wang Y, Tian Z, Leus G, Zhang G. Efficient angle estimation for MIMO systems via redundancy reduction representation. *IEEE Signal Process. Lett* (2022) 29:1052–6. doi:10.1109/lsp.2022.3164850

12. Cui C, Xu J, Gui R, Wang W, Wu W. Search-free DOD, DOA and range estimation for bistatic FDA-MIMO radar. *IEEE Access* (2018) 6:15431–45. doi:10.1109/access.2018.2816780

13. Xu KJ, Quan YH, Bie BW, Xing MD, Nie WK, Hanyu E. Fast direction of arrival estimation for uniform circular arrays with a virtual signal subspace. *IEEE Trans Aerospace Electron Syst* (2021) 57(3):1731–41. doi:10.1109/taes.2021.3050667

14. Chen Z, Chen P. Compressed sensing-based DOA and DOD estimation in bistatic co-prime MIMO arrays. In: 2017 IEEE Conference on Antenna Measurements & Applications (CAMA); 4-6 Dec. 2017; Tsukuba, Japan (2017). p. 297–300.

15. Roy R, Kailath T. ESPRIT-Estimation of signal parameters via rotational invariance techniques. *IEEE Trans Acoust Speech, Signal Process* (1989) 37(7):984–95. doi:10.1109/29.32276

16. Shi W, He C, Huang J, Zhang Q. Fast MUSIC algorithm for joint DOD-DOA estimation based on Gibbs Sampling in MIMO array. In: 2016 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC); 5-8 Aug. 2016; Hong Kong, China (2016). p. 1–4.

17. Li Z, Zhou C, Jia Y, Wang G, Yan C, Lv Y. DOD and DOA Estimation for bistatic co-prime MIMO array based on correlation matrix augmentation. In: 2019 International Conference on Control, Automation and Information Sciences (ICCAIS); 23-26 October 2019; Chengdu, China (2019). p. 1–5.

18. Fan T, Jiang H, Sun J. Ambiguity function-based ESPRIT-Root-MUSIC algorithm for DOD-DOA estimation in MIMO radar. In: International Conference on Radar Systems (Radar 2017); 23-26 Oct. 2017; Belfast, UK (2017). p. 1–4.

19. Zhao X, Guo C, Chu S, Zhang W. DOD and DOA estimation of coherent targets in bistatic MIMO radar based on DFSS_ESPRIT. In: 2021 13th International Symposium on Antennas, Propagation and EM Theory (ISAPE); Jan 12, 2021-Apr 12, 2021; Zhuhai, China (2021). p. 1–3.

20. Jiang H, Lu Y, Yao S. Random matrix based method for joint DOD and DOA estimation for large scale MIMO radar in non-Gaussian noise. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 20-25 March 2016; Shanghai, China (2016). p. 3031–5.

21. Tang WG, Jiang H, Zhang Q. Robust DOD and DOA estimation for bistatic MIMO radar in unknown mutual coupling and non-uniform noise. In: 2022 IEEE 12th Sensor Array and Multichannel Signal Processing Workshop (SAM); June 20-23, 2022; Trondheim, Norway (2022). p. 116–20.

22. Wen F, Zhang Z, Zhang G. Joint DOD and DOA estimation for bistatic MIMO radar: A covariance trilinear decomposition perspective. *IEEE Access* (2019) 7:53273–83. doi:10.1109/access.2019.2912842

23. Xu K, Xing M, Zhang R, Hanyu E, Sha M, Nie W, et al. High-accuracy DOA estimation algorithm at low SNR through exploiting a supervised index. *IEEE Trans Aerospace Electron Syst* (2022) 58(4):3658–65. doi:10.1109/taes.2022.3144121

24. Han J, Ng BP, Er MH. An adaptive orientational beamforming technique for narrowband interference rejection. In: 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 22 May 2022 – 27 May 2022; Singapore (2022). p. 4908–12.

25. Zhu D, Li B, Liang P. A novel hybrid beamforming algorithm with unified analog beamforming by subspace construction based on partial csi for massive MIMO-OFDM Systems. *IEEE Trans Commun* (2017) 65(2):594–607. doi:10.1109/tcomm.2016.2625794

26. Sun J, Fang W, Wu X, Palade V, Xu W. Quantum-behaved particle swarm optimization: Analysis of individual particle behavior and parameter selection. *Evol Comput* (2012) 20:349–93. doi:10.1162/evco_a_00049

27. Luo J, Shao Y, Liao X, Liu J, Zhang J. Complex permittivity estimation for cloths based on QPSO method over (40 to 50) GHz. *IEEE Trans Antennas Propagation* (2021) 69(1):600–5. doi:10.1109/tap.2020.3005032

28. Yu X, Liu B, Qiu G, Tian C. 2D-DOA efficient estimation algorithm based on IRD-MUSIC for frequency hopping signal. In: 2022 IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT); 12-14 October 2022; Dali, China (2022). p. 675–9.

29. Wu R, Zhang Y, Ni L, Zhang K, Liu N, Wan Q. DOA estimation for beam scanning radar based on sparse signal reconstruction with incomplete pulse. In: 2022 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC); October 25-27, 2022; Xi'an, Shaanxi, China (2022). p. 1–6.

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 22 February 2023
DOI 10.3389/fphy.2023.1142154

OPEN ACCESS

EDITED BY
Xukun Yin,
Xidian University, China

REVIEWED BY
Kaijie Xu,
University of Alberta, Canada
Zhonghao Li,
North University of China, China

*CORRESPONDENCE
Qi Li,
liq42@cast504.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Optics and Photonics, a section of the journal Frontiers in Physics

RECEIVED 11 January 2023
ACCEPTED 02 February 2023
PUBLISHED 22 February 2023

CITATION
Li Y, Yang W, Li Q, Chen J, Wang W, Li C and Duan C (2023), A discrete side-lobe clutter recognition method based on sliding filter response loss for space-based radar.
Front. Phys. 11:1142154.
doi: 10.3389/fphy.2023.1142154

A discrete side-lobe clutter recognition method based on sliding filter response loss for space-based radar

Yu Li†, Wenhai Yang†, Qi Li*, Jinming Chen, Weiwei Wang, Caipin Li and Chongdi Duan

Xi'an Institute of Space Radio Technology, Xi'an, China

Different from ground-based or airborne early warning radar, space-based radar (SBR) possesses large coverage capability. As a result, several discrete strong scatter points from the antenna side-lobe shares the same feature with the real targets in range-Doppler domain, which leads to false alarms when conducting constant false alarm rate (CFAR) detection process, and the detection performance with regard to SBR deteriorates seriously. In this paper, a discrete side-lobe clutter recognition method based on sliding filter response loss is proposed for space-based radar. Firstly, considering both the echo inhomogeneity and the limited degrees of freedom (DOFs) after dimension-reduced space-time adaptive processing (STAP), the sliding window design strategy is employed to segment range cells for the observation scene. Then, the images related to different range segments are registered after clutter suppression, in this way, the candidate target parameters, including the position information and the amplitude information are counted. On this basis, the reliable recognition scheme between the real target and the discrete side-lobe clutter can be realized by comparing these filter response losses. Compared with recent works, experimental results based on real measured data show that the proposed method significantly improves the fault-tolerant discrimination ability, which possesses high robustness in algorithm performance as well as good prospect in engineering application.

KEYWORDS

space-based radar, discrete side-lobe clutter, space-time adaptive processing, constant false alarm rate, filter response loss

1 Introduction

For space-based radar, due to its wide detection range, there exist a large number of discrete strong scatter points in the echo data, including iron towers, telegraph poles, tall buildings, isolated islands and marine targets [1]. These discrete side-lobe clutters, which come from the antenna side-lobe, cannot be distinguished from the real targets as they are located outside the main lobe clutter in Doppler domain. Space-time adaptive processing (STAP) [2–6] is currently one of the effective means for clutter suppression. However, the reason for the emergence of discrete side-lobe clutters in heterogeneous environment lies in the sample selection strategy when constructing the clutter covariance matrix (CCM). The amplitude of discrete side-lobe clutters is affected several factors, i.e., radar detection range, antenna side-lobe level, self-scattering cross section, and hence resulting in a certain

amplitude randomness after clutter suppression. That is, in order to alleviate the influence of non-independent identical distribution (IID) samples on clutter suppression performance, strong scatter points may be regarded as heterogeneous samples, and thus the corresponding resolution cells cannot be suppressed effectively for their CCM have a bad match with the cell under test (CUT) [7, 8].

Presently, the discrete side-lobe clutter recognition researches can be divided into two categories. The first one is the synthetic channel gain method [9, 10], in which the candidate targets are identified through comparing the output of the synthetic channel gain and that of the auxiliary channel gain. If the candidate target is located at the antenna beam pointing position, the synthetic channel gain is $10log_{10} N^2$ without considering the system errors, where N stand for the channel number. Instead, if the candidate target comes from the antenna side-lobe, the synthetic channel gain declines obviously for the channel amplitude cannot be coherently accumulated. Thus, discrete side-lobe clutters can be recognized according to the increased channel gain value. The other one is the filter response loss method [11], in which the energy loss of the candidate target are counted in the STAP process. Compared with the real target, discrete side-lobe clutters suffer more energy loss as the target steering vector is pointed at the main lobe area. On this basis, multiple space-time steering vector constraint methods are presented to identify discrete side-lobe clutter [12]. Firstly, two steering vectors related to the target and main-lobe clutter regions are employed to construct the optimal weight vectors. Then, the energy loss of candidate targets after STAP processing is compared, where these resolution cells with large energy loss are judged as the real targets, others are regarded as the discrete side-lobe clutters. To a certain extent, the above two methods provide an effective means for discrete side-lobe clutter recognition. However, the real targets generally do not come from the antenna beam center pointing direction, and hence the synthetic channel gain method cannot achieve the ideal value. In addition, although the discrete side-lobe clutters share apparent amplitude attenuation with regard to filter response loss method, there also exist energy loss for the real target. Therefore, the existing methods cannot significantly enhance the fault-tolerant capability when conducting discrete side-lobe clutter recognition process. It is necessary to further explore the robust discrete side-lobe clutter recognition method for space-based radar with wide area coverage. Aiming at the above problems, a discrete side-lobe clutter recognition method based on sliding filter response loss for space-based radar is proposed in this paper, in which the wide area detection ability and the sample inhomogeneity characteristics are combined, and the filter response losses corresponding to different segmented data are evaluated to achieve the recognition capability between the real targets and discrete side-lobe clutters. Meanwhile, this method introduces the sample segmentation registration strategy, and thus to alleviate the poor clutter suppression ability at the scene edge area, which possesses simple implementation and high robustness characteristics. Experimental results with airborne measured data verify the effectiveness of this method.

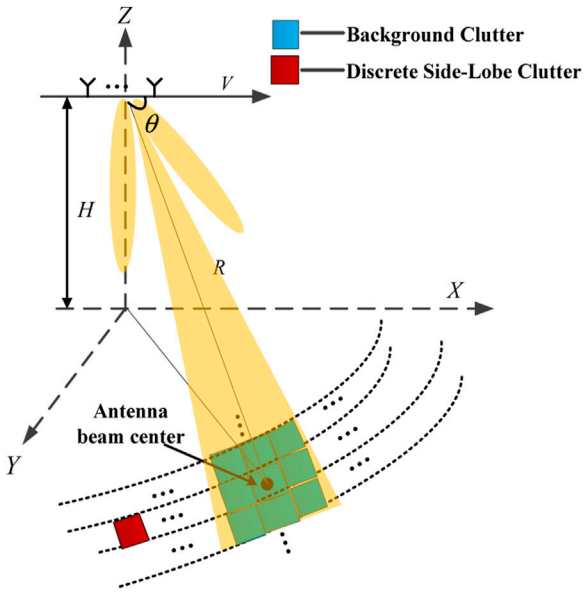


FIGURE 1
STAP model of SBR.

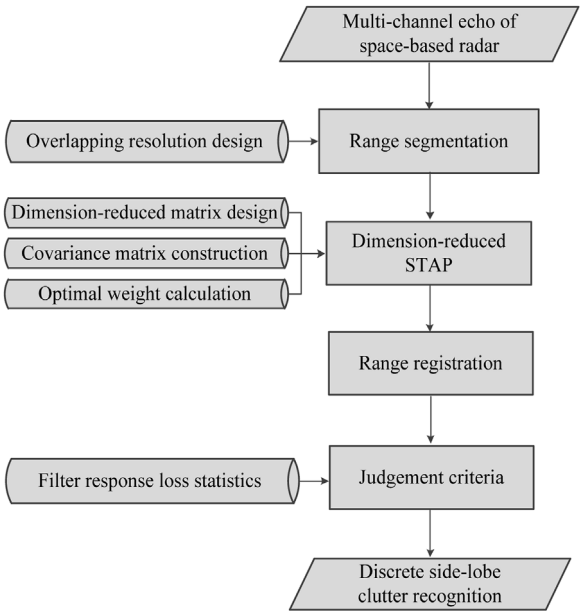


FIGURE 2
Flow chart of discrete side-lobe clutter recognition method.

2 STAP model

Space-based radar (SBR) generally operates under down-looking mode, and there exists strong ground clutter or sea clutter in the echo data. Besides, due to its wide coverage capability, a large number of isolated scattering points from the antenna side-lobe are collected by the receiver, as demonstrated in Figure 1. Thus, clutter suppression is important to meet the requirements of moving target detection.

TABLE 1 Airborne radar system parameters.

Parameters	Value	Parameters	Value
Wavelength	L	Platform velocity	100 m/s
Channel number	8	Bandwidth	25 MHz
Pulse number per CPI	250	PRF	2,500 Hz
Range gates	300	Platform height	3,000 m

Assume that C_l denotes the scatter number in the l th range bin, where $1 \le l \le L$. G_i and G_{tar} denote the complex amplitude of the i th scatter and the target, respectively. N_l represents the Gaussian white noise. Thus, the echo data of the l th range bin can be expressed as

$$\boldsymbol{x}_l = \sum_{i=1}^{C_l} \boldsymbol{G}_i \left(\boldsymbol{s}_t^i \otimes \boldsymbol{s}_s^i\right) + \boldsymbol{G}_{tar}\left(\boldsymbol{s}_t^{tar} \otimes \boldsymbol{s}_s^{tar}\right) + \boldsymbol{N}_l \tag{1}$$

where symbol $\otimes$ denotes the Kronecker product operation, and $j = sqrt(-1)$. Let the normalized temporal steering vector and spatial steering vector of clutter cell be $\boldsymbol{s}_t^i$ and $\boldsymbol{s}_s^i$ in turn. The antenna array is composed of N elements with an interval of d, and the number of pulses during one coherent processing interval (CPI) is K. Thus, the normalized temporal steering vector and spatial steering vector with regard to the target are given by

$$\boldsymbol{s}_t^{tar} = \left[1,\ exp\left[j2\pi\left(\frac{2V}{\lambda \cdot f_r} cos\,\theta_{tar} + \frac{2v}{\lambda \cdot f_r}\right)\right], ...,\right.$$
$$\left. exp\left[j2\pi\left(\frac{2V}{\lambda \cdot f_r} cos\,\theta_{tar} + \frac{2v}{\lambda \cdot f_r}\right)(K-1)\right]\right]^T \tag{2}$$

$$\boldsymbol{s}_s^{tar} = \left[1,\ exp\left[j2\pi\frac{d}{\lambda} cos\,\theta_{tar}\right], ...,\ exp\left[j2\pi\frac{d}{\lambda} cos\,\theta_{tar}(N-1)\right]\right]^T \tag{3}$$

In Eqs 2, 3, f_r stands for the pulse repetition frequency (PRF), V represents the platform velocity, λ is the signal wave length, θ_{tar} is the cone angle, v is the target velocity, exp and cos indicate the exponential operation and the cosine operation, respectively. The subscript T denotes the transpose operation. If $v = 0$, the steering vectors of clutter echo are consistent with that of the target. Instead, if $v \neq 0$, the temporal steering vector caused by the Doppler component will lead to the separation of the target and the echo data in space-time domain. However, the discrete side-lobe clutters come from the antenna side-lobe may result in similar distribution characteristics, compared with the real target, in which these candidate targets cannot be distinguished in the STAP process.

3 A discrete side-lobe clutter recognition method based on sliding filter response loss for space-based radar

For full-dimensional STAP methods, it is not conductive to real-time processing due to the high computational complexity in the order of $O(NK)^3$. Further, the number of independent identically distributed (IID) samples should be more than $2NK$ to minimize the optimal detection performance loss to less than 3 dB. However, the actual observation environment is difficult to meet this requirement, especially for the heterogeneous scene with massive discrete strong scattering points. In order to reduce the full-dimensional STAP computational complexity and achieve reasonable clutter suppression performance, the dimension-reduced strategy is employed. Assume that $\boldsymbol{T} \in \mathbb{C}^{NM \times PQ}$ is the dimension-reduced matrix, where P; Q denote the degrees of freedom (DOFs) with regard to the spatial domain and the temporal domain in turn. The adaptive weight based on the linear constraint minimum variance (LCMV) criterion can be rewritten as

$$\boldsymbol{w}_T = \boldsymbol{\mu}\boldsymbol{R}_T^{-1}\boldsymbol{s}_T \tag{4}$$

where $\boldsymbol{m} = \frac{1}{\boldsymbol{s}_T^H \boldsymbol{R}_T^{-1} \boldsymbol{s}_T}(\cdot)^{-1}$ is the matrix inverse operation. The dimension-reduced matrix $\boldsymbol{R}_T$ and the target steering vector s_T can be re-represented as

$$\boldsymbol{R}_T = \sum_{i=1}^{L} \left(\boldsymbol{T}^H \boldsymbol{x}_l\right)\left(\boldsymbol{T}^H \boldsymbol{x}_l\right)^H \tag{5}$$

$$\boldsymbol{s}_T = \boldsymbol{T}^H\left(\boldsymbol{s}_t^{tar} \otimes \boldsymbol{s}_s^{tar}\right) \tag{6}$$

Here, $(\cdot)^H$ stand for the conjugate transpose operation.

Without loss of generality, EFA methods are adopted in the dimension-reduced STAP. The filter response loss is defined as

$$\boldsymbol{Loss} = \frac{\boldsymbol{E\_out}}{\boldsymbol{E\_in}} \tag{7}$$

where E_in represents echo power of the target resolution cell before clutter suppression, while $\boldsymbol{E\_out}$ denotes that after clutter suppression. Let $\boldsymbol{x}$ be the echo data of the cell under detected (CUT) in range-Doppler domain, the above variables can be expressed as $\boldsymbol{E\_in} = |\boldsymbol{x}^H \boldsymbol{x}|$, $\boldsymbol{E\_out} = |\boldsymbol{w}_T^H \boldsymbol{x}|^2$.

Space-based radars possess wide coverage capability as its high orbit height, and the echo data shares strong fluctuation characteristics. As a result, the sliding window needs to be employed to segment the range rings in heterogeneous environment. On one hand, these resolution cells within a limited scene have relatively consistent distribution characteristics by means of range segmentation process. On the other hand, the number of available samples for constructing clutter covariance matrix (CCM) will be reduced after range segmentation strategy, which may lead to clutter suppression performance deterioration with regard to the scene edge resolution cells. In this paper, a discrete side-lobe clutter recognition method based on the sliding window is proposed for space-based radar system, in which both the sample number for building CCM and the filtering performance with regard to scene edge area are taken into account. Firstly, the sliding window is presented to segment the observation scene into the identical range intervals. Assuming these two sample sets are denoted as X_1 and X_2 before and after the sliding process, the corresponding CCMs are expressed as R_1 and R_2, and the optimal weight vectors are indicated as w_1 and w_2, respectively. Thus, the STAP results related to different range intervals are given by

$$\boldsymbol{result}_1 = \boldsymbol{w}_1^T \boldsymbol{X}_1 \tag{8}$$

$$\boldsymbol{result}_2 = \boldsymbol{w}_2^T \boldsymbol{X}_2 \tag{9}$$

For the two $result_1$ and $result_2$ generated by the sliding window, the range cells are registered according to mark the range cell number. In this way, the amplitudes of the resolution cells which

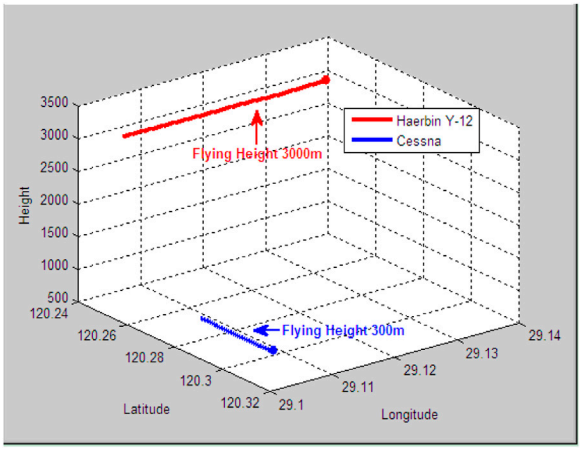


FIGURE 3
Target detection scheme design.

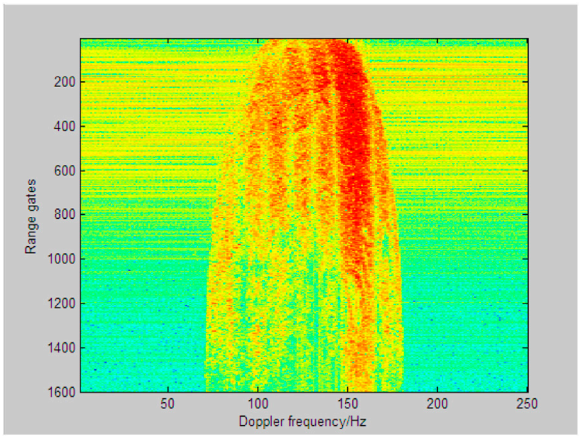


FIGURE 4
Range-Doppler data of channel 1.

exceed the CFAR threshold are added to the candidate target set [13]. Assume that judgment threshold of filter response loss is η. For any candidate target, the STAP filter response loss before and after the sliding process are $Loss_1$ and $Loss_2$, respectively, and the corresponding judgment criteria is given by

$$\text{candidate target} = \begin{cases} \text{discrete side} - \text{lobe clutters}, \ \min(\text{Loss}_1, \text{Loss}_2) > \eta \\ \text{discrete side} - \text{lobe clutters}, \{\max(\text{Loss}_1, \text{Loss}_2) > \eta > \min(\text{Loss}_1, \text{Loss}_2)\} \\ \qquad \&\& \left\{ \dfrac{\max(\text{Loss}_1, \text{Loss}_2) + \min(\text{Loss}_1, \text{Loss}_2)}{2} > \eta \right\} \\ \text{real target} \qquad , \{\max(\text{Loss}_1, \text{Loss}_2) > \eta > \min(\text{Loss}_1, \text{Loss}_2)\} \\ \qquad \&\& \left\{ \dfrac{\max(\text{Loss}_1, \text{Loss}_2) + \min(\text{Loss}_1, \text{Loss}_2)}{2} \le \eta \right\} \\ \text{real target} \qquad , \ \max(\text{Loss}_1, \text{Loss}_2) \le \eta \end{cases} \tag{10}$$

The background clutter suppression as well as the target energy accumulation can be achieved by employing the optimal weight vector when conducting STAP algorithm. Generally, the filter energy loss of the moving target is small, while that of the discrete side-lobe clutter is large. However, the moving target is not necessarily located in the antenna beam center, and the suppression degree of discrete side-lobe clutter show randomness due to the heterogeneous environment, which resulting in the limited fault tolerance with regard to filter response loss method for distinguishing the real targets from discrete side-lobe clutters [12].

Space-based radar possesses wide area coverage and complex regional distribution, in which the discrete side-lobe clutters mainly generate from the heterogeneous area with discrete strong scattering points. Considering that the samples corresponding to different range segments present quite different characteristics, the category judgment results of these samples are uncertain for a specific sample discrimination criterion. In this paper, the sliding window strategy is proposed to process the segmented echo data. As the clutter covariance matrices before and after the sliding window operation are constructed by adopting different sample sets, the filter response loss value in regard of a discrete side-lobe clutter suppression is also different. Thus, the maximum filter response loss of each discrete side-lobe clutter can be statistically obtained based on the above sample sets. Besides, the real target is located in the main antenna lobe direction, as a result, the target steering vectors are identical for different sample sets, and the filtering response losses of the real target is relatively small compared with discrete side-lobe clutters. If the candidate target within the CUT is a discrete side-lobe clutter, which comes from the antenna side-lobe direction, the target steering vector has a limited constraint ability to this area. However, the suppression ability to the discrete side-lobe clutter shows great fluctuation with regard to different samples sets. That is, for a specific candidate target, the significance of energy loss characteristics can be enhanced through analyzing the filter response losses. Conversely, if the candidate target within the CUT is a real target, which comes from the antenna main-lobe direction, the target steering vector has a strong constraint ability to this region. As the real target and its competitive clutter have obvious differences in spatial domain, the filter response losses of the real target can be maintained at a small level. Therefore, the proposed filter response loss method based on the given sliding window can effectively improve the diversity between the discrete side-lobe clutter and the real target. Especially for heterogeneous environment, this method has extremely robust recognition ability for discrete side-lobe clutters. Figure 2 demonstrates the flow chart of discrete side-lobe clutter recognition method.

4 Experimental results

4.1 Airborne scheme design

In this section, the effectiveness of the proposed method is validated by utilizing the real measured data with an airborne radar system. Radar system parameters are shown in Table 1.

In order to simulate the SBR operating mode more realistically, a drone aircraft is designed to fly below the

10.3389/fphy.2023.1142154

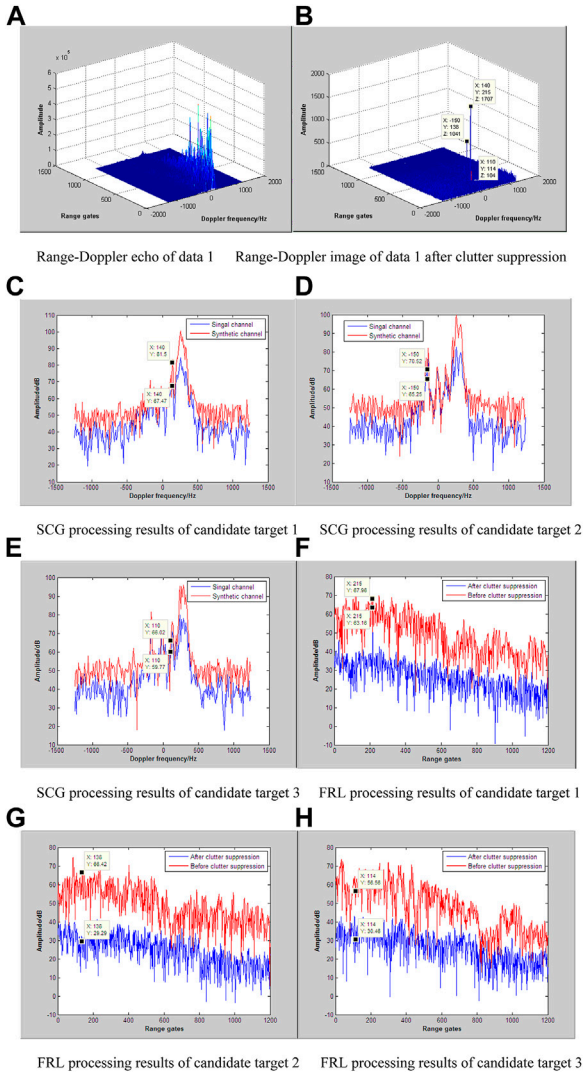


FIGURE 5
Processing results of data 1.

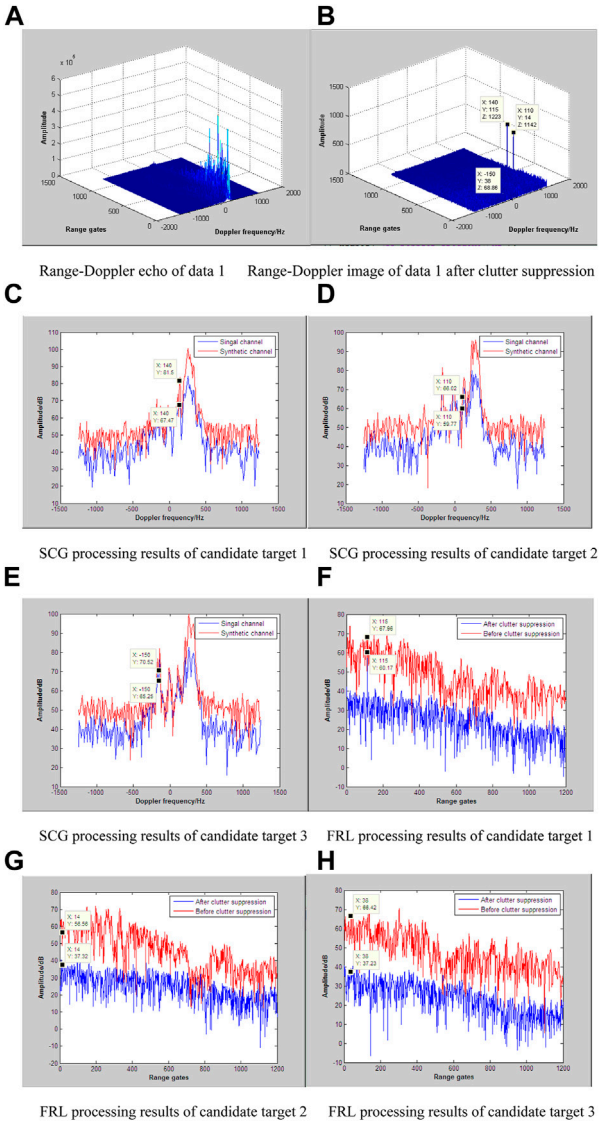


FIGURE 6
Processing results of data 2.

carrier aircraft, as shown in Figure 3. Here, Haerbin Y-12 and Cessna are selected as the carrier aircraft and the drone aircraft, respectively. Figure 4 shows the Range-Doppler data of channel 1 with 250 pulses. One can see that the interferences and discrete terrain clutters exist in most of the Range-Doppler image. As a result, the measured data is composed of real target echo, real clutter echo and real interference echo, with which the performance of the proposed method could be varified reliably.

4.2 Clutter suppression and discrete side-lobe clutter recognition

In this section, FSA algorithm is adopted as the dimension-reduced method for STAP. In order to achieve robust clutter suppression performance, the heterogeneous samples are eliminated when constructing the CCM. As a result, the real target and discrete side-lobe clutters may coexist in the image after clutter suppression. During the data processing procedure, the range segment unit is set to 400, and that of the overlapping range units in regard of adjacent data sets is set to 100. For the two data sets before and after the sliding window strategy, the corresponding clutter suppression results is shown in Figures 5, 6 respectively. From Figures 5B, 6B, there exist two candidate targets for different data set, where the candidate target coordinates of the first data segment are given by (140, 215), (−150, 138), and those of the second data segment are given by (140, 115), (110, 14). Considering that only one drone aircraft is arranged in the experiment, it is necessary to identify the categories of the candidate targets. By employing the range registration process, one candidate target is generated in each image, which are indicated as (110, 114), (−150, 38) in turn. Meanwhile, the candidate points (140, 215) within the first image and the candidate points (140, 115)within the second image has a

TABLE 2 Processing results of candidate targets with different methods.

	SCG (dB)	FRL (dB)	SFRL (dB)
Candidate target 1	14.0	6.3	7.8
Candidate target 2	5.3	27.1	29.2
Candidate target 3	6.3	19.2	26.1
Fault tolerance capability	7.7	12.9	18.3

good match after image registration. Therefore, there are three candidate targets in the observation scene. Figures 5C, 5H, 6C, 6H represent the processing results of these candidate targets with the Synthetic channel gain (SCG) method, the filter response loss (FRL) method and the sliding filter response loss (SFRL) method. Here, the sliding window size is set to 400 range cells, and the overlapping range number between adjacent sliding windows is set to 100.

The data processing results of the real measured data are shown in Table 2. According to the judging criteria, candidate target one is recognized as a real target, candidate target two and candidate target three are recognized as discrete side-lobe clutters. Among the above three methods, SFRL method has the best fault tolerance performance, while SCG method has the worst fault tolerance performance. Here, the fault tolerance capability is defined as the desirable dynamic interval of the decision threshold, that is, it can be regarded as the reliable threshold setting interval to distinguish the real target from the discrete side-lobe clutter.

5 Conclusion

For space-based radar, a discrete side-lobe clutter recognition method based on sliding filter response loss is demonstrated to separate the targets from abundant discrete side-lobe clutters in heterogeneous environment, in which the filter response losses corresponding to different segmented data are calculated, and thus the fault tolerance capability for real target recognition can be effectively enhanced. Meanwhile, the insufficient clutter suppression ability at the scene edge areas is significantly alleviated by means of the range segmentation strategy. Theoretical analysis and experimental results based on real measured data verify the reliability of the proposed method.

Data availability statement

The datasets presented in this article are not immediately available because of sensitive information. Requests to access the datasets should be directed to the corresponding author.

Author contributions

Conceptualization YL investigation JC methodology YL and WY Visualization WW validation QL writing-original draft, JC writing-review and editing CL and CD All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the National Natural Science Foundation of China (NNSFC) (No. 62107033) and the National Defense Science and Technology Foundation Strengthening Plan Fund (No. 2022-JCJQ-JJ-0202) and the Sustainedly Supported Foundation by National Key Laboratory of Science and Technology on Space Microwave (No. HTKJ2022KL504004) and the CAST Foundation of the Fifth Academy (No. Y22-CASTJJ-02) and the Outstanding Youth Foundation of the Fifth Academy, Aerospace Science and Technology Group (No. Y22-RCWYJQ1-01).

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

1. Li HY, Bao WW, Hu JF, Xie J, Liu R. A training samples selection method based on system identification for STAP. *Signal Process.* (2018) 142(3):119–24. doi:10.1016/j.sigpro.2017.07.008

2. Wu YF, Wang T, Wu JX, Duan J. Robust training samples selection algorithm based on spectral similarity for space-time adaptive processing in heterogeneous interference environments. *IET Radar Sonar & Navig* (2015) 9(7):778–82. doi:10.1049/iet-rsn.2014.0285

3. Guo JJ, Liao GS, Yang ZW, Du WT. Iterative weighted covariance matrix estimation method for STAP based on generalized inner products. *J Electro Inf Technol* (2014) 36(2):422–7. doi:10.3724/SP.J.1146.2013.00426

4. Yang XP, Liu YX, Hu XN. Robust generalized inner products algorithm using prolate spheroidal wave functions. In: Radar Conference (RADAR) on Aerospace, Components and Signal Processing; 2012; Atlanta, US (2012).

5. Sun GH, He ZS, Tong J, Zhang X. Knowledge-aided covariance matrix estimation via kronecker product expansions for airborne STAP. *IEEE Geosci Remote Sensing Lett* (2018) 15(4):527–31. doi:10.1109/lgrs.2018.2799329

6. Wen C, Peng JY, Zhou Y, Wu JX. Enhanced three-dimensional joint domain localized STAP for airborne FDA-MIMO radar under dense false-target

jamming scenario. *IEEE sensors J* (2018) 18(10):4154–66. doi:10.1109/jsen.2018.2820905

7. Duan CD, Li Y, Wang WW. An intelligent sample selection method for space-time adaptive processing in heterogeneous environment. *IEEE Access* (2019) 7(1):30321–30. doi:10.1109/access.2019.2902218

8. Jing H, Hu MK, Wang ZW. A improved knowledge-aided space time adaptive signal processing algorithm for MIMO radar. *J Electro Inf Technol* (2019) 41(4):795–800. doi:10.11999/JEIT180557

9. Shnidman DA, Toumodge SS. Sidelobe blanking with integration and target fluctuation. *IEEE Trans Aerospace Electron Syst* (2002) 38(3):1023–37. doi:10.1109/taes.2002.1039418

10. Narasimhan RS, Engadarajan AV, Ramakrishnan KR. Mitigation of sidelobe clutter discrete using sidelobe blanking technique in airborne radars. In: IEEE Aerospace Conference; 2018; Big Sky, US (2018).

11. Tian M, Yang ZW, Dang HX, Xu HJ, Huang PH. A two-step detector based on point spread function feature for multi-channel SAR-GMTI radar. In: 2016 CIE International Conference on Radar; 2016; Guangzhou, China (2016).

12. Wang WW, Duan CD, Zhang X. A discrete side-lobe clutter recognition method using space-time steering vectors for space-based radar system. *J Electro Inf Technol* (2020) 42(11):2592–2599. doi:10.11999/JEIT190562

13. He Y, Guan J, Meng XW. *Radar target detection and CFAR processing*. Beijing: Tsinghua University Press (2011). p. 40–50.

frontiers | Frontiers in Ecology and Evolution

TYPE Original Research
PUBLISHED 02 March 2023
DOI 10.3389/fevo.2023.1121189

OPEN ACCESS

EDITED BY
Huadan Zheng,
Jinan University, China

REVIEWED BY
Chong Jiang,
Guangdong Provincial Academy of Chinese Medical Sciences, China
Yongchun Zhou,
Northeastern University, China

*CORRESPONDENCE
Jiahong Zhang
✉ zhangjiahong@mail.cgs.gov.cn
Haiyan Zhang
✉ zhanghaiyan@igsnrr.ac.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Environmental Informatics and Remote Sensing, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 11 December 2022
ACCEPTED 06 February 2023
PUBLISHED 02 March 2023

CITATION
Guo Y, Liu X, Liu X, Zhang J, Zhang H, Fan J, Khan N and Ma J (2023) Quantitative assessment of the degree of harmony between humanity and nature for national parks in China: A case study of the Three-River-Source National Park. *Front. Ecol. Evol.* 11:1121189. doi: 10.3389/fevo.2023.1121189

Quantitative assessment of the degree of harmony between humanity and nature for national parks in China: A case study of the Three-River-Source National Park

Yanming Guo[1,2,3†], Xiaojie Liu[1†], Xiaohuang Liu[1,4], Jiahong Zhang[2*], Haiyan Zhang[1,3*], Jiangwen Fan[1], Nawab Khan[5] and Jiliang Ma[6]

[1]Key Laboratory of Natural Resource Coupling Process and Effects, Ministry of Natural Resources, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China, [2]China Aero Geophysical Survey and Remote Sensing Center for Natural Resources, Beijing, China, [3]College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, China, [4]Natural Resources Comprehensive Investigation Command Center, China Geological Survey, Beijing, China, [5]College of Management, Sichuan Agricultural University, Chengdu, China, [6]Institute of Agricultural Economics and Development, Chinese Academy of Agricultural Sciences, Beijing, China

Introduction: National parks, defined as the mainstay of the nature reserve system in China, pursue to achieve scientific protection and rational utilization of natural resources. However, eco-environmental and socioeconomic benefits are rarely considered together. Hereby, how to quantitatively express the relationship between humanity and nature in national parks needs further exploration. We selected the Three-River-Source National Park (TRSNP), China's largest national park by area and the world's highest altitude national park, as a representative case to construct an evaluation model for the degree of harmony between humanity and the nature of the national alpine ecological park.

Methods: Based on the field survey data, the meteorological data, the remote sensing data, and the socioeconomic data, the study used the model inversion and the spatial analysis methods to quantitatively evaluate the degree of harmony consisting of 12 indexes from a fresh perspective of a combination of the ecological environment and social economy. Considering the TRSNP establishment in 2016 as the time node, we assessed and compared the degree of harmony between humanity and nature during the dynamic baseline period (2011–2015) and the evaluation period (2016–2020).

Results: The results show that the degree of harmony between humanity and nature showed a gradual upward trend from the northwest to the southeast in the TRSNP. Compared with the dynamic baseline period, the eco-environmental and socioeconomic levels of the evaluation period were increased by 34.48 and 5.46%, respectively. Overall, the degree of harmony between humanity and nature visibly increased by 23.38%.

Discussion: This study has developed a novel comprehensive method for evaluating national parks at the regional scale for the win-win goal of both protection and development, and it provides a theoretical basis for effective planning and management policies for national parks.

KEYWORDS

harmony between humanity and nature, quantitative assessment, nature reserve, sustainable development, Three-River-Source National Park

1. Introduction

China is one of the nations with the highest levels of biodiversity (Gorenflo et al., 2012; Ma et al., 2017), with approximately 10% of higher plants (Ren et al., 2019) and 22% of terrestrial vertebrates (Xue and Zhang, 2019). With rapid economic and social developments, the natural ecosystem faces a conflict between the protection of natural resources and their development and utilization. Natural reserve systems that are the core carriers of ecological constructions in China can offer protection to maintain biodiversity, natural landscapes, and vital ecosystems (Juffe-Bignoli et al., 2014; Ren and Guo, 2021), while the traditional nature reserve model has been unable to meet the current developmental requirements (Rennie, 2006). Therefore, a new nature reserve system was established considering both ecological protection and rational utilization (Liu et al., 2021). As the mainstay of the nature reserve system in China, a national park is a specific land or marine area established and managed by the state with clearly defined boundaries and aimed mainly to protect the authenticity and integrity of significant natural ecosystems over a large area (General Office of the CPC and Central Committee, 2017). The establishment goals of national parks in China not only strengthen the protection of the authenticity and integrity of the natural ecosystem but also emphasize benefits to society, which coordinate the relationship between ecological environmental protection and the production and life of farmers and herdsmen. The ultimate purpose of national parks is to maintain the harmonious coexistence and sustainable development of humanity and nature, to build a nature reserve system with Chinese characteristics with national parks as its core, and to promote the development of a beautiful China [National Development and Reform Commission (NDRC), 2018; Xinhua News Agency, 2022]. International national parks are also focused on the relationship between nature conservation and local economic development (Zhu et al., 2022). For example, France, one of the first countries in the world to create national parks, during the construction of its national parks paid much attention to the synergistic development of cities and parks to build a sustainable national park development model (Yang, 2022). Significantly, China's national parks permit indigenous people to engage in production and operation activities during the park's construction and build a mechanism for joint construction and sharing (Zhao, 2019). Specifically, national parks are not only important ecological function areas but also production and living spaces for locals. Against the backdrop of building China into a great modern socialist nation in all respects, a set of national park construction ideas harmonizing the ecological, production, and living considerations are in demand. This requires narrowing the striking contrast between economic growth and environmental degradation (Ouyang et al., 2020) and achieving the harmonious coexistence of humanity and nature. Consequently, the scientific and exhaustive assessments of the changes in the harmony between humanity and nature brought about by the establishment of China's national parks are valuable for guiding the future construction and governance of national parks and facilitating the development of ecological civilization.

Since the United States set up Yellowstone National Park in 1872 as the world's first national park, more than 5,000 national parks have been established in more than 100 countries and regions (Zhang et al., 2022). Currently, the evaluation research on national parks often separates the eco-environmental and socioeconomic evaluations. The ecological aspects include the evaluation of ecosystem integrity and authenticity (Jiang et al., 2021; Zhao, 2021), ecological sensitivity (Yang et al., 2017; Meng et al., 2018), ecological health (Su, 2019), and ecosystem service value (Li et al., 2016; Zhang and Zhang, 2019), whereas the social aspects mainly include the evaluation of national park management efficiency (Rudnicki et al., 2005; Heiland et al., 2020), resident livelihood flexibility (Meng and Chen, 2019), and resident sustainable livelihoods (Shang and Cao, 2019; Yu et al., 2020). The evaluation research also includes an assessment of economic aspects such as suitability for research and study travel (Chen et al., 2020) and for recreational use (Xiao et al., 2019). In this type of research, the calculation methods of the weight of the indicator system are mainly divided into the subjective weighting method [Delphi method (Li and Cong, 2021); the analytic hierarchy process (AHP) (Tang et al., 2010); the expert scoring method (Yang and Zhu, 2016)]; and the objective weighting methods including the entropy weight method (EWM) (Zhou et al., 2021) and the gray clustering method (Lu and Li, 2018). In addition, the approaches used are the Rapid Assessment and Prioritization of Protected Area Management (Nchor and Ogogo, 2013) and the Enrichment Evaluation Combined with a Stochastic Multi-Criteria Acceptability Analysis approach and a preference-ranked organizational approach (PROMETHEE) (An et al., 2019). Although these evaluation methods are classic and have many practical bases, this kind of evaluation work needs further exploration.

The Three-River-Source region is a significant ecological security barrier and plateau biological germplasm resource bank, which is of high importance to China and even the world (Fan et al., 2011). Due to its vital ecological status, existing studies of effectiveness evaluations in this region focus on the ecological aspect without considering the balanced development of a fundamental social economy. For instance, the effectiveness evaluation of the ecological environmental protection and construction projects in the Three-River-Source region concluded that the ecological environment had partially improved and the deterioration had been contained initially (Shao et al., 2016). Other small-scale eco-environment assessments of the Three-River-Source National Park (TRSNP) mainly include a comparative analysis of the changes in grass yields and livestock pressure for the grassland ecosystems (Zhang et al., 2014), health status assessment of the wetland ecosystems (Jia et al., 2011), and ecological health levels within the ecological protection comprehensive experimental zone (Huang, 2017). As China's largest and world's highest-altitude national park, a pilot TRSNP was first established in 2016 and then officially set up the first batch of national parks in 2021 (Zhang et al., 2022). The evaluation research objects have gradually shifted to the scale of the TRSNP, but the quantitative evaluation research is still just focused on ecology. For example, Cao et al. (2019) explored the spatiotemporal differentiation characteristics of the ecological functions through GIS spatial analysis and other means;

Su (2019) assessed the ecological health of the TRSNP utilizing the PSR model and AHP; Fu et al. (2021) established the indicator system for the evaluation of the eco-environmental protection effects of the TRSNP. The evaluation of the socioeconomic aspects and inclusive benefits of the TRSNP is still at the qualitative analysis level (Zhao et al., 2020; Zhao and Yang, 2021). In conclusion, the quantitative evaluation research of the TRSNP scale needs further exploration to improve the quality of the scale. Simultaneously, to meet the profound purpose of the harmonious coexistence of humanity and nature pursued by the construction of national parks in China, it is essential to express the relationship between the ecological environment and social economy. At the method level, due to the inconsistency between the simulation scale of the eco-environmental index and the statistical caliber of the socioeconomic index, the question of how to quantitatively express the relationship between ecological conservation and community development needs to be answered.

This study takes the TRSNP as a case study area to bridge the current research gaps. We intend to develop a quantitative method for evaluating the degree of harmony between humanity and nature in pilot or official national parks of China with a focus on the balance between ecological conservation and community development objectives. First, this study designs the degree of harmony between humanity and nature (DHHN) index based on the eco-environmental level (EEL) and the socioeconomic level (SEL). To realize this method, we broke through the limitations of administrative boundaries to integrate the socioeconomic data from the Statistical Yearbook with the eco-environmental data in the same spatial and temporal resolutions. Moreover, EWM and AHP were used to comprehensively calculate the index weight. Then, considering the TRSNP establishment in 2016 as the time node, we analyzed the changes in the spatiotemporal variation characteristics of EEL, SEL, and DHHN during the dynamic baseline period (2011–2015) and the evaluation period (2016–2020). Finally, we provide some recommendations for future layouts and designs of national parks to achieve both ecological protection and local development goals.

2. Materials and methods

2.1. Study area

The TRSNP is in the Southern Qinghai Province (32°23′–36°48′ N, 89°51′–99°15′ E) and in the hinterland of the Qinghai–Tibet Plateau (Figure 1). Its total area is 123,100 km^2, involving the four counties, Qumalai, Zhiduo, Zaduo, and Maduo, and the Hoh Xil Nature Reserve, with a total of 12 towns and 53 administrative villages. The TRSNP is mainly characterized by mountain ranges and alpine canyons, including the Kunlun Mountains, the Bayan Har Mountains, and the Tanggula Mountains with an average elevation of more than 4,500 m and a maximum elevation of 6,726 m. Due to its high altitude and hinterland location, the TRSNP has a typical plateau continental climate with the annual average temperatures between −5.6 and 7.8°C and annual precipitation between 170 and 700 mm. Grassland, mainly including alpine meadows and alpine steppe, occupies 61.54% of the study area and is the dominant ecosystem type. In 2020, the average GDP of the TRSNP was 801.44 million yuan, with a total population of 168,000.

The TRSNP is a vital repository of natural capital and ecosystem service flows for a substantial portion of Qinghai and China (Ouyang et al., 2019). There are several rivers, swamps, and lakes in the territory, of which 167 have an area of more than 1 km^2 (Figure 1). It provides approximately 40 billion m^3 of water annually downstream and is known as the "water tower" of East and Southeast Asia (Shao et al., 2016). The TRSNP is the source of the Yangtze, Huanghe, and Lancang rivers, correspondingly including three parks: the Yangtze River Source Region (YRSR), the Huanghe River Source Region (HRSR), and the Lancang River Source Region (LRSR). The TRSNP is also a global biodiversity hotspot and a repository for alpine biological germplasm, including 760 kinds of vascular plants and 270 kinds of wild vertebrates. Based on the goal of overall ecosystem protection and system restoration, the functional regionalization of the TRSNP is divided into core protection areas and general control areas.

2.2. Data source

In this study, the data used include basic geographic, meteorological, remote sensing, field survey data, and socioeconomic data (Table 1) from 2011 to 2020. Basic geographic data, including the Digital Elevation Model (DEM), the administrative division boundary, and the TRSNP boundary, were obtained from the relevant official management departments. The meteorological data were provided by 107 weather stations from the China Meteorological Science Data Sharing Service Network (http://cdc.cma.gov.cn/) and 167 weather stations from the local weather department. The remote sensing data were obtained from the MODIS and mainly include NDVI (Normalized Difference Vegetation Index) and NPP (Net primary productivity). The LUCC (Land Use and Land-Cover Change) data were from the 1:100,000 scale land use dynamic database established by Liu et al. (2014) from the late 1980s to 2020 and included six first-level types and 25 second-level types of forest land, grassland, cultivated land, industrial and mining land, residential land, urban and rural lands, unused land, and water bodies. Through field sampling and lab analysis, our research team obtained 475 1 m × 1 m quadrats at the peak of vegetation growth from mid-July 2010 to mid-August 2020. The population density data and the GDP (gross domestic product) density data were from the existing datasets downloaded from the Resource and Environmental Science Data and the Center of the Chinese Academy of Sciences (https://www.resdc.cn). Other socioeconomic data were from the Statistical Yearbooks of the provinces and counties and the Local Livestock House.

2.3. Evaluation method for the degree of harmony between humanity and nature

2.3.1. Construction of an evaluation indicator system

Based on the borrowed experience and regional background features, we designed an indicator system for evaluating the degree

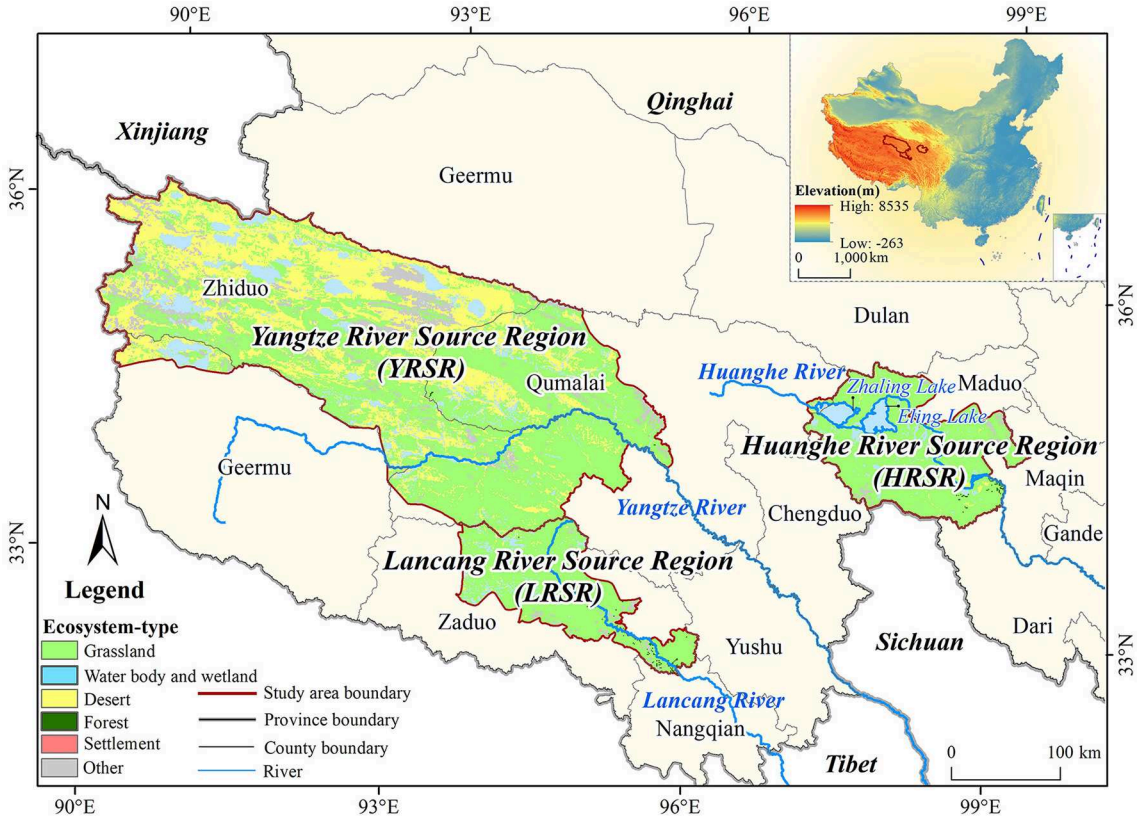


FIGURE 1
Overview of the Three-River-Source National Park (TRSNP).

of harmony between humanity and nature (DHHN) for the TRSNP that includes two aspects: the eco-environmental level (EEL) and the socioeconomic level (SEL). The EEL is to evaluate the core purpose of establishing the national parks system and the SEL is to evaluate the construction of an ecological environment. The ultimate purpose is to realize the harmonious coexistence of humanity and nature. Therefore, the constructions of the EEL and SEL indicators should always focus on protecting the authenticity and integrity of the TRSNP natural ecosystem and representing and reflecting the social and economic developments of the TRSNP as comprehensively as possible, respectively.

First, the subsequent indicator screening process should adhere to the following guidelines: (1) scientifically, referring to the experts' research and scholars in related fields combined with field research to screen the indicators systematically; (2) pertinence, screening indicators according to the purpose of the TRSNP construction and the regional characteristics; (3) accessibility, which requires a relatively mature means of acquiring the data required by the monitoring system for national parks in China (Jiao et al., 2022); (4) quantifiability, requiring each indicator to obtain quantitative results through scientific operations to ensure the practical significance of the evaluation system. According to the aforementioned principles, this study finally selected two categories, three levels, and twelve indexes involving the ecological environment and social economy to build a DHHN system for the pilot implementation of the TRSNP system (Table 2). Referring to the TRSNP-related government documents, this study considers the pilot establishment of the TRSNP in 2016 as an important time node and selects 5 years before and after and considers the dynamic baseline period as 2011–2015 and the evaluation period as 2016–2020.

The EEL has three characteristics, namely ecosystem qualities, ecosystem services, and landscape pattern metrics. For ecosystem qualities, fractional vegetation cover (FVC) reflects the vegetation density, the size of the photosynthetic area, and the growth status of the vegetation community (Mu et al., 2012). The Net Primary Productivity (NPP) can represent the efficiency of plant fixation, conversion of photosynthesis products, and determining the available substances and energy (Zhu et al., 2007). For ecosystem services, water conservation, the core ecological service of the TRSNP, reflects the comprehensive function produced by the interaction of vegetation, water, and soil (Gong et al., 2017); soil retention reflects the ability of the ecosystems to reduce soil erosion (Hu et al., 2014); and wind erosion prevention is one of the most important ecosystem services in arid and semi-arid regions and reflects the ability of the region to retain soil and prevent dust storms (Gao et al., 2013). For landscape pattern metrics, habitat fragmentation has become the primary threat to biodiversity and Shannon's diversity index (SHDI) can represent the degree of habitat fragmentation at the landscape level. The greater the SHDI, the better the landscape fragmentation degree, the key reason for biodiversity loss (Ding et al., 2005).

TABLE 1 Data collection for DHHN assessment.

Data category	Specific criteria	Spatial resolution	Temporal resolution	Format	Source
Basic geographic data	Administrative division boundary	N/A	Latest	Vector	National Basic Geographic Information Center (http://www.ngcc.cn/ngcc/)
	TRSNP boundary	N/A	N/A	Vector	Administration bureau of Three-River-Source National Park
	DEM	1 km	Yearly	Raster	Resource and Environmental Science Data and Center of Chinese Academy of Sciences (https://www.resdc.cn)
Meteorological data	Temperature	N/A	Daily	Text	China Meteorological Science Data Sharing Service Network (http://cdc.cma.gov.cn/)
	Precipitation	N/A	Daily	Text	China Meteorological Science Data Sharing Service Network (http://cdc.cma.gov.cn/)
	Relative humidity	N/A	Daily	Text	China Meteorological Science Data Sharing Service Network (http://cdc.cma.gov.cn/)
	Radiation	N/A	Daily	Text	China Meteorological Science Data Sharing Service Network (http://cdc.cma.gov.cn/)
	Wind direction/speed	N/A	Daily	Text	China Meteorological Science Data Sharing Service Network (http://cdc.cma.gov.cn/) and local weather departments,
Remote sensing data	NDVI	250 m	16-day	Raster	MOD13Q1, NASA (http://www.gscloud.cn/)
	NPP	250 m	16-day	Raster	MOD17A3HGF, NASA (http://www.gscloud.cn/)
	LUCC	1:100,000	5-yearly	Raster	Land use dynamic database established by Liu et al. (2014)
Field sampling data	Aboveground biomass	1 m	Yearly	Excel	Our research team
	Belowground biomass	1 m	Yearly	Excel	Our research team
	Vegetation coverage	1 m	Yearly	Excel	Our research team
	Grassland type	1 m	Yearly	Excel	Our research team
Socio-economic data	Population density data	1 km	5-yearly	Raster	Resource and Environmental Science Data and Center of Chinese Academy of Sciences (https://www.resdc.cn)
	GDP density data	1 km	5-yearly	Raster	Resource and Environmental Science Data and Center of Chinese Academy of Sciences (https://www.resdc.cn)
	Population employed in the tertiary industry	County	Yearly	Excel	Statistical Yearbook of provinces and counties
	Savings deposit balance of residents	County	Yearly	Excel	Statistical Yearbook of provinces and counties
	Livestock number	County	Yearly	Excel	Local Livestock House
	Beds number of medical and health institutions	County	Yearly	Excel	Statistical Yearbook of provinces and counties

N/A, not applicable.

Among the SEL indicators, population density is an important indicator of the population distribution of a region (Wu, 2011); medical service capacity enjoyed by the unit population reflects the patient acceptance ability of the medical institutions in a region, and the essential social development goals of the TRSNP are to build more high-quality and efficient medical service capacities (Xie et al., 2018). The population employed in the tertiary industry reflects the economic structure and the industrial development of the tertiary industry in a region (Liu, 2008), the gross domestic product (GDP) density is an important indicator of a region's economic situation and level of development (Liang and Xu, 2013), the savings deposit balance of the residents reflects the actual economic status and consumption capacity of the residents in a region (Zhu and Cheng, 2016), and the livestock density reflects the intensity of grassland grazing of the TRSNP animal husbandry production (Wu et al., 2021) and puts great pressure directly on the grassland ecosystem (Wei et al., 2020) in the TRSNP.

2.3.2. Data processing and standardization

2.3.2.1. Processing of the eco-environmental data

We primarily checked the meteorological data to control for quality and replace missing values with observations from nearby areas. Then, they were interpolated in space to 1 × 1 km

TABLE 2 The DHHN index system of the Pilot Three-River-Source National Park System.

Target layer	Criterion layer	Indicator layer	Unit	Effect direction
Eco-environmental level	Ecosystem qualities	FVC	%	Positive
		NPP	g·C/m^2	Positive
	Ecosystem services	Water conservation	m^3/km^2	Positive
		Soil retention	t/ha	Positive
		Wind erosion prevention	t/ha	Positive
	Landscape pattern metrics	SHDI	–	Negative
Socio-economic level	Social situation	Population density	person/km^2	Negative
		Medical service capacity	pcs./person	Positive
		Population employed in tertiary industry	person/km^2	Positive
	Economic situation	GDP density	10,000 yuan/km^2	Positive
		Savings deposit balance of residents	10,000 yuan/km^2	Positive
		Livestock density	SHU/km^2	Negative

grid data by the optimized thin-plate smooth spline function of the ANUSPLIN software with DEM. The meteorological spatial data after interpolation were regarded as the key intermediate parameters of water conservation, soil retention, and wind erosion prevention.

The FVC was calculated by the dimidiate pixel model based on the NDVI data after projection transformation, mosaicking, cropping, maximum value synthesis, S-G filtering, etc. After batch pretreatment, the NPP value was obtained by adding up to 23 periods of data over the whole year. The ecosystem services, i.e., water conservation, soil retention, and wind erosion prevention, were calculated using the precipitation storage method, the revised universal soil loss equation (RUSLE), and the revised wind erosion equation (RWEQ), respectively, which were shown in the Supplementary material (Figure 2). The ground sampling data were used to verify the FVC and NPP values in this study. Considering the spatial distribution of quadrats and the representativeness of grassland types, the quadrat data are strictly screened and standardized before modeling. The remaining quadrats may reflect the main grassland types and multiyear grass yields in the quadrat representative area. SHDI was used to calculate the LUCC data by the Fragstats software.

2.3.2.2. Processing of the socioeconomic data

Some socioeconomic data at the county level, such as the livestock number and the number of beds in the medical and health institutions, are obtained from the Statistical Yearbook. While the function zone boundaries of the TRSNP do not coincide with the administrative division boundaries of the counties and townships, the statistical data also do not reflect the spatial differences to a certain extent and reduce the accuracy of the evaluation.

In this study, we selected the number of beds in the medical and health institutions to represent the medical service capacity. First, using network crawling, the bed counts of all the medical and health institutions of the four counties related to the TRSNP are obtained; however, the impact of the counties other than the four counties is not considered. Second, we queried the relevant information of each medical institution to get the establishment time and the corresponding medical institution in the *medical institutions' classification management standard*. Then, according to the *National Medical and Health Service System Planning Outline*, the medical coverage theory capacities of the medical and health institutions of different levels in the set buffer (50 km, 100 km, 150 km, 200 km, and 500 km) were defined (Table 3) and the theoretical capacities of the medical service institutions in the various regions of the TRSNP were obtained. Using the theoretical capacities of the medical service institutions in different regions, the bed counts of the medical and health institutions in the four counties from 2011 to 2020 can be rostered and then divided by the population density of the TRSNP to obtain the spatial distribution data on the actual medical service capacity (Figure 3). The technical process figure is shown in the Supplementary material. The medical service capacity is estimated using the following equations:

$$NBMI_{ij} = NBMI_{cau} \times \frac{MSTA_{ij}}{MSTA_{cau}} \quad (1)$$

$$MSC_{ij} = \frac{NBMI_{ij}}{POP_{ij}}, \quad (2)$$

where $NBMI_{ij}$ is the bed count of the medical and health institutions (pcs.) in the grid in row i and column j with the grid size 1 km × 1 km; $NBMI_{cau}$is the statistical value of the bed count (pcs.) of the medical and health institutions in the county-level administrative regions where the grid unit is located; $MSTA_{ij}$ is the theoretical medical service capacity (%) of the grid unit; $MSTA_{cau}$ is the theoretical total medical service capacity (%) of the county-level administrative unit that contains the grid unit; MSC_{ij} is the medical service capacity (pcs./person) of the grid in row i and column j; and POP_{ij} is the population of that grid unit (person).

As the distribution of livestock quantity has a strong correlation with the altitude, settlements, water sources, and grassland conditions (Qiao et al., 2017), this study selected the distance from settlements, altitude, NPP, and distance from a water source as the main influencing factors of the spatial distribution of livestock. In addition, due to the complete prohibition of productive animal husbandry activities in the core conservation area and the implementation of a strict balance of grass and livestock in the general area

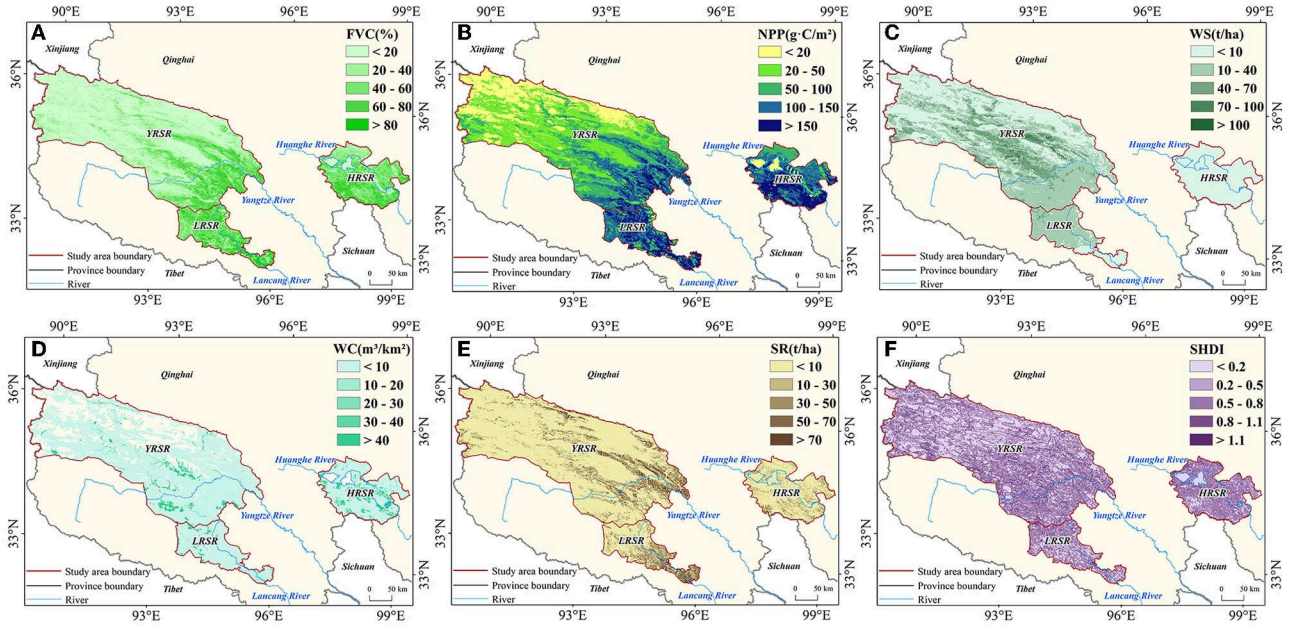


FIGURE 2
Spatial distribution of mean values of the ecological data from 2000 to 2020. **(A)** Fractional vegetation cover (FVC); **(B)** Net primary productivity (NPP); **(C)** Water conservation (WC); **(D)** Soil retention (SR); **(E)** Wind erosion prevention (WEP); and **(F)** Shannon's diversity index (SHDI).

TABLE 3 Theoretical coverage of medical services in different medical and health institutions at different distances.

Distance (km)	0–50	50–100	100–150	150–200	200–500	>500
Coverage capacity of primary hospitals (%)	100	70	50	25	10	5
Coverage capacity of secondary hospitals (%)	100	70	50	25	10	5

[National Development and Reform Commission (NDRC), 2018], this study distributed the number of livestock after 2016 in the remaining area except for the core conservation area. First, the four influencing factors were rasterized on the grasslands and normalized by calculating the deviating Euclidean distance. At the same time, a hierarchical model is constructed and the weights of the seven influencing factors are obtained by expert scoring. By multiplying the normalized results of different factors with the corresponding weight, the livestock activity intensity scores of different grids can be obtained, and by combining them with the livestock quantities of the four counties from 2011 to 2020, the livestock quantity distribution of the TRSNP can be obtained (Figure 4). The technical process figure is shown in the Supplementary material. The calculation formula is as follows:

$$LD_{ij} = LD_{cau} \times \frac{LAI_{ij}}{LAI_{cau}}, \quad (3)$$

where LD_{ij} is the livestock number (SHU) of the grid in row i and column j with the grid size 1 × 1 km; LD_{cau} is the statistical value of the livestock number (SHU) at the county level; LAI_{ij} is the livestock intensity of this grid; and LAI_{cau} is the total livestock intensity at the county level in which that grid unit is located.

2.3.2.3. Data standardization

The DHHN assessment was performed both at the 1-km spatial resolution and the annual temporal resolution. All data involved in the evaluation system were unified at the temporal and spatial resolutions, the coordinate system, and the data format. Furthermore, there are dimensional differences between the various evaluation indicators of the DHHN. Each index value uses extreme difference analysis for the standardization of the treatment to 0–1. For positive indicators, the calculation formula is as follows:

$$X_{i,j} = \frac{x_{i,j} - x_{i,\min}}{x_{i,\max} - x_{i,\min}} \quad (4)$$

For negative indicators, the calculation formula is as follows:

$$X_{i,j} = \frac{x_{i,\max} - x_{i,j}}{x_{i,\max} - x_{i,\min}}, \quad (5)$$

where $X_{i,j}$ represents the standardized value of indicator i in year j; $x_{i,j}$ represents the actual value of indicator i in year j; and $x_{i,\min}$ and $x_{i,\max}$ represent the minimum and maximum values observed in all actual measurements of indicator i during 2011–2020.

2.3.3. Weight determination of the indices

Determination of index weight is important for final evaluation

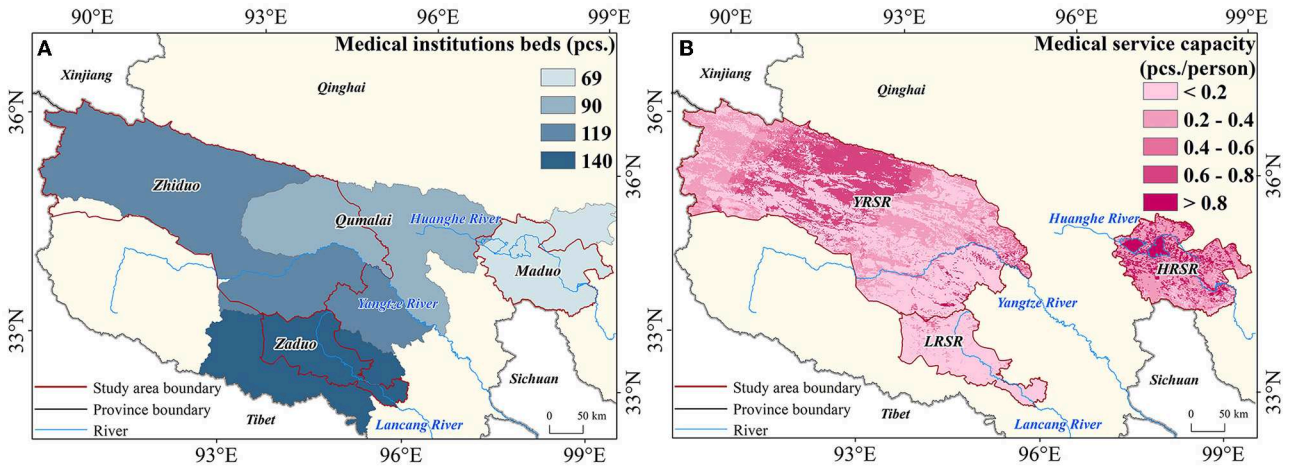


FIGURE 3
Comparison of the bed count of the medical and health institutions at the county level **(A)** and the 1 × 1 km level **(B)**.

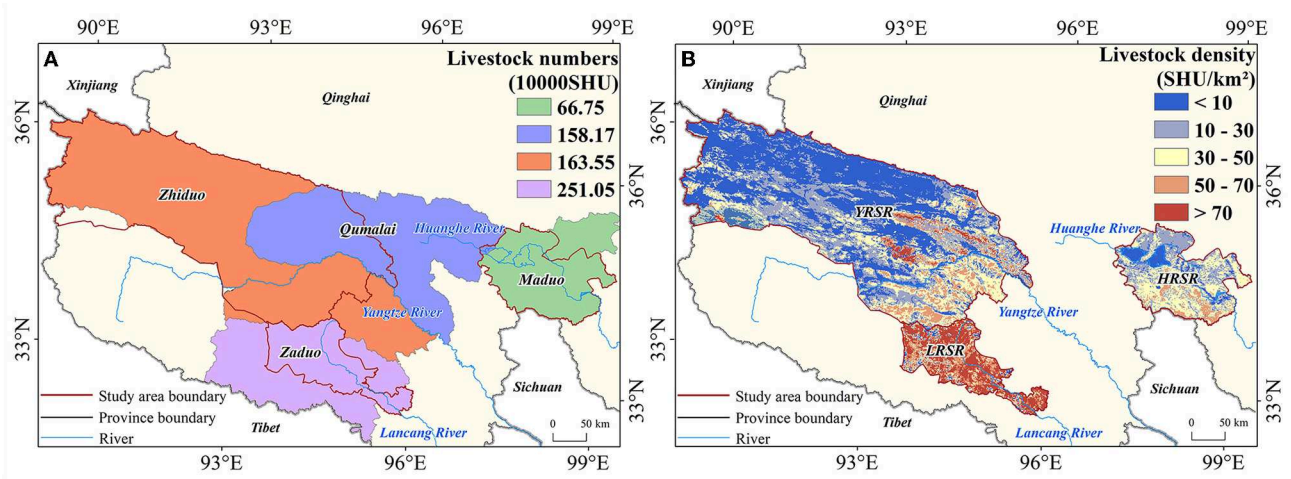


FIGURE 4
Comparison of the livestock number at the county scale **(A)** and the 1-km scale **(B)**.

results and this study uses EWM and AHP to determine the weight of each indicator. The EWM is an objective weighing approach that quantifies the indicator's weight by calculating the indicator's information entropy (Ding et al., 2017). When the indicator information entropy is larger, it means that the system carries less information and corresponding indicator weight is smaller (Chen et al., 2009; Chen, 2019). While this approach can eliminate the influence of man-made subjective consciousness compared with the subjective weighting method, the disadvantage is due to the lack of horizontal comparison among the indicators. While the weight completely depends on the data samples with high accuracy requirements, AHP combines the qualitative and quantitative approaches.It constructs each influencing factor of the complex system a multilevel analysis structure model, analyzes and calculates the various levels, and then obtains the weights (Chang and Jiang, 2007). The advantage of this method is that the calculation is concise and clear and more attention should be paid to the essence of the research problem. It has fewer requirements for data but its disadvantages have strong subjectivity. Therefore, AHP and EWM are used together to provide each other the data and make the final weight calculation more scientific and reasonable.

For the EWM to calculate the proportion of the value of the indicator j of sample i:

$$p_{ij} = \frac{X_{i,j}}{\sum_{i=1}^{n} X_{i,j}} \tag{6}$$

For calculating the entropy of the indicator j:

$$e_{ij} = -\frac{1}{\ln n} \sum_{i=1}^{n} p_{ij} \ln p_{ij} \tag{7}$$

For calculating the weight of the indicator j:

$$w_j = \frac{1-e_j}{\sum_{j=1}^{m}(1-e_j)} \tag{8}$$

For the AHP, a judgment matrix is constructed by comparing the two indicators at the same level by expert scoring and then constructing the judgment matrix of the level:

$$U = \begin{bmatrix} u_{11} & \cdots & u_{1n} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nn} \end{bmatrix}, \quad (9)$$

where U is the judgment matrix; u_{ij} is the importance value of u_i relative to u_j; 1–9 and its reciprocals are the importance scales; and n is the number of indicators at this level.

Weight calculation by the eigenvector method is performed by calculating the eigenvector corresponding to the maximum characteristic root based on the judgment matrix and then normalizing it to obtain the weight vector. The calculation formula is as follows:

$$U_w = \alpha_{\max} W, \quad (10)$$

where W is the eigenvector; U_w is the judgment matrix with the eigenvector W; and $\alpha_{\max}$ is the maximum characteristic root.

The consistency test is performed by checking whether the weight vector obtained passes the consistency test. First, the general consistency index of the judgment matrix is calculated using the following formula:

$$CI = \frac{\alpha_{\max} - n}{n-1}, \quad (11)$$

where CI is the general consistency index of the judgment matrix and n is the number of factors in the AHP structure.

Second, the average random consistency index RI corresponding to the number of factors was found (Table 4):

Finally, the random consistency ratio (CR) of the judgment matrix was calculated using the formula

$$CR = \frac{CI}{RI} \quad (12)$$

It is considered that the consistency of the judgment matrix U is acceptable when the calculated CR of the judgment matrix (U) is <0.1. Otherwise, it is necessary to make an appropriate adjustment to the judgment matrix (U) to make its consistency meet the requirements.

After calculating the weights of the evaluation indicator system using the EWM and AHP values, the final weights of the evaluation indicator system are determined by averaging the weights obtained from the two methods (Table 5).

2.3.4. Model for the degree of harmony between humanity and nature

The connotation of the harmonious degree of humanity and nature (HDMN) model multiplies the weights of the indicator system by the standardized value of the corresponding indicators and then accumulates the corresponding EEL, SEL, and DHHN values to carry out subsequent comparison and analysis. The model can be mathematically expressed as follows:

$$EEL = \sum_{i=1}^{m} (EW_i \times EI_i) \quad (13)$$

$$SEL = \sum_{j=1}^{n} (SW_j \times SI_j) \quad (14)$$

$$HDMN = Cw_e \times EEL + Cw_s \times SEL \quad (15)$$

where EW_i is the weight of each indicator of the ecological environment; SW_j is the weight of each indicator of social economy; Cw_e is the weight of EEL; Cw_s is the weight of SEL; EI_i is the result of the standardization of the eco-environmental indicators; and SI_j is the result of the standardization of socioeconomic indicators.

3. Results

3.1. Spatiotemporal variation characteristics of eco-environmental level

In terms of spatial distribution, the EEL of the TRSNP in the baseline and evaluation periods showed a trend of gradual increase from the northwest to the southeast (Figures 5A, B). The EELs of the two periods were best in the LRSR, followed by the HRSR; the YRSR has the worst EEL. Among them, in the baseline period, the EEL of the LRSR was 0.61, while those of the HRSR and the YRSR were 0.52 and 0.28, respectively, and the EEL of the entire TRSNP was 0.47. During the evaluation period, the EELs of the LRSR, the HRSR, and the YRSR were 0.72, 0.64, and 0.35, respectively, and the EEL of the entire TRSNP was 0.63.

From the perception of time change, according to the statistical results, the study area where the EEL of the TRSNP increased accounted for 40.03% of the study area. Mainly, in the southeast, the EELs of 5−20% and 4.10% regions increased by 35.93% and >20%. The EEL of the remaining regions was unchanged. The largest area accounted for 53.34% change and the smallest area accounted for 6.63% In the northwest region of the TRSNP, −20 to −5% of the regions accounted for a 6.61% reduction. A reduction >-20% accounted for only 0.02% (Figure 5C). Among the three regions, the YRSR has more areas with reduced or unchanged EEL in the western region, while the areas with EEL increasing gradually increase eastward; most of the EEL of the LESR is increasing, and a few areas with a decrease are scattered among them. The HRSR showed that the EEL of the southern regions mostly increased and it remained unchanged or gradually increased along the northern regions (Figure 5C).

3.2. Spatiotemporal variation characteristics of the socioeconomic level

From the perspective of spatial distribution, the SEL of the TRSNP in the baseline and evaluation periods are similar to those of the EEL, showing a gradually increasing trend from the northwest to the southeast (Figures 6A, B). The difference is that the SEL can be seen to have strong county characteristics. In both periods, the LRSR in Zaduo had the best SEL, followed by the HRSR in Maduo, and the worst SEL was the YRSR in Zhiduo and Qumalai; however, the SEL of Qumalai was better than Zhiduo. In the baseline period, the SEL of the LRSR was 0.84, while the SEL of the HRSR and the YRSR were 0.48 and 0.26, respectively, and the SEL of the whole TRSNP was 0.53. During the evaluation period, the SEL of the LRSR, the HRSR, and the YRSR were 0.90, 0.43, and 0.33, respectively, and the SEL of the whole TRSNP was 0.56.

In terms of the time change, the SEL of 48.82% of the total study area increased which is mainly distributed in the

TABLE 4 RI values for the different number of factors.

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
RI	0	0	0.52	0.89	1.12	1.26	1.36	1.41	1.46	1.49	1.52	1.54	1.56	1.58	1.59

TABLE 5 Weight corresponding to every indicator of the evaluation indicator system.

Target layer	AHP	EWM	Average	Index layer	AHP	EWM	Average
Eco-environmental level	80.00%	59.26%	69.63%	FVC	10.95%	8.22%	9.58%
				NPP	15.79%	11.15%	13.47%
				Water conservation	38.88%	14.21%	26.54%
				Soil retention	4.59%	4.76%	4.68%
				Wind erosion prevention	5.90%	7.40%	6.65%
				SHDI	3.89%	13.52%	8.70%
Socio-economic level	20.00%	40.74%	30.37%	Population density	2.14%	11.23%	6.69%
				Medical service capacity	4.79%	2.01%	3.40%
				Population employed in tertiary industry	1.05%	8.67%	4.86%
				GDP density	3.16%	6.78%	4.97%
				Savings deposit balance of residents	7.27%	1.37%	4.32%
				Livestock density	1.59%	10.67%	6.13%

Zhiduo county area of the YRSR, that of 35.59% increased by 5–20%, which is mainly distributed in the LRSR, and the SEL of 13.23% of the total study area increased by more than 20%. The unchanged area accounted for 28.47% of the TRSNP and 22.71% of the YSSR mainly distributed in the Qumalai County, and −20–5% of the areas of the HRSR decreased by 15.81% and that of −20% decreased by 6.90% (Figure 6C).

3.3. Quantitative assessment of the degree of harmony between humanity and nature

In terms of spatial distribution, the DHHN of the TRSNP during the baseline and evaluation periods show a gradual growth trend from the northwest to the southeast and from the north to the south (Figures 7A, B). Among the three regions, the one with the largest DHHN during the baseline and evaluation periods was the LRSR followed by the HRSR. The YRSR had the smallest DHHN. During the baseline period, the DHHN of the LRSR is 0.68 and those of the HRSR and the YRSR were 0.51 and 0.27, respectively. The DHHN of the TRSNP was 0.49. During the evaluation period, the DHHN of the LRSR was 0.84 and those of the HRSR and the YRSR were 0.63 and 0.33, respectively. The DHHN of the TRSNP was 0.60. Among the main ecosystems of the TRSNP, the grassland ecosystem had the largest DHHN during both periods (the baseline period: 0.52, the evaluation period: 0.66) followed by the water body and wetland ecosystems (the baseline period: 0.28, the evaluation period: 0.33). The desert ecosystem had the smallest DHHN (the baseline period: 0.16, the evaluation period: 0.18) (Figure 8A).

From the perspective of the temporal variation, the DHHN of the whole TRSNP evaluation period increased by 23.38% compared with the baseline period. The areas where the DHHN had increased account for 53.87% of the TRSNP and are mainly distributed in the southeast of TRSNP. A total of 36.45% of the area underwent an increase in the range of 5–20% and 17.42% underwent an increase in the range of more than 20%. A total of 37.42% of the areas that have constant DHHN are mainly distributed in the northwest region. The 8.71%, 8.68%, and 0.03% of the areas with reduction are scattered in the northwest of the TRSNP, accounting for −20–5%, and by more than −20%. The growth rate of the DHHN of the YRSR is 21.93% which was mainly distributed in the southeast, while most areas in the northwest show a constant decreasing trend. Many areas that underwent a decrease in DHHN are located in Qumalai. The DHHN of almost the whole area of LRSR had increased and that during the evaluation period was increased by 24.05% compared with the baseline period. The DHHN of the HRSR is 23.26%, the area of comprehensive performance growth gradually decreases, and the areas undergoing constant decrease gradually increase from north to south, among which the DHHN of the Zhaling lake and the Eling Lake decrease significantly (Figure 7C). Among the three main ecosystem types, the grassland ecosystem with the best DHHN has the largest growth rate of 26.55%, the water body and wetland ecosystem types have a growth rate of 17.84%, and the desert ecosystem type with the worst comprehensive performance has the smallest growth rate of 10.84% (Figure 8B).

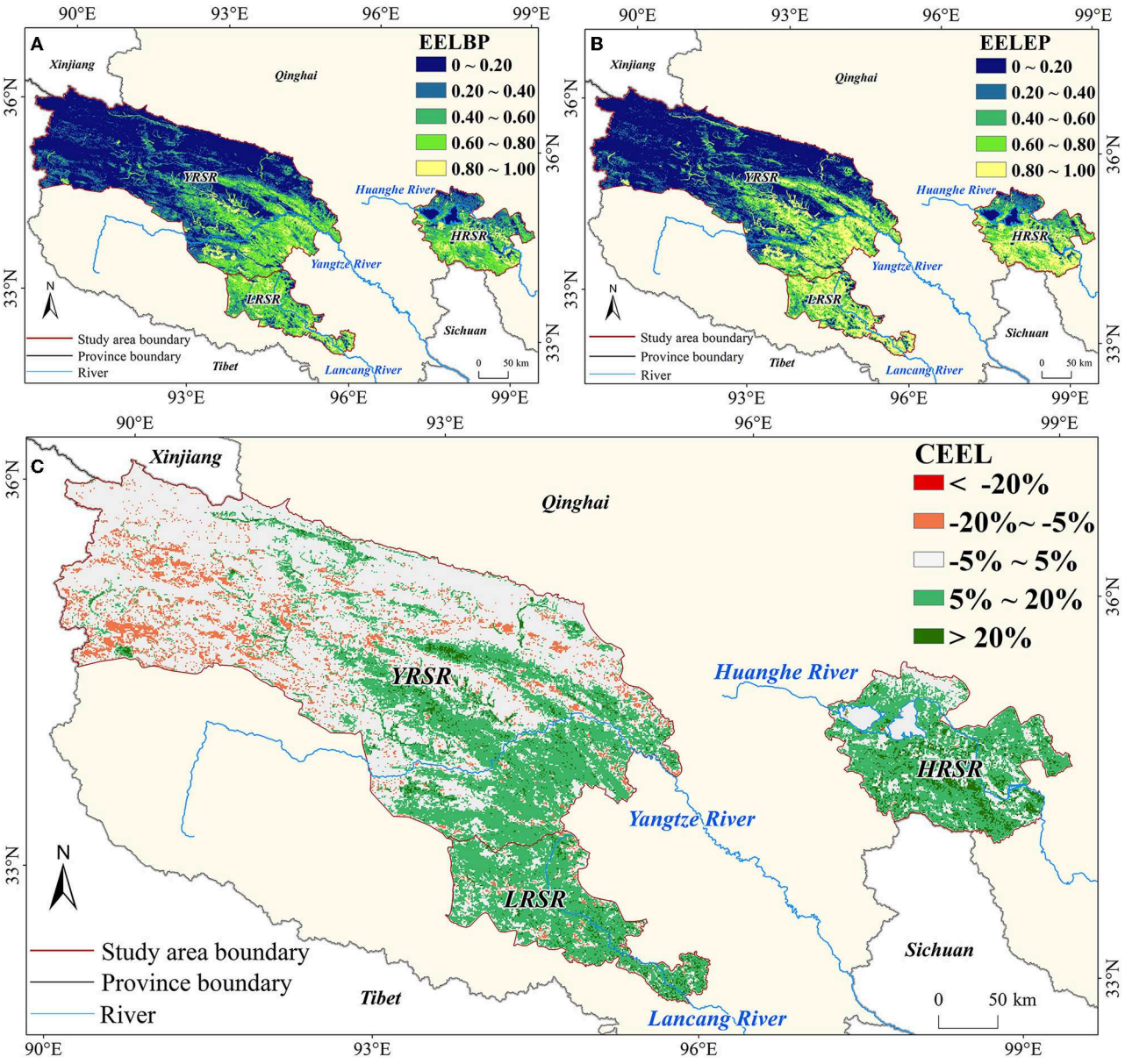


FIGURE 5
Distribution and change of the eco-environmental level (EEL) in the TRSNP. Eco-environmental level during the baseline period (EELBP) **(A)**; Eco-environmental level during the evaluation period (EELEP) **(B)**; Change in the eco-environmental level (CEEL) **(C)**.

4. Discussion

4.1. Implications of the spatiotemporal variation characteristics of the EEL, SEL, and DHHN

The assessment of the implementation effectiveness of establishing national parks is often used as a basis for the management organization to make further decisions (Yuh et al., 2019). From the outcomes of the EEL, SEL, and DHHN, the aforementioned three evaluation indicators during the baseline and evaluation periods showed a gradually increasing trend. This shows that the quality of the eco-environment continues to recover and improve, but the socioeconomic and social situation also improves simultaneously. Long-term overgrazing leads to serious grassland degradation, resulting in a loss of biodiversity and ecosystem services (Xin, 2014; Zhang and Jin, 2021). Since 2000, China has undertaken a variety of ecological projects to restore overgrazed and degraded grasslands, maintain forests and wetlands, and restore watershed ecosystem services, i.e., *the Letter on Please Consider Establishing the Qinghai Three-River-Source Nature Reserve as Soon as Possible*, *the Master Plan for the Ecological Protection and Construction of the Three-River-Source Nature Reserve in Qinghai*, and *the Overall Plan of the Qinghai Three-River-Source National Ecological Protection Comprehensive Experimental Zone*. Although research shows that natural factors,

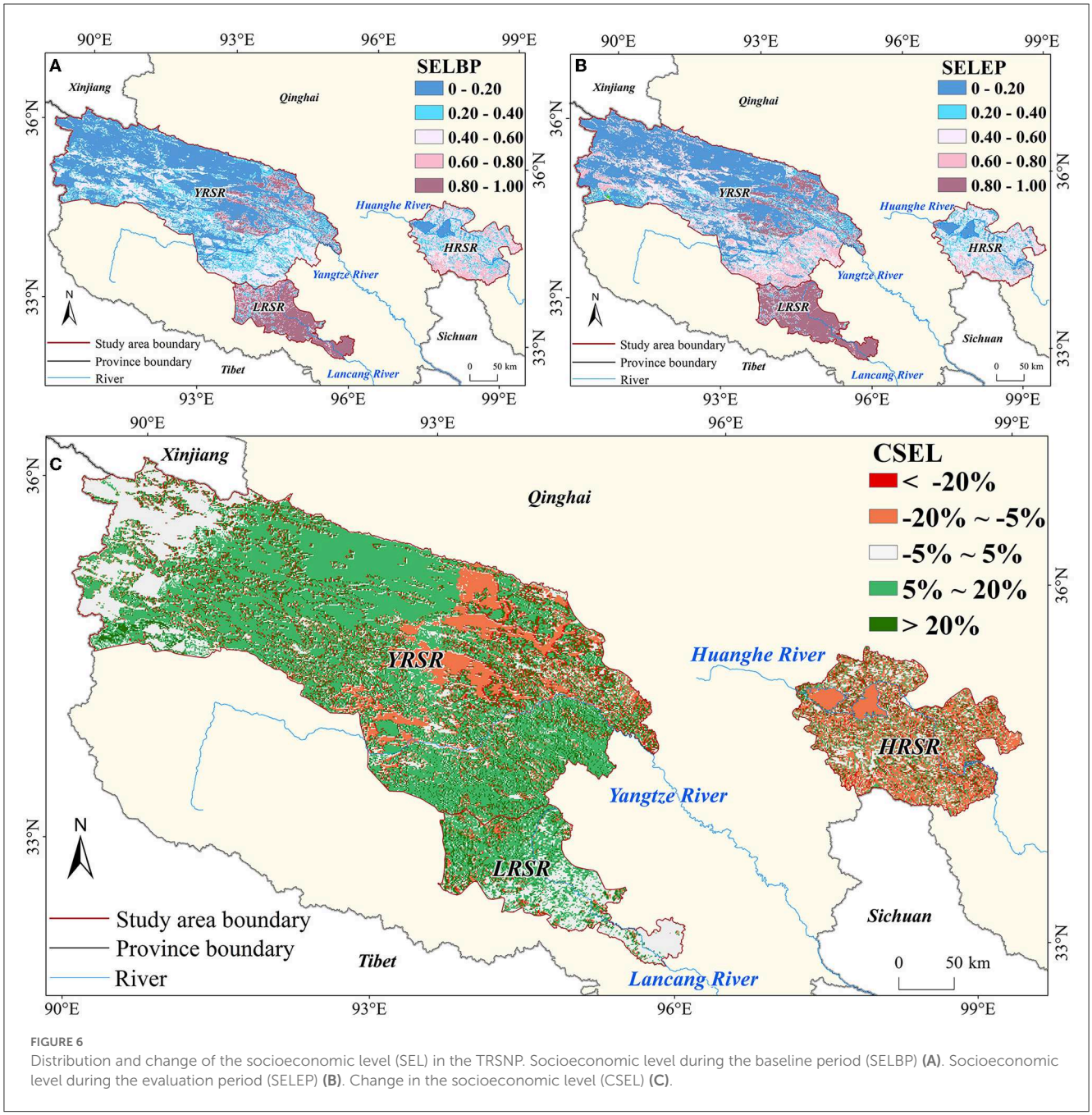


FIGURE 6
Distribution and change of the socioeconomic level (SEL) in the TRSNP. Socioeconomic level during the baseline period (SELBP) **(A)**. Socioeconomic level during the evaluation period (SELEP) **(B)**. Change in the socioeconomic level (CSEL) **(C)**.

such as climate warming have contributed to a certain extent to the improvement of various ecological indicators in the TRSNP, the contribution of humans to improving the harmony between humanity and nature is also crucial (Shao et al., 2016). The greater increase in the combined benefits of the indicators in the southeast region than in the northwest region of the park with more human intervention may also indicate a positive human role. The TRSNP innovatively implements the "one household, one post" ecological management and protection post system, covering all herders in the region. As of 2020, there were 16,621 herdsmen and 17,211 ecological management personnel in the region. The goal of the harmonious coexistence of humanity and nature has made new progress.

We analyzed the spatiotemporal variation characteristics of the EEL, SEL, and DHHN in the TRSNP. The aforementioned three evaluation indicators showed a consistent spatial distribution pattern from the northwest to the southeast in space. The areas with increases in the aforementioned three evaluation indicators were mainly concentrated in the southeast, while those with no change or a decrease were distributed in the northwest of TRSNP. Specifically, the aforementioned three evaluation indicators of the TRSNP are spatially manifested in areas with better social and economic developments and the baseline status and improvements in the ecological environment are also in a better state simultaneously. This is consistent with Liu et al. (2021) conclusion that the intensity of anthropogenic disturbance and the NDVI, NPP, and GPP of the

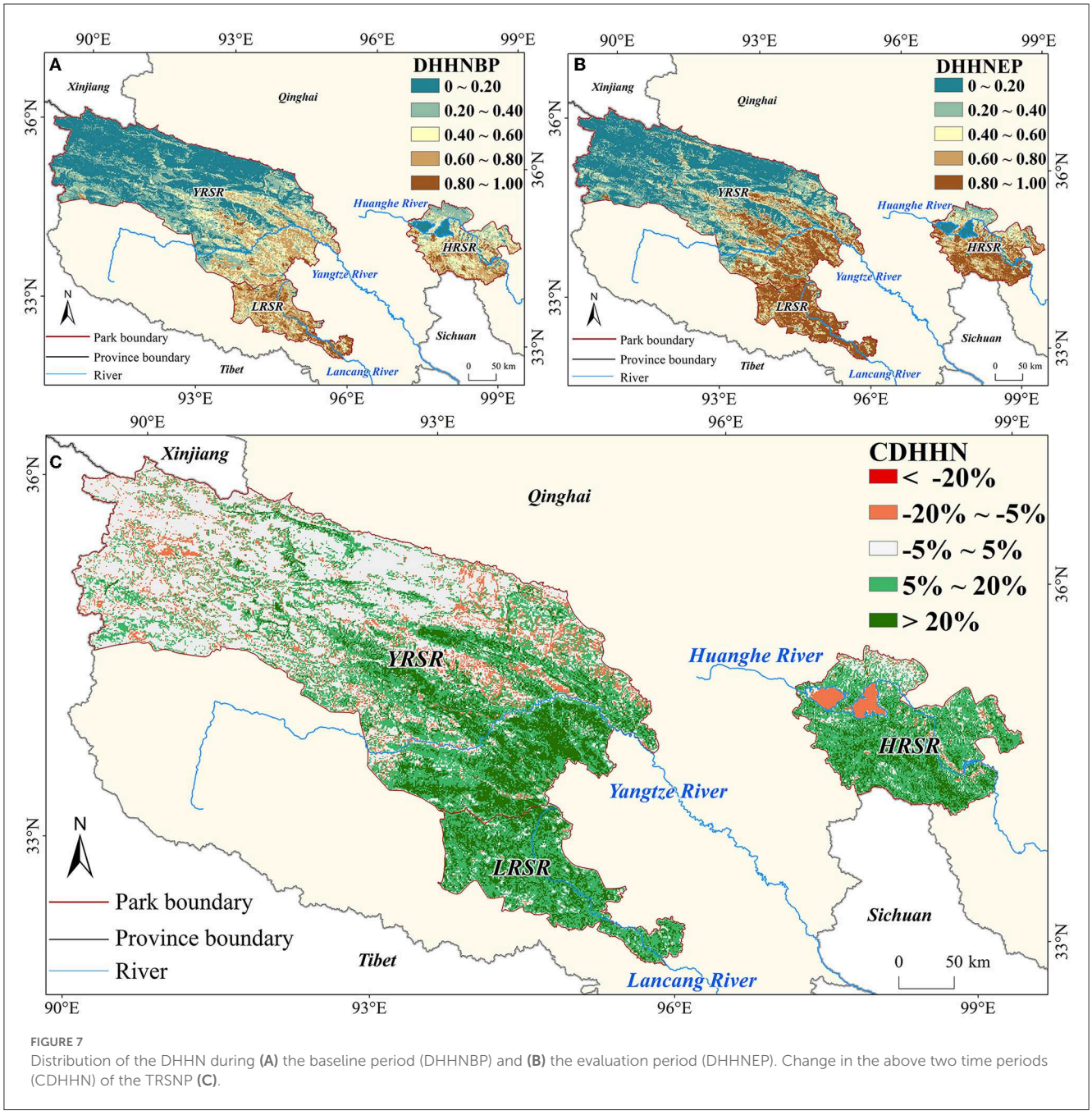


FIGURE 7
Distribution of the DHHN during **(A)** the baseline period (DHHNBP) and **(B)** the evaluation period (DHHNEP). Change in the above two time periods (CDHHN) of the TRSNP **(C)**.

Three-River-Source region all gradually increased from northwest to southeast. The reason for this result may be that the southwest monsoon of the Indian Ocean can travel northward along the Hengduanshan Mountain canyon up to the southeast of the TRSNP in the summer, making it a relatively humid area with better climatic conditions such as water and heat on the Qinghai–Tibet Plateau and an important distribution area in the alpine meadows. In addition, this region's altitude is lower than that of the northwest, and its ecological environment is in relatively good condition. It belongs to an area with a concentration of farmers and herdsmen. Therefore, the local government and residents pay more attention to improving the ecological environment and socioeconomic conditions. Consequently, a situation of simultaneous preference for the ecological environment, socioeconomic conditions, and improvement of the southeastern part of the TRSNP has been formed.

4.2. Suggestions for the sustainable development of national parks

The result comparisons of the EEL, SEL, and DHHN of the TRSNP during the baseline and evaluation periods show that the levels of the three aspects have increased by more than 20%. The TRSNP has initially completed the pilot task. However, it is worth noting that the absolute value of the three evaluation

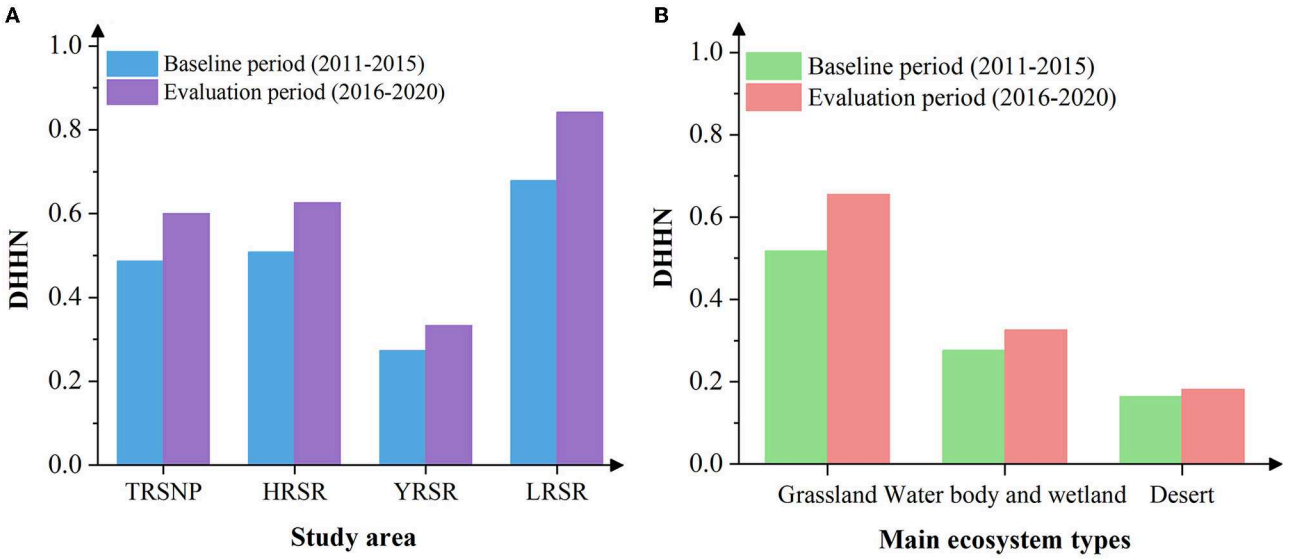


FIGURE 8
Comparison of the DHHN of different areas **(A)** and different ecosystem types **(B)**.

indicators is not high, especially as the difference between the YRSR and the other two regions is still large. Moreover, some eco-environmental and socioeconomic indicators fluctuate or even decrease. Regarding the eco-environmental indicators, windbreak and sand fixation in the TRSNP during the evaluation period decreased compared to that during the baseline period (Figure 9). Some studies (Cao et al., 2019) also showed that wind erosion prevention in the TRSNP decreased from 2000 to 2015, which may be attributed to the decrease in wind speed in the TRSNP and a decline in vegetation coverage in some areas. In addition, from the collected statistical data, the study found that the GDP growth rate and population growth rate in some years of the socioeconomic indicators also decreased, which may be due to the government's failure to properly coordinate the balanced relationship between the protection of the ecological environment and the socioeconomic development during the early stages of the TRSNP pilot system.

Thus, it can be seen that the harmonious relationship between humanity and nature needs to be further optimized in this region. The win–win vision of protection and development still requires the joint efforts of all parties, the government, and the residents (Ouyang et al., 2021). Therefore, the construction of the TRSNP still needs continuous investment and persistence. In addition to focusing on the improvement and upgrading of the aforementioned reduced aspects, it is also necessary to take upgrading measures according to different regional developmental conditions. For example, in the YRSR, due to the low absolute EEL value, priority must be given to the protection and regional ecological environment development, especially in the northwest of the region, and appropriate ecological conservation and restoration measures must be taken to improve the region's ecology. In the LRSR, due to the good ecological environment and socioeconomic conditions, it is necessary to carry out pilot programs of innovative development models for national parks under the development requirements, explore new models of coordinated development of the ecological environment and social economy, and provide new ideas and demonstrations for the construction of other national parks. The EEL of the HRSR region was mainly improved, while the SEL decreased significantly. Therefore, it is necessary for this region to carry out industrial transformation and upgrade properly based on the premises of good ecological and environmental protection and fully utilize the TRSNP and its local characteristics, such as characteristic folk culture industries, traditional Chinese medicine, and the Tibetan medicine industry.

The advantages of the local eco-environment can also be optimized by developing the ecological experience industry and effectively improving the socioeconomic situation of the TRSNP and the wellbeing of the people. For the entire TRSNP, since one of the ultimate goals of national park construction is to promote the harmonious coexistence of man and nature, both the ecological protection and the socioeconomic development of TRSNP must be pursued. On the one hand, managers must continuously improve and innovate existing systems and mechanisms, including optimizing ecological management and protection post mechanisms, wildlife accident compensation mechanisms, and hierarchical management systems and mechanisms. Local governments should always take the construction of laws and regulations as the core and constantly explore new models of harmonious coexistence of humanity and nature. On the other hand, the development of TRSNP should be more integrated into the scientific and technological support system, create a "cloud" platform for data sharing, improve the "integration of heaven and earth" ecological environment monitoring and evaluation system and the data integration and sharing mechanism, promote information interaction, and lead regional development with science and technology.

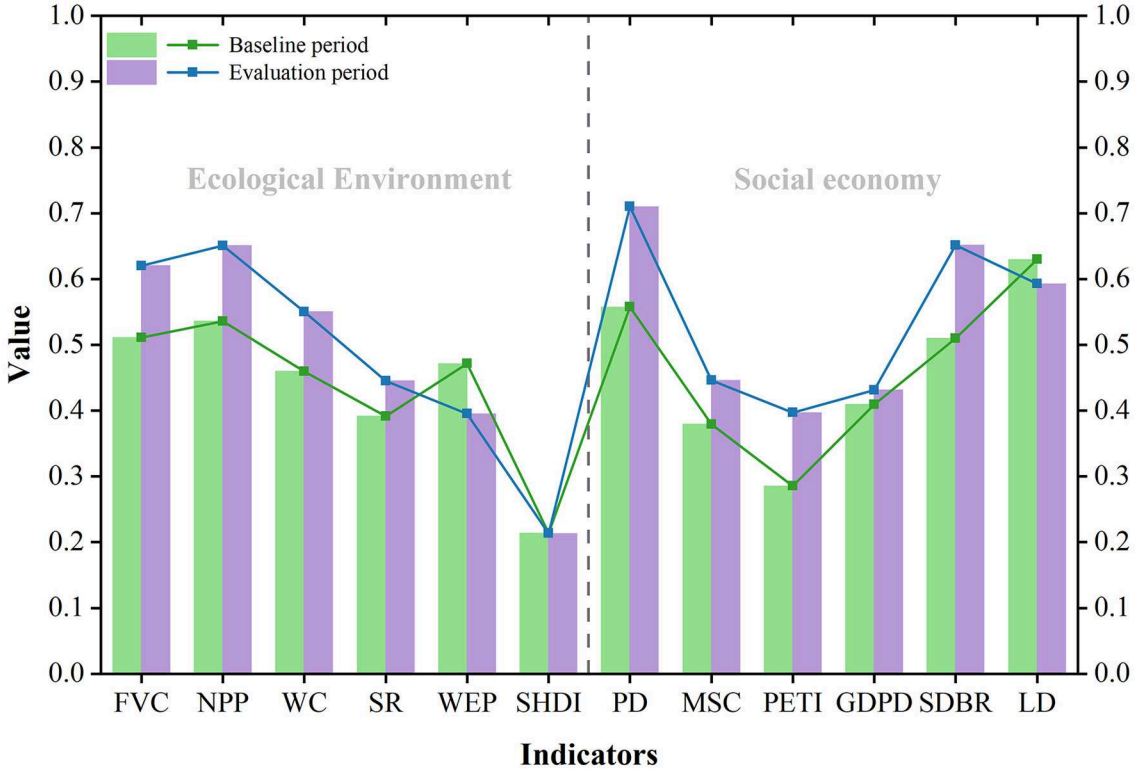


FIGURE 9
Changes of 12 indicators during the baseline and evaluation periods. Fractional vegetation cover (FVC), net primary productivity (NPP), water conservation (WC), soil retention (SR), wind erosion prevention (WEP), and Shannon's diversity index (SHDI) constitute the ecological environment level. The socioeconomic level includes population density (PD), medical service capacity (MSC), population employed in the tertiary industry (PETI), gross domestic product density (GDPD), savings deposit balance of residents (SDBR), and livestock density (LD).

4.3. Limitations and future perspectives

Previous evaluations often focused on the livelihood of ecological migrants, including a wide range of socioeconomic issues (Xin et al., 2016; Feng, 2017; Ma et al., 2021) or the alpine fragile ecological rating index system, in which the indicators measuring vegetation and climate fit with this study, and also include the indicators of population and livestock population (Yu and Lu, 2011). To achieve the goal of a comprehensive evaluation, this study not only selected the ecological index system that fits well with the index system for assessing the ecological effectiveness of the Three-River-Source ecological protection and construction project (Shao et al., 2016) but also considered the aspects of population, resources, environment, and economy (Cheng and Shen, 2000). In addition, the results of the study were more consistent compared to other relevant studies (Shao et al., 2016; Cao et al., 2019). However, considering the importance of biodiversity conservation in national parks, the evaluation index system constructed in this study is yet to include wild animals and plant species because of the limitations in data acquisition. Next, the boundary of the TRSNP does not coincide with the county administrative boundaries. Although this study successfully rostered the eco-environment and socioeconomic indicators to a scale of 1 × 1 km, indeterminacy may still remain. In addition, upon analyzing the results, the current study focuses on the macro-benefit analysis of the ecological environment and social economy and there are deficiencies in the spatial presentation and data analysis of the changes in the microindices.

Furthermore, due to the limitations of statistical data disclosure and access, this study currently selects the decade from 2011 to 2020 as the time series of the study, which is illustrative but may not be statistically significant. Meanwhile, ecological and social projects implemented before 2011 also have an impact on the degree of harmony between human and nature, and hence, the choice of 5 years as the reference period may affect the accuracy of the results. Subsequently, based on the more open and accurate data of the TRSNP, related research can build a more comprehensive index system with various aspects of the ecological environment and social economy and continuously improve the scientificity and accuracy of the evaluation of the construction effect of the TRSNP.

5. Conclusion

In this study, we examined the goals of balancing ecological conservation and human activity in national parks and propose a quantitative model to assess the degree of harmony between humanity and nature for the construction of national parks. Using the inversion model and the spatial analysis method, the pattern of the evolution characteristics of the DHHN from 2011 to 2020 were comparatively analyzed. The research shows that the DHHN

showed a trend of gradual increase from the northwest to the southeast. Compared with the pilot baseline period (2011–2015), the DHHN of the TRSNP has been greatly improved during the evaluation period (2016–2020), especially in the southeast. It indicates that the pilot national park has initially achieved the goal of balancing protection and development. However, while the DHHN has improved, some indicators of the TRSNP also fluctuate or even decline. Consequently, it is necessary to propose suitable models for protection and development and focus on coordinating and balancing the contradiction between eco-environmental protection and socioeconomic development by following the specific conditions of each region. It is committed to achieving the sustained improvement of the ecological environment, the harmonious coexistence of humanity and nature, and the sustainable development of the TRSNP. This study better enriches the relevant evaluation research on the TRSNP system pilot and also provides a good reference value for the exploration and construction of a national park system in China. We are currently applying this methodology to the TRSNP and extending the time span of the study for quantitatively evaluating the effectiveness of establishing the national park system.

Data availability statement

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

Author contributions

YG: data curation, investigation, methodology, and writing—original draft. XiaojL: investigation, conceptualization, and supervision. XiaohL: resources and supervision. JZ: resources, supervision, and funding acquisition. HZ: resources, visualization, supervision, validation, and funding acquisition. JF: methodology and supervision. NK: review, editing, and analysis. JM: visualization and methodology. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the National Key Research and Development Program (2021YFD1300501), the Second Tibetan Plateau Scientific Expedition Program (2019QZKK0608), the Remote Sensing Mapping of the Geological Interpretation Base Map of the Lancang-Mekong River Region (300012000000212171), the Strategic Priority Research Program A of the Chinese Academy of Sciences (XDA20090200), and the National Natural Science Foundation of China (42007429).

Acknowledgments

We are very grateful to the experts who put forward suggestions for the measurement of indicator weight. We also would like to express our sincere thanks to reviewers for their valuable comments.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2023.1121189/full#supplementary-material

References

An, L. T., Markowski, J., Bartos, M., Rzenca, A., and Namiecinski, P. (2019). An evaluation of destination attractiveness for nature-based tourism: recommendations for the management of national parks in Vietnam. *Nat. Conservat*. 32, 51–80. doi: 10.3897/natureconservation.32.30753

Cao, W., Liu, L. L., Wu, D., and Huang, L. (2019). Spatial and temporal variations and the importance of hierarchy of ecosystem functions in the Three-river-source National Park. *Acta Ecologica Sinica*. 39, 1361–1374. doi: 10.5846/stxb201807311629

Chang, J. E., and Jiang, T. L. (2007). Research on the weight of coefficient through analytic hierarchy process. *J. Wuhan Univ. Technol. Mater. Sci. Ed.* 1,153–156.

Chen, D. J., Zhong, L. S., and Xiao, L. L. (2020). Construction and empirical analysis of the suitability evaluation of study travel development in national park. *Acta Ecologica Sinica*. 40, 7222–7230. doi: 10.5846/stxb201904300887

Chen, M. X., Lu, D. D., and Zhang, H. (2009). Comprehensive evaluation and the driving factors of china's urbanization. *Acta Geographica Sinica*. 64, 387–398.

Chen, W. Y. (2019). *Research on Comprehensive Performance Evaluation of Z Power Supply Branch of Jiangxi Power Grid*. Jiangxi: East China University of Technology.

Cheng, S. K., and Shen, L. (2000). Approach to dynamic relationship between population resources environment and development of the Qinghai-Tibet Plateau. *J. Natural Res*. 15, 297–304.

Ding, L. Z., Xu, G. F., Lu, J. B., Zhang, D. S., and Huang, B. F. (2005). Landscape fragmentation and its effect on biodiversity. *Forest Sci. Technol.*. 32, 45–49. doi: 10.3969/j.issn.1001-7380.2005.04.017

Ding, X., Chong, X., Bao, Z., Xue, Y., and Zhang, S. (2017). Fuzzy comprehensive assessment method based on the entropy weight method and its application in the water environmental safety evaluation of the Heshangshan drinking water source area, three gorges reservoir area, China. *Water*. 9, 329. doi: 10.3390/w9050329

Fan, J. W., Shao, Q. Q., Wang, J. B., Chen, Z. Q., and Zhong, H. P. (2011). An analysis of temporal-spatial dynamics of grazing pressure on grassland in three rivers headwater region. *Chinese J. Plant Ecol*. 33, 64–72.

Feng, W. H. (2017). *Study on Ecological Migration in the Source Area of Three Rivers*. Beijing, China: Party School of the CPC Central Committee.

Fu, M. D., Liu, W. W., Li, B. Y., Ren, Y. H., Li, S., Bai, X., et al. (2021). Construction and application of an evaluation index system for ecological and environmental protection effectiveness of national parks. *Chin. J. Appl. Ecol*. 40, 4109–4118. doi: 10.13292/j.1000-4890.202112.023

Gao, J. L., Hao, Y. G., Ding, G. D., Liu, F., Xin, Z. M., Xu, J., et al. (2013). Primary assessment on the wind-breaking and sand-fixing function of the vegetation and its value in Ulan Buh desert ecosystem. *J. Arid Land*. 27, 41–46. doi: 10.13448/j.cnki.jalre.2013.12.008

General Office of the CPC and Central Committee (2017). General plan for establishing national park system. Available online at: http://www.gov.cn/gongbao/content/2017/content_5232358.htm (accessed September 26, 2017).

Gong, S. H., Xiao, Y., Zheng, H., Xiao, Y., and Ouyang, Z. Y. (2017). Spatial patterns of ecosystem water conservation in China and its impact factors analysis. *Acta Ecologica Sinica*. 37, 2455–2462. doi: 10.5846/stxb201512012406

Gorenflo, L. J., Suzanne, R., Mittermeier, R. A., Walker-Painemilla, K. (2012). Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. *PNAS*. 109, 8032–8037. doi: 10.1073/pnas.1117511109

Heiland, S., May, A., and Scherfose, V. (2020). Evaluation of the management effectiveness of german national parks—experiences, results, lessons learned and future prospects. *Sustainability*. 12, 7135. doi: 10.3390/su12177135

Hu, S., Cao, M. M., Liu, Q., Zhang, T. Q., Qiu, H. Y., Liu, W., et al. (2014). Comparative study on the soil conservation function of InVEST model under different perspectives. *Geographical Res*. 33, 2393–2406. doi: 10.11821/dlyj201412016

Huang, X. Y. (2017). *Ecosystem Health Assessment and Ecological co-mpensation standards for the Comprehensive Test Areas of Sanjiangyuan National Natural Reserve*. Qinghai: Qinghai Normal University.

Jia, H. C., Cao, C. X., Ma, G. R., Bao, D. M., Wu, X. B., Xu, M., et al. (2011). Assessment of wetland ecosystem health in the source region of yangtze, yellow and yalu Tsangpo Rivers of Qinghai Province. *Wetland Sci*. 9, 209—217. doi: 10.13248/j.cnki.wetlandsci.2011.03.001

Jiang, Y. F., Tian, J., Zhao, J. B., and Tang, X. P. (2021). The connotation and assessment framework of national park ecosystem integrity: a case study of the Amur Tiger and Leopard National Park. *Biodivers. Sci*. 29, 1279–1287. doi: 10.17520/biods.2021319

Jiao, W. J., Liu, X. Y., and He, S. Y. (2022). Establishing an ecological monitoring system for national parks in China: a theoretical framework. *Ecol. Indic*. 143, 109414. doi: 10.1016/j.ecolind.2022.109414

Juffe-Bignoli, D., Burgess, N. D., Bingham, H., Belle, E., Lima, M. D., Deguignet, M., et al. (2014). *Protected Planet Report 2014: Tracking progress towards global targets for protected areas*. Cambridge: UNEP-WCMC.

Li, C. L., and Cong, L. (2021). Research on the legislative dilemma and restrictive factors of national park law in China Based on Delphi method. *Chin. Landscape Architect*. 37, 104–108. doi: 10.19775/j.cla.2021.05.0104

Li, L., Lin, H. L., and Gao, Y. (2016). Emergy analysis of the value of grassland ecosystem services in the Three Rivers Source Region. *Acta Pratacultu-rae Sinica*. 25, 34–41. doi: 10.11686/cyxb2015387

Liang, Y. J., and Xu, Z. M. (2013). A case study in Ganzhou District, Zhan-gye municipality. *J. Glaciol. Geocryol*. 35, 249–254. doi: 10.7522/j.issn.1000-0240.2013.0030

Liu, D. P., Oouyang, Z. Y., Zhang, Y. J., Zou, H. F., Zhong, L. S., Xu, J. L., et al. (2021). Development of natural protected areas in China: opportunities and challenges. *Nat. Protect. Areas*. 1, 1–12.

Liu, J. Y., Kuang, W. H., and Zhang, Z. X. (2014). Spatiotemporal characteristics, patterns and causes of land use changes in China since the late 1980s. *Acta Geographica Sinica*. 69, 3–14. doi: 10.1007/s11442-014-1082-6

Liu, S. S. (2008). *The Influence of Population Density on the Development Level of Tertiary Industry*. Guangdong: Sun Yat-sen University,.

Lu, J. H., and Li, X. (2018). Comprehensive evaluation model of national park based on grey clustering. *Forestry Econ*. 40, 22–27. doi: 10.13843/j.cnki.lyjj.2018.05.004

Ma, K., Shen, X., Grumbine, R. E., and Corlett, R. (2017). China's biodiversity conservation research in progress. *Biol. Conserv*. 210, 1–2. doi: 10.1016/j.biocon.2017.05.029

Ma, T., Min, Q. W., Xu, K., and Sang, W. G. (2021). Resident willingness to pay for ecotourism resources and associated factors in Sanjiangyuan National Park, China. *J. Res. Ecol*. 12, 693–706. doi: 10.5814/j.issn.1674-764x.2021.05.012

Meng, P., Wang, L., and Zhao, Y. C. (2018). Ecological sensitivity in national park planning in the Beijing-Tianjin-Hebei region. *Forest Res. Manag*. 2, 98–124. doi: 10.13466/j.cnki.lyzygl.2018.02.016

Meng, Y., and Chen, W. K. (2019). Evaluation of National Park Residents' livelihood resilience in developing countries–a case study of the Giant Panda National Park, China. *Basic Clin. Pharmacol. Toxicol*. 125, 83–84.

Mu, S. J., Li, J. L., Chen, Y. Z., Gang, C. C., Zhou, W., and Ju, W. M. (2012). Spatial differences of variations of vegetation coverage in inner Mongolia during 2001-2010. *Acta Geographica Sinica*. 67, 1255–1268. doi: 10.11821/xb201209010

National Development and Reform Commission (NDRC) (2018). *Master plan of Three-River-Source National Park*. Available online at: http://www.gov.cn/xinwen/2018-01/17/content_5257568.htm (accessed January 12, 2018).

Nchor, A., and Ogogo, A. (2013). Rapid assessment of protected area pressures and threats in nigeria national parks. *Global J. Agricultural Sci*. 11, 63–72. doi: 10.4314/gjass.v11i2.1

Ouyang, Z. Y., Song, C. S., Zheng, H., Polasky, S., Xiao, Y., Bateman, I. J., et al. (2020). Using gross ecosystem product (GEP) to value nature in decision making. *Proc Natl Acad Sci USA*. 117, 14593–14601. doi: 10.1073/pnas.1911439117

Ouyang, Z. Y., Xu, W. H., and Zang, Z. H. (2021). Suggestions on improving the management system of national parks. *Biodivers Sci*. 29, 272–274. doi: 10.17520/biods.2021083

Qiao, Y. X., Zhu, H. Z., Shao, X. M., and Zhong, H. P. (2017). Research on gridding of livestock spatial density based on multi-source information. *Sci. China Technol. Sci*. 49, 53–59.

Ren, H., and Guo, Z. (2021). Progress and prospect of biodiversity conservation in China. *Ecol.Sci*. 40, 247–252.

Ren, H., Qin, H., Ouyang, Z., Wen, X. Y., Jin, X. H., Liu, H., et al. (2019). Progress of implementation on the Global Strategy for Plant Conservation in (2011–2020) China. *Biol. Conserv*. 230, 169–178. doi: 10.1016/j.biocon.2018.12.030

Rennie, A. (2006). The importance of national parks to nation-building: support for the National Parks Act (2000) in the Scottish Parliament. *Scottish Geograph. J*. 122, 223–232. doi: 10.1080/00369220601100091

Rudnicki, A., Shivatzki, S., Beyer, L. A., Takada, Y., Raphael, Y., Bovo, R., et al. (2005). *Assessment of National Park Management Effectiveness: A Case Study of Khao Yai National Park*. Thailand: Proceedings of 43rd Kasetsart University Annual Conference. p. 449–460.

Shang, T. T., and Cao, Y. K. (2019). The evaluation and analysis on sustainable livelihoods of residents in Northeast Tiger and Leopard National Park. *Forest. Econ*. 41, 17–22. doi: 10.13843/j.cnki.lyjj.2019.10.003

Shao, Q. Q., Fan, J. W., Liu, J. Y., Huang, L., Cao, W., Xu, X. L., et al. (2016). Assessment on the effects of the first-stage ecological conservation and restoration project in Sanjiangyuan region. *Acta Geographica Sinica*. 71, 3–20. doi: 10.11821/dlxb201601001

Su, X. Y. (2019). *Ecological Health Assessment of Sanjiangyuan National Park*. Qinghai: Qinghai Normal University.

Tang, F. L., Zhang, J. C., Yang, Y. M., and Wang, M. J. (2010). Study on evaluation system of National Park. *Ecol. Environ. Sci*. 19, 2993–2999. doi: 10.16258/j.cnki.1674-5906.2010.12.026

Wei, W., Zhang, K., and Zhou, J. (2020). Review and prospect of human-land relationship in Three River Headwaters Region: based on the perspective of people, events, time and space. *Adv. Earth Sci*. 35, 26–37. doi: 10.11867/j.issn.1001-8166.2020.010

Wu, X., Liu, F. G., Liu, L. S., Liu, F., and Caixiang, C. M. (2021). Changes and spatial characteristics of livestock population in Tibetan Plateau. *Ecol. Sci*. 40, 38–47.

Wu, X. Y. (2011). The influences of population density on the developmental level of tertiary industry: an empirical analysis from China. *J. Foshan Uni*. 29, 37–40. doi: 10.3969/j.issn.1008-018X.2011.03.009

Xiao, L. L., Zhong, L. S., Yu, H., and Zhou, R. (2019). Assessment of recreational use suitability of Qianjiangyuan National Park Pilot under the zoning constraints. *Acta Ecologica Sinica*. 39, 1375–1384. doi: 10.5846/stxb201808241811

Xie, Y. Y., Zhou, N. X., Ma, H. H., and Ma, Z. F. (2018). Spatial distribution characteristics and influencing factors of health services in China. *J. Cent. South Univ (Natural Sciences)*. 52, 713–722. doi: 10.19603/j.cnki.1000-1190.2018.05.019

Xin, R. P., Han, Z. Q., and Li, W. B. (2016). A study on the livelihoods of ecological migrant families at the source of three rivers: based on the field survey in Yushu, Qinghai. *J. Gansu Univ*. 01, 119–126.

Xin, Y. (2014). The degradation trend of natural grassland in Qinghai Province. *Qinghai Pratacult*. 23, 46–53.

Xinhua News Agency (2022). *Report of The 20th National Congress of the Communist Party of China (CPC)*. Available online at: http://www.gov.cn/xinwen/2022-10/25/content_5721685.htm (accessed October 10, 2022).

Xue, D., and Zhang, Y. (2019). Achievement and outlook of biodiversity conservation in China. *Environ. Protect*. 47, 38–42.

Yang, C. Y. (2022). A study of international experience in national park construction: the case of French National Parks. *City*. 2022, 18–24.

Yang, Q. Z., Li, T. T., Wang, Z. X., Lin, L. Q., Peng, Q. Q., Lin, B. J., et al. (2017). Integrated assessment on ecological sensitivity for Shennongjia National Park. *J. Hubei Uni. (Natural Science)*. 39, 455–461. doi: 10.3969/j.issn.1000-2375.2017.05.004

Yang, Z. J., and Zhu, Y. (2016). Study on the effectiveness evaluation and countermeasure of zoning management of Meili Snow Mountain National Park. *Ecol. Econ*. 32, 201–204. doi: 10.3969/j.issn.1671-4407.2016.10.043

Yu, B. H., and Lu, C. H. (2011). Assessment of ecological vulnerability on the Tibetan Plateau. *Geograph. Res*. 30, 2289–2295.

Yu, P., Zhang, J. H., Wang, Y. R., Wang, C., and Zhang, H. M. (2020). Can tourism development enhance livelihood capitals of rural households? Evidence from Huangshan National Park adjacent communities, China. *Sci. Total Environ*. 748, 141099. doi: 10.1016/j.scitotenv.2020.141099

Yuh, Y. G., Dongmo, Z. N., N'Goran, P. K., Ekodeck, H., Mengamenya, A., Kuehl, H., et al. (2019). Effects of land cover change on great apes distribution at the Lobéké National Park and its surrounding forest management units, South-East Cameroon. a 13 year time series analysis. *Sci Rep*. 9, 1445. doi: 10.1038/s41598-018-36225-2

Zhang, L. X., Fan, J. W., Shan, Q. Q., Tang, F. P., Zhang, H. Y., and Li, Y. Z. (2014). Changes in grassland yield and grazing pressure in the Three Rivers headwater region before and after the implementation of the eco-restoration project. *Acta Prataculturae Sinica*. 23, 116–123.

Zhang, M. S., Zhou, Y. Q., and Sheng, M. Y. (2022). Thoughts and suggestions on the establishment of nature reserve system withnational park as the main body. *Ecol. Sci*. 41, 237–247. doi: 10.14108/j.cnki.1008-8873.2022.06.028

Zhang, X., and Jin, X. (2021). Vegetation dynamics and responses to climate change and anthropogenic activities in the Three-River Headwaters Region, China. *Ecol. Indic*. 131, 108223. doi: 10.1016/j.ecolind.2021.108223

Zhang, Y., and Zhang, C. N. (2019). Evaluation of ecosystem cultural services in Qilian Mountain National Park, Qinghai Province. *Environ. Protect*. 47, 56–60. doi: 10.14026/j.cnki.0253-9705.2019.14.012

Zhao, X. J. (2019). Construction of China's National Park Management System. *Social Scientist*. 07, 70−74.

Zhao, X. Q. (2021). The five integrative management strategies of Sanjiangyuan National Park. *Biodiversity Science*. 29, 301–303. doi: 10.17520/biods.2021023

Zhao, X. Q., Chen, X. J., and Xian, Y. J. (2020). Dialogue: the value of Sanjiangyuan National Park. *Man Biosphere*. 1, 44–49. doi: 10.3969/j.issn.1009-1661.2020.04.010

Zhao, Z. C., and Yang, R. (2021). The concept of national park authenticity and integrity in China and its evaluation framework. *Biodiver. Sci*. 29, 1271–1278. doi: 10.17520/biods.2021287

Zhou, K., Liu, H. C., Fan, J., and Yu, H. (2021). Environmental stress intensity of human activities and its spatial effects in the Qinghai-Tibet Plateau national park cluster: a case study in Sanjiangyuan region. *Acta Ecologica Sinica*. 41, 268–279. doi: 10.5846/stxb202003310766

Zhu, H. G., Zhao, M. H., Chen, Y. R., and Zhang, Y. T. (2022). Community governance in national parks: international experience and enlightenment. *World Forest. Res*. 35, 1–6. doi: 10.13348/j.cnki.sjlyyj.2022.0072.y

Zhu, J. M., and Cheng, F. Y. (2016). Study on the determinants of chinese residents' savings deposit balance. *Econ. Vision*. 2, 58–67. doi: 10.3969/j.issn.1672-3309(s).2016.02.08

Zhu, W. Q., Pan, Y. Z., and Zhang, J. S. (2007). Estimation of net primary productivity of chinese terrestrial vegetation based on remote sensing. *J. Plant Ecol*. 1, 413–424. doi: 10.17521/cjpe.2007.0050

frontiers | Frontiers in Environmental Science

TYPE Original Research
PUBLISHED 10 March 2023
DOI 10.3389/fenvs.2023.1097249

OPEN ACCESS

EDITED BY
Kaijie Xu,
University of Alberta, Canada

REVIEWED BY
Xiaotong Zhang,
Beijing Normal University, China
Rui Zhang,
Xidian University, China

*CORRESPONDENCE
Xinchang Zhang,
zhangxc@gzhu.edu.cn

SPECIALTY SECTION
This article was submitted to Environmental Informatics and Remote Sensing, aection of the journal Frontiers in Environmental Science.

RECEIVED 13 November 2022
ACCEPTED 07 February 2023
PUBLISHED 10 March 2023

CITATION
Ruan Y, Ruan B, Xin Q, Liao X, Jing F and Zhang X (2023), phenoC++: An open-source tool for retrieving vegetation phenology from satellite remote sensing data.
Front. Environ. Sci. 11:1097249.
doi: 10.3389/fenvs.2023.1097249

phenoC++: An open-source tool for retrieving vegetation phenology from satellite remote sensing data

Yongjian Ruan[1,2], Baozhen Ruan[1], Qinchuan Xin[3], Xi Liao[1], Fengrui Jing[4] and Xinchang Zhang[1]*

[1]School of Geography and Remote Sensing, Guangzhou University, Guangzhou, China, [2]Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shenzhen, China, [3]School of Geography and Planning, Sun Yat-Sen University, Guangzhou, China, [4]Department of Geography, University of South Carolina, Columbia, SC, United States

Satellite-retrieved vegetation phenology has great potential for application in characterizing seasonal and annual land surface dynamics. However, obtaining regional-scale vegetation phenology from satellite remote sensing data often requires extensive data processing and computation, which makes the accurate and rapid retrieval of regional-scale phenology a challenge. To retrieve vegetation phenology from satellite remote sensing data, we developed an open-source tool called phenoC++, which uses parallel technology in C++. phenoC++ includes six common algorithms: amplitude threshold (AT), first-order derivative (FOD), second-order derivative (SOD), third-order derivative (TOD), relative change rate (RCR), and curvature change rate (CCR). We implemented the proposed phenoC++ and evaluated its performance on a site scale with PhenoCam-observed phenology metrics. The result shows that SOS derived from MODIS images by phenoC++ with six methods (i.e., AT, FOD, SOD, RCR, TOD, and CCR) obtained r-values of 0.75, 0.76, 0.75, 0.76, 0.64, and 0.67, and RMSE values of 21.36, 20.41, 22.38, 19.11, 33.56, and 32.14, respectively. Satellite-retrieved EOS by phenoC++ with six methods obtained r-values of 0.58, 0.59, 0.57, 0.56, 0.36, and 0.40, and RMSE values of 52.43, 46.68, 55.13, 49.46, 71.13, and 69.34, respectively. Using PhenoCam-observed phenology as a baseline, SOS retrieved by phenoC++ was superior to MCD12Q2, while EOS retrieved by phenoC++ was slightly inferior to that of MCD12Q2. Moreover, compared with MCD12Q2 on a regional scale, phenoC++-retrieved vegetation phenology yields more effective pixels. The innovative features of phenoC++ are 1) integrating six algorithms for retrieving SOS and EOS; 2) quickly processing data on a large scale with simple input startup parameters; 3) outputting phenology metrics in GeoTIFF format image, which is more convenient to use with other geospatial data. phenoC++ could aid in investigating and addressing large-scale phenology problems of the ecological environment.

KEYWORDS

vegetation phenology, satellite remote sensing data, PhenoCam, C++ language, the contiguous United States

1 Introduction

Vegetation phenology is a first-order control of ecosystem productivity that reflects and affects the physiological, physical, and chemical processes of vegetation ecosystems. It also plays a key role in energy exchange and affects the carbon balance of the earth system (Foley et al., 1996; Ganguly et al., 2010; Elmore et al., 2012; Zhang et al., 2020; Gao et al., 2021). It is an essential component of the earth's ecology (DeFries et al., 1995; Foley et al., 1996). In the context of global warming, characterizing seasonal and annual land surface dynamics is critical for monitoring climate change scenarios (Peng et al., 2017; Caparros-Santiago et al., 2021; Kollert et al., 2021).

Satellite remote sensing data with rich historical records and suitable spatial-temporal resolutions are frequently employed to monitor the variation of vegetation phenology (Ganguly et al., 2010; Peng et al., 2017; Ao et al., 2020; Peng et al., 2021). Satellite-retrieved phenology has great application potential in characterizing both seasonal and annual land surface dynamics (De Beurs and Henebry, 2008; Zhang et al., 2018), such as monitoring climate–vegetation interaction and extreme events, modeling carbon cycles, crop-type discrimination, crop-yield estimation, and land cover mapping (Walker et al., 2012; Elmore et al., 2017; Bolton et al., 2020; Peng et al., 2021). Most studies based on satellite remote sensing data (such as AVHRR, MODIS, and SPOT VGT) have developed related algorithms for retrieving regional- or global-scale phenological products, such as amplitude threshold (AT) (ORNL Distributed Active Archive Center, Fischer, 1994; Zhou et al., 2016), first-order derivative (FOD) (Yu et al., 2003), second-order derivative (SOD) (Sakamoto et al., 2005), third-order derivative (TOD) (Tan et al., 2011), relative change rate (RCR) (Piao et al., 2006), and curvature change rate (CCR) (Zhang et al., 2003). Previous researchers integrated one or more phenological retrieval algorithms into software toolkits such as TIMESAT (Jönsson and Eklundh, 2004), phenofit (Kong et al., 2022), and the phenor R package (Hufkens et al., 2018). These are of great significance and aid in vegetation phenology. However, these software tools frequently require complex parameters to be entered, are slow to process data at runtime, lack open-source code, or lack ground-based observations to evaluate their performance results.

In the past, few ground-based observational datasets of vegetation phenology were publicly available online, preventing researchers from accessing ground validation data to evaluate the performance of retrieving algorithms (Zhou et al., 2016). Recently, some vegetation phenology datasets obtained from ground-based observations have been directly downloaded from the internet, including United States of America National Phenology Network (USA-NPN) data resources, the Pan-European Phenological database (PEP725) (Templ et al., 2018), and the PhenoCam Dataset (Seyednasrollah et al., 2019). The PhenoCam Dataset mainly applies red–green–blue (RGB) digital cameras to record the timing of the specific phenophases of plants, which differs from the USA-NPN and PEP725, which are collated by human observation. The use of digital cameras to observe changes in the phenological states of vegetation may reduce the uncertainty caused by non-uniformity compared to traditional human observations (Menzel, 2002; Richardson et al., 2018) and may be more appropriate for evaluating the performance of satellite-based vegetation phenology.

Our study developed an open-source tool called phenoC++ for retrieving vegetation phenology. The innovation of this tool is threefold. First, it integrates six algorithms for retrieving data at the start of season (SOS) and end of season (EOS). Second, phenoC++ quickly processes data, and the input startup parameters are simple. Third, it outputs phenology metrics in GeoTIFF format images, which are more convenient to use with other geospatial data. Furthermore, in this study, we implement this open-source tool to retrieve vegetation phenology and evaluate its performance on both site and regional scales using PhenoCam-observed phenophases and existing MODIS phenology products.

2 Data and methods

2.1 The algorithms of phenoC++ for retrieving vegetation phenology

The open-source tool phenoC++ includes six methods that correspond to different algorithms for retrieving vegetation phenology from satellite remote sensing data (see Table 1. These six methods have been widely used to retrieve SOS and EOS from remote sensing time-series data such as EVI2, NDVI, and LAI. In this study, we employed the EVI2 time series data from MOD09Q1 as input data for phenoC++ and obtained the EVI2 time series data from MOD09Q1 using the following equation:

$$EVI2 = 2.5\frac{P_N - P_R}{P_N + 2.4P_R + 1}, \quad (1)$$

where EVI2 denotes a two-band enhanced vegetation index, P_N denotes the near-infrared band reflectance, and P_R denotes the red band reflectance.

We used the Savitzky–Golay (S-G) method (Savitzky and Golay, 1964) to remove the outliers of the EVI2 time series and interpolated the eight-day EVI2 time series to one-day intervals. We divided the EVI2 time series of a growing season into a growth period and a dormancy period and used Eq. 2 to fit them.

$$y(t) = \frac{c}{1 + e^{a+bt}} + d, \quad (2)$$

where t denotes the day of the year, y (t) denotes the EVI2 value on date t, and a, b, c, and d are fitting parameters. We used this data series when fitting the EVI2 of vegetation dormancy with Eq. 2.

2.2 Test region

We selected the conterminous United States as the test region; it has a rich distribution of vegetation types, such as a large number of deciduous broadleaf forests distributed in its east and west, and vast grassland, shrub, and some deciduous broadleaf forests distributed in its center. In addition, many digital camera observation stations for tracking vegetation phenology have been established in the conterminous United States, and data from these stations can be used to evaluate the performance of vegetation phenology extracted from satellite remote sensing data using phenoC++.

TABLE 1 Description of the six algorithms for extracting the start-of-season (SOS) and end-of-season (EOS) from the EVI2 time series.

Method	Index	SOS	EOS	Reference
Amplitude threshold	AT	$0.2 \times (\max(\mathrm{EVI2}_s) - \min(\mathrm{EVI2}_s))$	$0.2 \times (\max(\mathrm{EVI2}_a) - \min(\mathrm{EVI2}_a))$	Zhou et al. (2016)
First-order derivative	FOD	$\max(dt(\mathrm{EVI2}_s))$	$\min(dt(\mathrm{EVI2}_a))$	Yu et al. (2003)
Second-order derivative	SOD	$\max(dt^2(\mathrm{EVI2}_s))$	$\max(dt^2(\mathrm{EVI2}_a))$	Sakamoto et al. (2005)
Relative change rate	RCR	$\max(\frac{\mathrm{EVI2}_s(t+1)-\mathrm{EVI2}_s(t)}{\mathrm{EVI2}_s(t)})$	$\min(\frac{\mathrm{EVI2}_a(t+1)-\mathrm{EVI2}_a(t)}{\mathrm{EVI2}_a(t)})$	Piao et al. (2006)
Third-order derivative	TOD	$\max(dt^3(\mathrm{EVI2}_s))$	$\min(dt^3(\mathrm{EVI2}_a))$	Tan et al. (2011)
Curvature change rate	CCR	$\max(K_s'(t))$	$\max(K_a'(t))$	Zhang et al. (2003)

Note: $\mathrm{EVI2}_s$ represents EVI2 at the periods of sustained increase for phenological cycles (i.e., spring): $\mathrm{EVI2}_a$ represents EVI2 at the periods of sustained decrease for phenological cycles (i.e., autumn); t represents the day of the year in the EVI2 time series; K_s' represents the rate of change of the curvature of the logistic-fitted EVI2 time series during the periods of sustained increase for phenological cycles (i.e., spring); K_a' represents the rate of change of curvature of the logistic-fitted EVI2 time series during the periods of sustained decrease for phenological cycles (i.e., autumn) (for more details, please see Zhang et al. ,2003); max () and min () denote the maximum and minimum of the time series data, respectively; dt (), dt^2 (), and dt^3 () denote obtaining the first-order, second-order, and third-order derivatives of time series data, respectively.

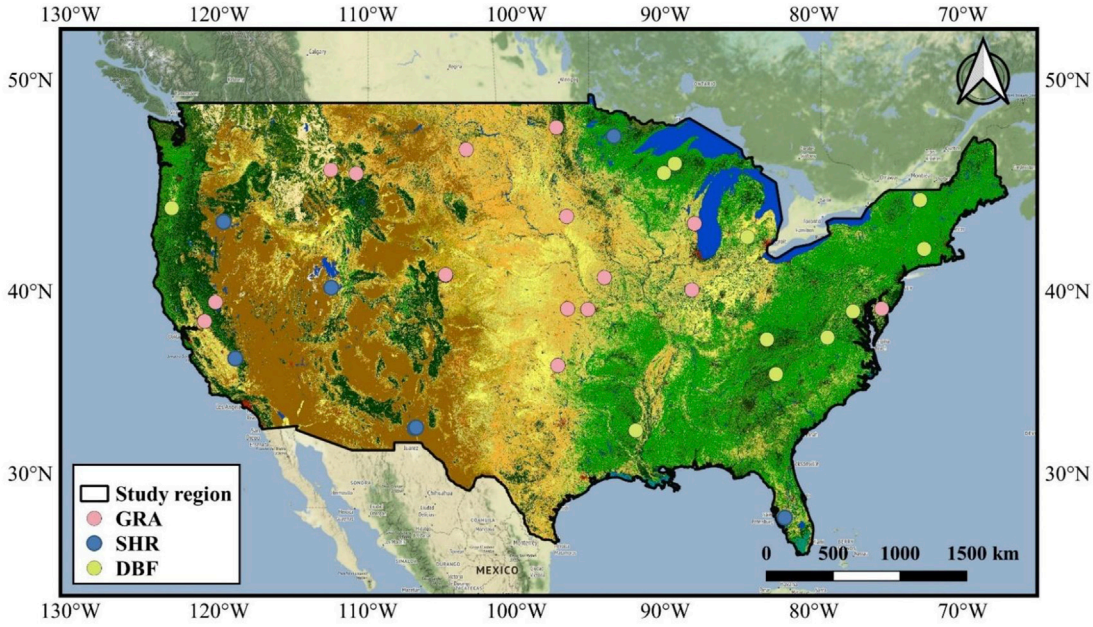


FIGURE 1
Test region and locations of selected PhenoCam sites, including deciduous broadleaf forest (DBF), shrubland (SHR), and grassland (GRA).

2.3 MODIS data

The MODIS/Terra surface reflectance eight-day 250-m product (MOD09Q1, Version 6) (Vermote et al., 2015) was used to retrieve the vegetation phenology in the developed open-source tool. MOD09Q1 includes two surface reflectance bands (red band, 620–670 nm; near-infrared band, 841–876 nm) with eight-day temporal resolution and an approximately 250-m spatial resolution. The MODIS/Terra and Aqua Land Cover Dynamics Yearly L3 Global 500 m SIN Grid (MCD12Q2, Version 6), which provides vegetation phenology over global land surfaces (Friedl et al., 2019), was used as reference data for evaluating the performance of the developed open-source tool. Both MOD09Q1 and MCD12Q2 are available from https://search.earthdata.nasa.gov/. We used the greenup layer (i.e., SOS) and the dormancy layer (i.e., EOS) of MCD12Q2 for analysis. In the MCD12Q2 product, SOS and the EOS would be obtained when EVI2 crosses 15% of the segment EVI2 amplitude for the first or last time, respectively.

2.4 PhenoCam data

We used PhenoCam Dataset v2.0 (Seyednasrollah et al., 2019) as validation data for evaluating the performance of vegetation phenophases extracted from the MOD09Q1 EVI2 time series by phenoC++. The PhenoCam sites were established in 2008 and use networked digital cameras to track vegetation phenology. There are

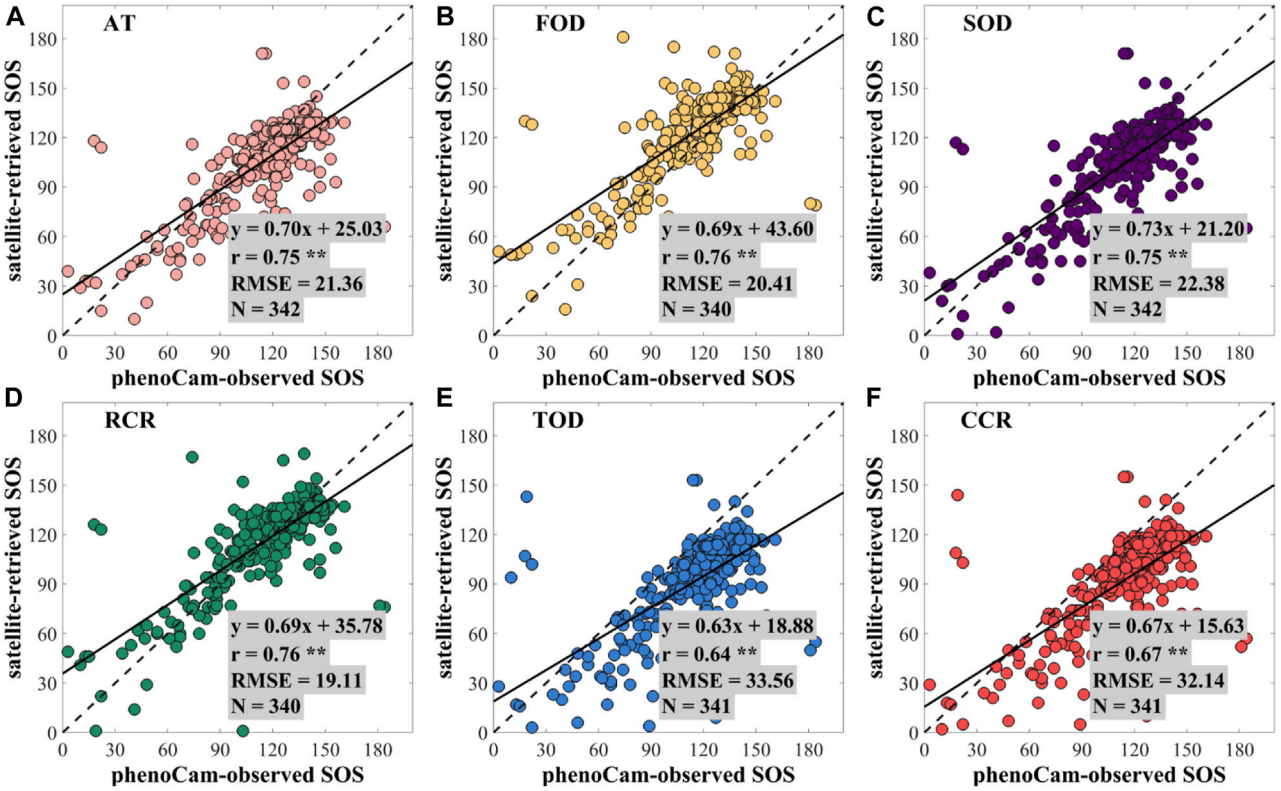


FIGURE 2
Scatterplots for the regression analysis between the PhenoCam-observed SOS and the satellite-retrieved SOS using the methods of **(A)** amplitude threshold (AT), **(B)** first-order derivative (FOD), **(C)** second-order derivative (SOD), **(D)** relative change rate (RCR), **(E)** third-order derivative (TOD), and **(F)** curvature change rate (CCR). ** denotes a *p*-value of two-tailed Student's t-tests of <0.01. N denotes the number of PhenoCam observation sites.

500 sites, particularly located in North America. The PhenoCam Dataset v2.0 (Seyednasrollah et al., 2019) can be obtained from https://phenocam.sr.unh.edu/webcam/. We used the phenology records of 28 deciduous broadleaf forest (DBF) sites, 43 shrubland (SHR) sites, and 38 grassland (GRA) sites from 2009 to 2018 for satellite-retrieved phenology validation (as shown in Figure 1). In the PhenoCam Dataset, three threshold values (i.e., 10%, 25%, and 50%) of the green chromatic coordinate index mean (G_{cc}-mean) were employed to retrieve vegetation phenophases. Among them, the threshold value of 50% of the vegetation index time series is rarely used to retrieve vegetation phenophases from remote sensing data, while the SOS and EOS retrieved by threshold values of 10% and 25% of the vegetation index time series are similar (Ruan et al., 2021). Therefore, SOS and EOS of the PhenoCam Dataset v2.0, which were retrieved from the threshold values of 25% of G_{cc}-mean amplitude, were selected for validation.

3 Results

3.1 Performance of phenoC++ on a site scale

After obtaining the EVI2 time series from MOD09Q1, we used phenoC++ with the six methods to retrieve the vegetation phenology from 2009 to 2018 and evaluated the retrieved vegetation phenology using PhenoCam-observed phenology. Pearson correlation coefficient and root-mean-square error (RMSE) were used to evaluate the performance of vegetation phenology retrieved by the open-source tool. Figure 2 shows the scatterplots for the regression result between the PhenoCam-observed SOS and the satellite-retrieved SOS using multiple methods. In Figure 2, most points are around 1: 1, demonstrating that SOS derived from two independent datasets is consistent. Compared with PhenoCam-observed SOS, the SOS derived from MOD09Q1 EVI2 by phenoC++ with six methods obtained r-values of 0.75, 0.76, 0.75, 0.76, 0.64, and 0.67, and RMSE values of 21.36, 20.41, 22.38, 19.11, 33.56, and 32.14, respectively. The TOD method produced the worst SOS, while the other methods performed similarly to each other.

Figure 3 shows the scatterplots for the regression analysis between the PhenoCam-observed EOS and the satellite-retrieved EOS using phenoC++ with six methods. Most points in Figure 3 are also around 1: 1, but they are more discrete than the SOS scatter points, as shown in Figure 2. Compared with PhenoCam-observed EOS, the EOS derived from MOD09Q1 EVI2 by phenoC++ with six methods obtained r-values of 0.58, 0.59, 0.57, 0.56, 0.36, and 0.40, and RMSE values of 52.43, 46.68, 55.13, 49.46, 71.13, and 69.34, respectively. The CCR method obtained the worst EOS, and the

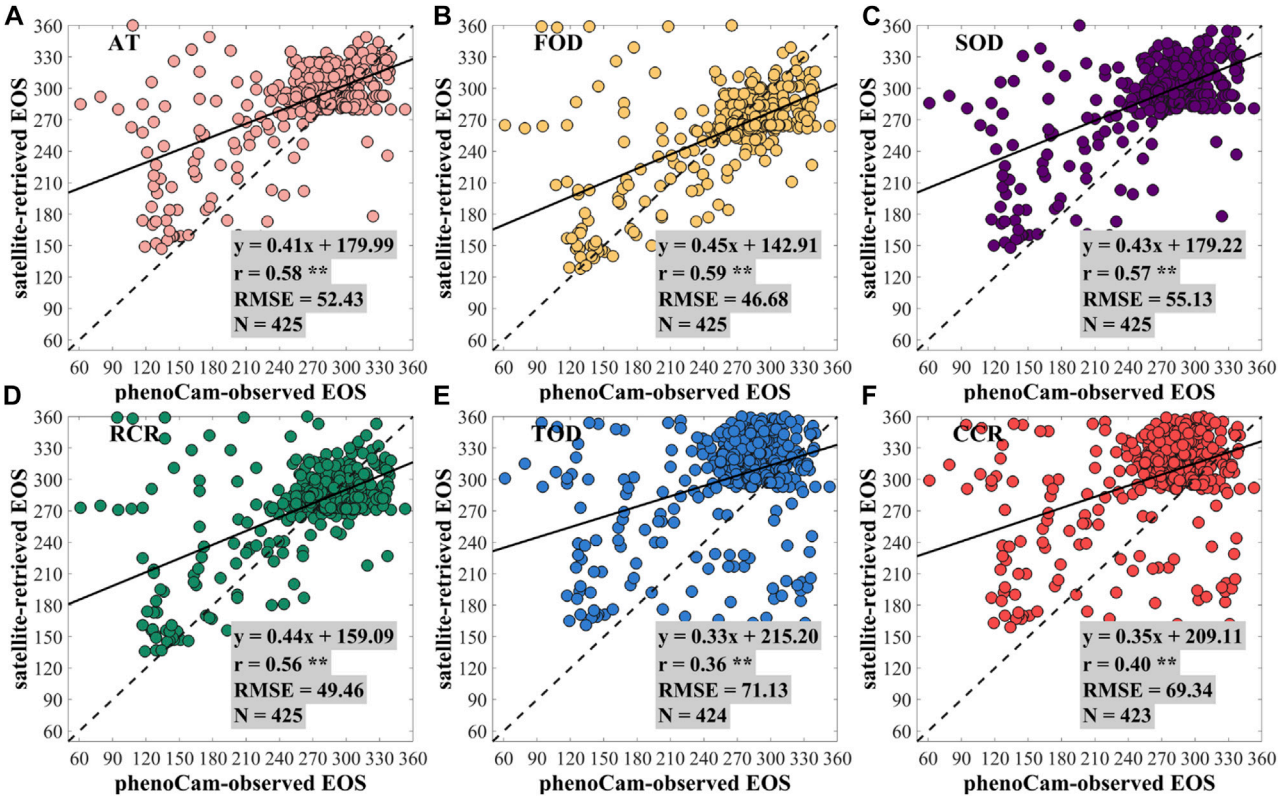


FIGURE 3
Scatterplots for the regression analysis between the PhenoCam-observed EOS and the satellite-retrieved EOS using the methods of **(A)** amplitude threshold (AT), **(B)** first-order derivative (FOD), **(C)** second-order derivative (SOD), **(D)** relative change rate (RCR), **(E)** third-order derivative (TOD), and **(F)** curvature change rate (CCR). ** denotes a *p*-value of two-tailed Student's t-tests of <0.01. N denotes the number of PhenoCam observation sites.

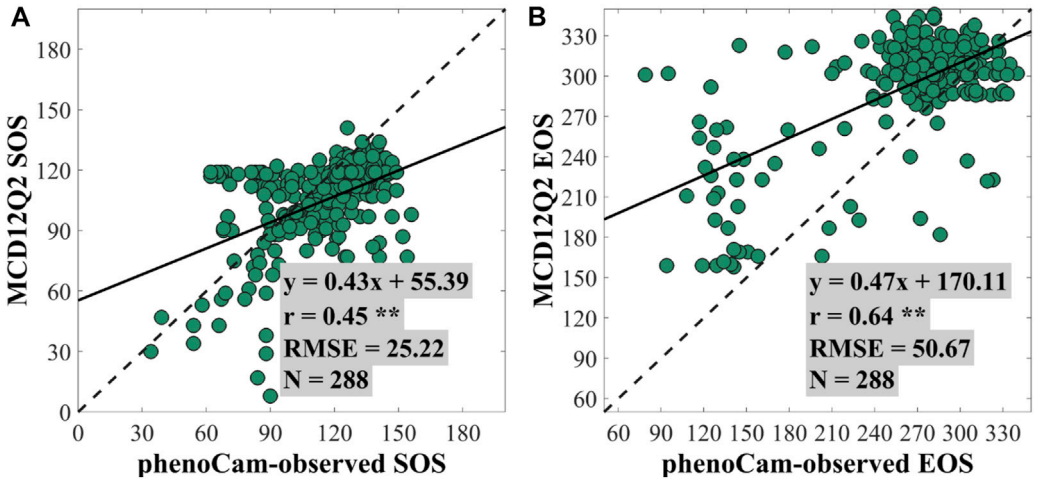


FIGURE 4
Scatterplots for the regression analysis between the PhenoCam-observed SOS/EOS and the SOS/EOS of MCD12Q2 at **(A)** the start of season (SOS) and **(B)** the end of season (EOS). ** denotes a *p*-value of two-tailed Student's t-tests of <0.01. N denotes the number of PhenoCam observation sites.

other methods had similar performance to each other. Figures 2 and 3 show that, compared with PhenoCam-observed phenology, the accuracy of satellite-retrieved SOS is slightly higher than that of satellite-retrieved EOS, which is consistent with previous results reported by Li et al. (2019). This may be caused by the different sensitivities of G_{cc} and EVI2 for detecting vegetation growth

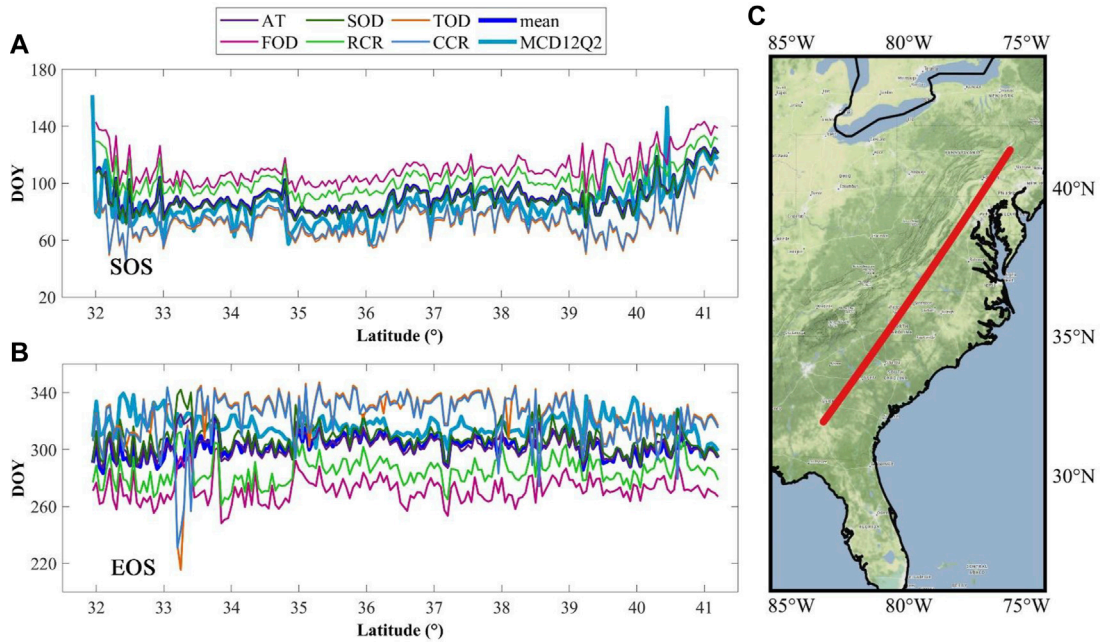


FIGURE 5
Profile of the start of season (SOS) and the end of season (EOS) in 2017 derived from two different data sources. **(A)** Profile of SOS; **(B)** profile of EOS; **(C)** location of the profile. The average value of phenophases was calculated every 0.05° along the profile. MCD12Q2 denotes the phenophases obtained from the MCD12Q2 product; AT, FOD, SOD, RCR, TOD, and CCR denote the phenophases estimated from the MOD09Q1 EVI2 time series using corresponding methods, and "mean" denotes the average value of the phenophases estimated from the fused EVI2 time series by six methods.

changes. During vegetation growth, G_{cc} and EVI2 both rapidly increase. EVI2 decreases slightly during vegetation dormancy, but G_{cc} rapidly decreases if leaf color changes.

Figure 4 illustrates the regression analysis of SOS/EOS obtained from MCD12Q2 and the PhenoCam Dataset. We used it to indirectly evaluate the performance of phenoC++. The regression analysis indicated that the SOS of MCD12Q2 achieved an r-value of 0.45 and an RMSE value of 25.22, and the EOS of MCD12Q2 achieved an r-value of 0.64 and an RMSE value of 50.67. Comparison with the PhenoCam Dataset found that the SOS of MCD12Q2 is unfavorable with the SOS retrieved by the phenoC++. The least desirable method of phenoC++ (TOD) obtained an r-value of 0.64, and its performance shows that it is better than the MCD12Q2 SOS, with an r-value of 0.45. For EOS, that of MCD12Q2 (r = 0.64; RMSE = 50.67) is better than the EOS obtained by phenoC++. Nevertheless, the SOS/EOS for retrieval by phenoC++ (N ≈ 420) obtained more effective pixels in the PhenoCam observation sites than MCD12Q2 (N = 288). Through the regression analyses between the phenoC++-retrieved phenology/MCD12Q2 with PhenoCam-observed phenology, we found an RMSE range of about 20–55 days. This is similar to the results of Xin, *et al.* [42], who used satellite-retrieved phenology for comparison with US National Phenology Network data (RMSE ~25–55 days), and Li, *et al.* [43], who used 30 m fine-resolution satellite-retrieved phenology to compare PhenoCam-observed phenology (RMSE about 25 days; r of SOS is 0.66 and r of EOS is 0.43). The main reason for the large RMSE could be that the observation scale between the satellite and PhenoCam is inconsistent, and the observation of spectral difference between the satellite sensor and PhenoCam-camera.

Figure 5 shows the profile of phenophases obtained from MOD09Q1 by phenoC++ with six methods and MCD12Q2. The average SOS estimated by the six methods and the AT and SOD methods were similar to the MCD12Q2 SOS. Compared with the MCD12Q2 SOS, the FOD and RCR methods were overestimated, while the TOD and CCR methods were underestimated. Meanwhile, the average EOS estimated by the six methods and the AT and SOD methods were similar to the MCD12Q2 EOS. Compared with the MCD12Q2 EOS, the TOD and CCR methods were overestimated, while the FOD and RCR methods were underestimated. Overall, the phenophases (SOS and EOS) estimated by phenoC++ were in line with those of MCD12Q2 with reliable performance.

3.2 The performance of phenoC++ on a regional scale

Figure 6 shows the spatial distribution for SOS over the conterminous United States, including the MCD12Q2 SOS and the SOS obtained from the MOD09Q1 EVI2 time series by phenoC++ with six methods. The SOS values obtained from MOD09Q1 EVI2 (Figures 6B–G) are close to MCD12Q2 SOS (Figure 6A) on a regional scale, while the former SOS is generally slightly earlier than the latter SOS in the southwest of the USA and slightly more delayed than the latter SOS in the northeast. Compared with Figure 6A, the SOS retrieved by phenoC++ shows high-value (red) pixels in North Dakota, South Dakota,

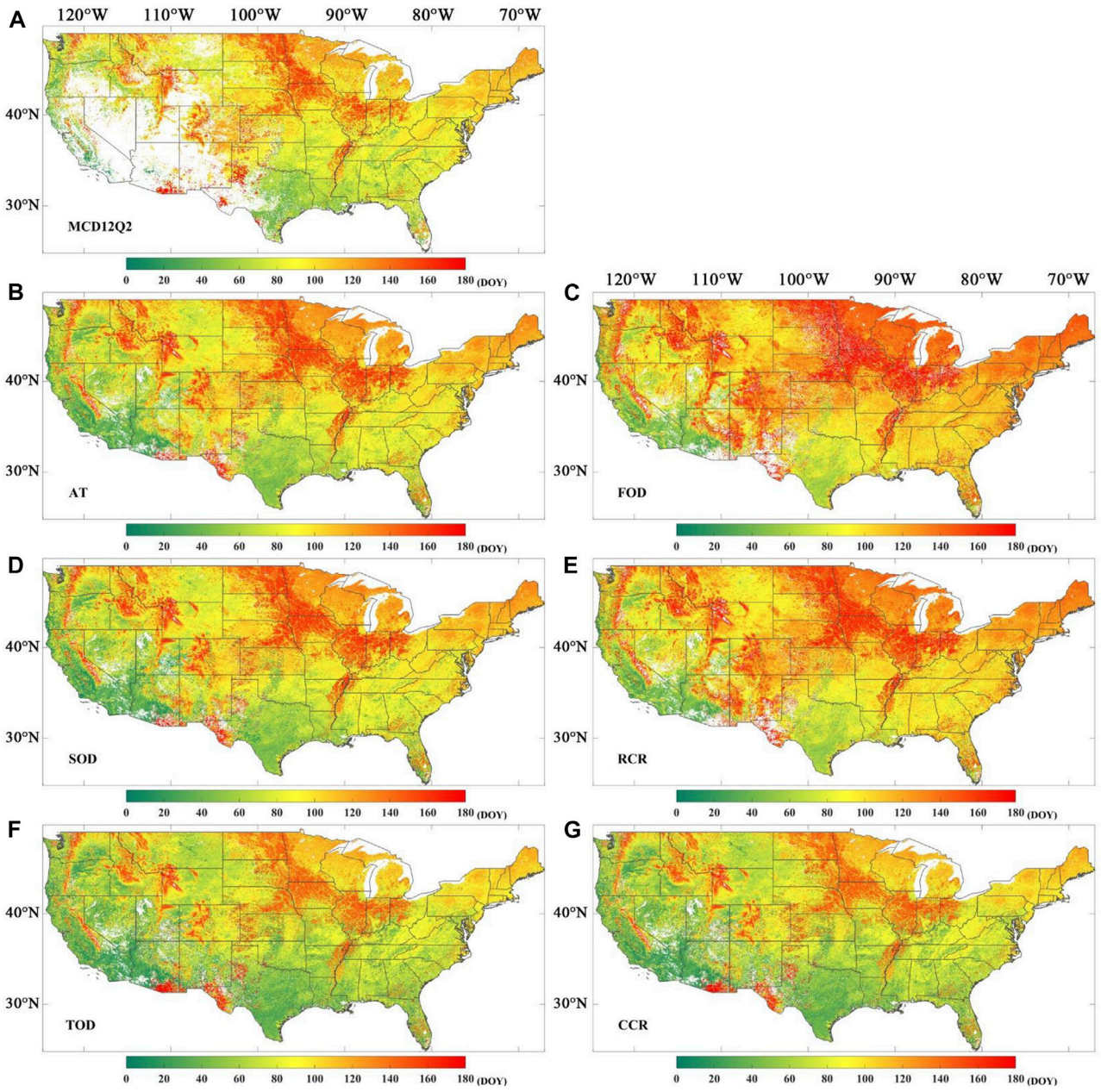


FIGURE 6
Spatial distribution for the start of season (SOS) in 2017. **(A)** SOS obtained from MCD12Q2; **(B)** SOS derived from MOD09Q1 EVI2 time series by amplitude threshold (AT); **(C)** SOS derived from MOD09Q1 EVI2 time series by first-order derivative (FOD); **(D)** SOS derived from MOD09Q1 EVI2 time series by second-order derivative (SOD); **(E)** SOS derived from MOD09Q1 EVI2 time series by relative change rate (RCR); **(F)** SOS derived from MOD09Q1 EVI2 time series by third-order derivative (TOD); **(G)** SOS derived from MOD09Q1 EVI2 time series by curvature change rate (CCR).

Minnesota, Iowa, Illinois, Indiana, Ohio, and Pennsylvania, especially in the results from the FOD and RCR methods. From Figures 6A–G, the SOS in the southwest of the USA (less than 80 days) is earlier than in the northeast of the USA (more than 80 days, mostly more than 140 days). This shows that the vegetation season begins earlier in the southwest of the USA than that in the northeast. In addition, the MCD12Q2 SOS (Figure 6A) shows more data gaps in the southwest of the USA, which is not shown in the SOS retrieved by phenoC++. This phenomenon could be caused by the MOCD12Q2 algorithm's inability to effectively retrieve phenological dates from remote sensing data in places with relatively sparse vegetation.

Figure 7 shows the spatial distribution of the EOS over the conterminous United States, including the MCD12Q2 EOS and the EOS obtained from the MOD09Q1 EVI2 time series with phenoC++ with six methods. The EOS values obtained from MOD09Q1 EVI2 (Figures 7B–G) are close to MCD12Q2 EOS (Figure 7A) on the regional scale, especially in Figures 7B and 7D. The EOS obtained from MOD09Q1 EVI2 with the FOD

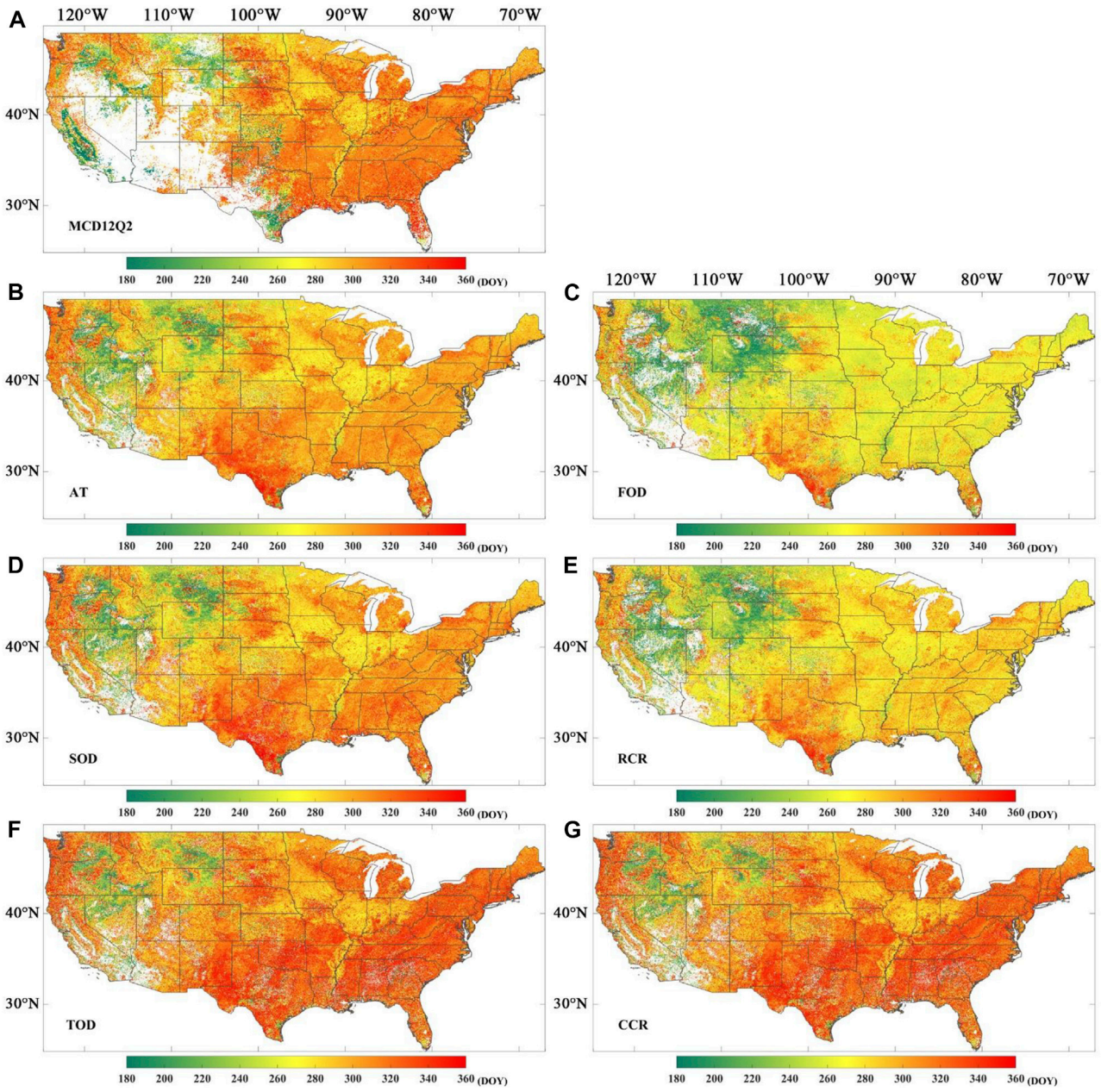


FIGURE 7
Spatial distribution for the end of season (EOS) in 2017. **(A)** EOS obtained from MCD12Q2; **(B)** EOS derived from MOD09Q1 EVI2 time series by amplitude threshold (AT); **(C)** EOS derived from MOD09Q1 EVI2 time series by first-order derivative (FOD); **(D)** EOS derived from MOD09Q1 EVI2 time series by second-order derivative (SOD); **(E)** EOS derived from MOD09Q1 EVI2 time series by relative change rate (RCR); **(F)** EOS derived from MOD09Q1 EVI2 time series by third-order derivative (TOD); **(G)** EOS derived from MOD09Q1 EVI2 time series by curvature change rate (CCR).

(Figure 7C) and RCR (Figure 7E) methods are slightly earlier than the EOS of MCD12Q2, and the EOS obtained from MOD09Q1 EVI2 with the TOD (Figure 7F) and CCR (Figure 7G) methods are generally slightly delayed than the EOS of MCD12Q2. From Figures 7A–G, the EOS in the northwest of the USA (about 240 days) is earlier than that in the southeast (more than 300 days, mostly more than 320 days). This suggests that the vegetation season ends later in the southeast of the USA than that in the northwest. In addition, like Figure 6A, Figure 7A also has data gaps in the southwest of the USA.

4 Discussion

Previous research has provided phenology-retrieval tools, such as in the R package (Hufkens et al., 2018; Kong et al., 2022) and the MATLAB platform (Jönsson and Eklundh, 2004). However, their stability and speed in large-scale vegetation phenology retrieval are not as effective as programs written in C++. This study provides a C++ compiled running program (phenoC++), which is compiled under the Linux system. If users need to run it under Windows, they must install the C++ and GDAL environments before compiling it. phenoC++ only

retrieves two commonly used and critical phenology metrics (i.e., SOS and EOS). If users need to retrieve other phenology metrics or add more algorithms, the source code of phenoC++ can be modified.

We compared the performance of TIMESAT software for retrieving the phenology metrics to that of phenoC++. Areas of 802 × 642 pixels and 5164 × 4193 pixels (about the area of Texas) from MOD09Q1 were used to test the performance of TIMESAT software and phenoC++. The test computer was configured as follows. The computer system was Ubuntu 16.04, and the CPU model was Intel(R) Xeon(R) Platinum 8269CY CPU @ 2.50 GHz (52 threads) with 512 GB of computer memory. After data preprocessing, TIMESAT ran for approximately 2.5 min in the 802 × 642-pixel region, occupying 0.2 GB of memory, while in the 5164 × 4193-pixel test, the running time was approximately 25 min and the memory occupied 0.4 GB. After data preprocessing, the phenoC++ test in the 802 × 642-pixel region took approximately 0.63 min and used 0.2 GB of memory, while the test for 5164 × 4193 pixels took approximately 24.5 min and used 4.8 GB of memory. This shows that the running speeds of the TIMESAT and phenoC++ are similar and that phenoC++ occupies more memory when the same number of CPU threads are used. This is because TIMESAT uses multiple steps to retrieve vegetation phenology, whereas phenoC++ is integrated, which is more user-friendly. Compared with TIMESAT, phenoC++ also has the following advantages. 1) phenoC++ outputs six algorithms' results for SOS or EOS, while TIMESAT only outputs threshold method results for SOS and EOS. 2) phenoC++ outputs phenology metrics in GeoTIFF format images with geographical coordinates, which are more convenient for use with other geospatial data, while TIMESAT outputs binary files. 3) The code of phenoC++ is open source, while TIMESAT is not. In addition, we also tested phenoC++ for retrieving vegetation phenology from MOD09Q1 EVI2 with 250-m spatial resolution over the USA (22,731 × 9774 pixels). The run time of phenoC++ was approximately 4.5 h, and the memory occupied was approximately 57.9 GB. Therefore, phenoC++ is an efficient and easy-to-use software tool.

5 Conclusion

Vegetation phenology is a first-order control on ecosystem productivity, so its accurate and rapid retrieval from satellite remote sensing data is key to understanding the feedback between the climate and the biosphere. We developed phenoC++, a tool that uses parallel C++ technology to retrieve start-of-season (SOS) and end-of-season (EOS) vegetation data from satellite time series data. Compared to traditional tools, the innovative features of phenoC++ include 1) integrating six algorithms for retrieving SOS and EOS; 2) rapid large-scale data processing with simple input startup parameters; 3) outputting phenology metrics as GeoTIFF format images, which are more convenient to use with other geospatial data.

We implemented phenoC++ to quickly and easily obtain the spatial distribution of SOS and EOS at 250-m spatial resolution over the conterminous United States using MODIS time-series data. We then evaluated phenoC++ performance for retrieving SOS and EOS on a site scale using PhenoCam Dataset v2.0. The results show that SOS derived from MODIS images by phenoC++ with six methods obtained r-values of 0.75, 0.76, 0.75, 0.76, 0.64, and 0.67, and RMSE values of 21.36, 20.41, 22.38, 19.11, 33.56, and 32.14, respectively. The satellite-retrieved EOS by phenoC++ with six methods obtained r-values of 0.58, 0.59, 0.57, 0.56, 0.36, and 0.40, respectively, and RMSE values of 52.43, 46.68, 55.13, 49.46, 71.13, and 69.34, respectively. Using PhenoCam-observed phenology as a baseline, SOS retrieved by phenoC++ outperforms MCD12Q2 SOS, while EOS retrieved by phenoC++ is slightly inferior to that of MCD12Q2 EOS. Moreover, compared to MCD12Q2, phenoC++-retrieved vegetation phenology yielded more effective pixels on a regional scale.

phenoC++ can rapidly produce robust vegetation phenology on a large scale. The SOS and EOS spatial distribution information on vegetation is more easily accessible through phenoC++, which will help solve large-scale ecological phenology problems.

Data availability statement

The original contributions presented in the study are included in the article/supplementary materials, further inquiries can be directed to the corresponding author.

Author contributions

YR and BR conceived and designed phenoC++ with contributions from QX and XZ. YR developed the application programming by parallel C++; BR analyzed the data and interpreted the results. YR drafted the manuscript. All authors commented on and approved the final manuscript.

Funding

The Project Supported by the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Guangzhou Science and Technology project (grant no. 2023A04J1541), and the National Natural Science Foundation of China (grant no. 42071441).

Acknowledgments

The authors thank the editor and reviewers for their constructive comments.

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

References

Ao, Z., Sun, Y., and Xin, Q. (2020). Constructing 10-m NDVI time series from landsat 8 and sentinel 2 images using convolutional neural networks. *IEEE Geoscience Remote Sens. Lett.* 18, 1461–1465. doi:10.1109/lgrs.2020.3003322

Bolton, D. K., Gray, J. M., Melaas, E. K., Moon, M., Eklundh, L., and Friedl, M. A. (2020). Continental-scale land surface phenology from harmonized Landsat 8 and Sentinel-2 imagery. *Remote Sens. Environ.* 240, 111685. doi:10.1016/j.rse.2020.111685

Caparros-Santiago, J. A., Rodriguez-Galiano, V., and Dash, J. (2021). Land surface phenology as indicator of global terrestrial ecosystem dynamics: A systematic review. *Isprs J. Photogrammetry Remote Sens.* 171, 330–347. doi:10.1016/j.isprsjprs.2020.11.019

De Beurs, K. M., and Henebry, G. M. (2008). Northern annular mode effects on the land surface phenologies of northern eurasia. *J. Clim.* 21 (17), 4257–4279. doi:10.1175/2008jcli2074.1

Defries, R. S., Field, C. B., Fung, I., Justice, C. O., Los, S., Matson, P. A., et al. (1995). Mapping the land surface for global atmosphere-biosphere models: Toward continuous distributions of vegetation's functional properties. *J. Geophys. Res. Atmos.* 100 (D10), 20867–20882. doi:10.1029/95jd01536

Elmore, A. J., Guinn, S. M., Minsley, B. J., and Richardson, A. D. (2012). Landscape controls on the timing of spring, autumn, and growing season length in mid-Atlantic forests. *Glob. Change Biol.* 18 (2), 656–674. doi:10.1111/j.1365-2486.2011.02521.x

Elmore, A. J., Nelson, D., Guinn, S. M., and Paulman, R. (2017). Landsat-based phenology and tree ring characterization. *East. U. S. For.*, 1984–2013.

Foley, J. A., Prentice, I. C., Ramankutty, N., Levis, S., Pollard, D., Sitch, S., et al. (1996). An integrated biosphere model of land surface processes, terrestrial carbon balance, and vegetation dynamics. *Glob. Biogeochem. Cycles* 10 (4), 603–628. doi:10.1029/96gb02692

Friedl, M., Gray, J., and Sulla-Menashe, D. (2019). Land surface phenology from MODIS: Characterization of the Collection 5 global land cover dynamics product. *Remote Sens. Environ.* 114 (8), 1805–1816. doi:10.1016/j.rse.2010.04.005

Gao, X., Gray, J. M., and Reich, B. J. (2021). Long-term, medium spatial resolution annual land surface phenology with a Bayesian hierarchical model. *Remote Sens. Environ.* 261, 112484. doi:10.1016/j.rse.2021.112484

Hufkens, K., Basler, D., Milliman, T., Melaas, E. K., and Richardson, A. D. (2018). An integrated phenology modelling framework in r. *Methods Ecol. Evol.* 9 (5), 1276–1285. doi:10.1111/2041-210x.12970

Jönsson, P., and Eklundh, L. (2004). TIMESAT—A program for analyzing time-series of satellite sensor data. *Comput. Geosciences* 30 (8), 833–845. doi:10.1016/j.cageo.2004.05.006

Kollert, A., Bremer, M., Löw, M., and Rutzinger, M. (2021). Exploring the potential of land surface phenology and seasonal cloud free composites of one year of Sentinel-2 imagery for tree species mapping in a mountainous region. *Int. J. Appl. Earth Observation Geoinformation* 94, 102208. doi:10.1016/j.jag.2020.102208

Kong, D., Mcvicar, T. R., Xiao, M., Zhang, Y., Peña-Arancibia, J. L., Filippa, G., et al. (2022). phenofit: An R package for extracting vegetation phenology from time series remote sensing. *Methods Ecol. Evol.* 13 (7), 1508–1527. doi:10.1111/2041-210x.13870

Li, X., Zhou, Y., Meng, L., Asrar, G. R., Lu, C., and Wu, Q. (2019). A dataset of 30 m annual vegetation phenology indicators (1985–2015) in urban areas of the conterminous United States. *Earth Syst. Sci. Data* 11 (2), 881–894. doi:10.5194/essd-11-881-2019

Menzel, A. (2002). Phenology: Its importance to the global change community. *Clim. Change* 54 (4), 379–385. doi:10.1023/a:1016125215496

ORNL Distributed Active Archive Center, Fischer A (1994). A model for the seasonal variations of vegetation indices in coarse resolution data and its inversion to extract crop parameters. *Remote Sens. Environ.* 48 (2), 220–230. doi:10.1016/0034-4257(94)90143-0

Peng, D., Wang, Y., Xian, G., Huete, A. R., Huang, W., Shen, M., et al. (2021). Investigation of land surface phenology detections in shrublands using multiple scale satellite data. *Remote Sens. Environ.* 252, 112133. doi:10.1016/j.rse.2020.112133

Peng, D., Zhang, X., Wu, C., Huang, W., Gonsamo, A., Huete, A. R., et al. (2017). Intercomparison and evaluation of spring phenology products using National Phenology Network and AmeriFlux observations in the contiguous United States. *Agric. For. Meteorology* 242, 33–46. doi:10.1016/j.agrformet.2017.04.009

Piao, S., Fang, J., Zhou, L., Ciais, P., and Zhu, B. (2006). Variations in satellite-derived phenology in China's temperate vegetation. *Glob. Change Biol.* 12 (4), 672–685. doi:10.1111/j.1365-2486.2006.01123.x

Richardson, A. D., Hufkens, K., Milliman, T., Aubrecht, D. M., Chen, M., Gray, J. M., et al. (2018). Tracking vegetation phenology across diverse North American biomes using PhenoCam imagery. *Sci. Data* 5 (1), 180028. doi:10.1038/sdata.2018.28

Ruan, Y., Zhang, X., Xin, Q., Sun, Y., Ao, Z., and Jiang, X. (2021). A method for quality management of vegetation phenophases derived from satellite remote sensing data. *Int. J. Remote Sens.* 42 (15), 5811–5830. doi:10.1080/01431161.2021.1931534

Sakamoto, T., Yokozawa, M., Toritani, H., Shibayama, M., Ishitsuka, N., and Ohno, H. (2005). A crop phenology detection method using time-series MODIS data. *Remote Sens. Environ.* 96 (3), 366–374. doi:10.1016/j.rse.2005.03.008

Savitzky, A., and Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 36 (8), 1627–1639. doi:10.1021/ac60214a047

Seyednasrollah, B., Young, A. M., Hufkens, K., Milliman, T., Friedl, M. A., Frolking, S., et al. (2019). *PhenoCam Dataset v2.0: Vegetation Phenology from Digital Camera Imagery, 2000-2018*. Oak Ridge, TN, United States: ORNL DAAC. doi:10.3334/ORNLDAAC/1674

Tan, B., Morisette, J. T., Wolfe, R. E., Gao, F., Ederer, G. A., Nightingale, J., et al. (2011). An enhanced TIMESAT algorithm for estimating vegetation phenology metrics from MODIS data. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 4 (2), 361–371. doi:10.1109/jstars.2010.2075916

Templ, B., Koch, E., Bolmgren, K., Ungersböck, M., Paul, A., Scheifinger, H., et al. (2018). Pan European phenological database (PEP725): A single point of access for European data. *Int. J. Biometeorology* 62 (6), 1109–1113. doi:10.1007/s00484-018-1512-8

Vermote, E., Walker, J. J., De Beurs, K. M., Wynne, R. H., and Gao, F. (2015). *MOD09Q1 MODIS/Terra Surface Reflectance 8-Day L3 Global 250m SIN Grid V006. NASA EOSDIS Land Processes DAAC*, 117, 381–393.Evaluation of Landsat and MODIS data fusion products for analysis of dryland forest phenology*Remote Sens. Environ.*

Yu, F., Price, K. P., Ellis, J., and Shi, P. (2003). Response of seasonal vegetation development to climatic variations in eastern central Asia. *Remote Sens. Environ.* 87 (1), 42–54. doi:10.1016/s0034-4257(03)00144-5

Zhang, X., Friedl, M. A., Schaaf, C. B., Strahler, A. H., Hodges, J. C. F., Gao, F., et al. (2003). Monitoring vegetation phenology using MODIS. *Remote Sens. Environ.* 84 (3), 471–475. doi:10.1016/s0034-4257(02)00135-9

Zhang, X., Liu, L., Liu, Y., Jayavelu, S., Wang, J., Moon, M., et al. (2018). Generation and evaluation of the VIIRS land surface phenology product. *Remote Sens. Environ.* 216, 212–229. doi:10.1016/j.rse.2018.06.047

Zhang, X., Wang, J., Henebry, G. M., and Gao, F. (2020). Development and evaluation of a new algorithm for detecting 30 m land surface phenology from VIIRS and HLS time series. *Isprs J. Photogrammetry Remote Sens.* 161, 37–51. doi:10.1016/j.isprsjprs.2020.01.012

Zhou, D., Zhao, S., Zhang, L., and Liu, S. (2016). Remotely sensed assessment of urbanization effects on vegetation phenology in China's 32 major cities. *Remote Sens. Environ.* 176, 272–281. doi:10.1016/j.rse.2016.02.010

frontiers | Frontiers in Environmental Science

TYPE Original Research
PUBLISHED 12 April 2023
DOI 10.3389/fenvs.2023.1131897

OPEN ACCESS

EDITED BY
Huadan Zheng,
Jinan University, China

REVIEWED BY
Kun Liu,
Chengdu University of Information Technology, China
Gaopeng Lu,
University of Science and Technology of China, China

*CORRESPONDENCE
Qilin Zhang,
qlzhang@nuist.edu.cn

SPECIALTY SECTION
This article was submitted to Environmental Informatics and Remote Sensing, a section of the journal Frontiers in Environmental Science

RECEIVED 26 December 2022
ACCEPTED 03 April 2023
PUBLISHED 12 April 2023

CITATION
Zhou J, Zhang Q, Zhang J, Dai B, Li J, Wang Y and Gu J (2023), Evaluation and revision of long-range single-site lightning location accuracy considering the time delay of ground wave. *Front. Environ. Sci.* 11:1131897. doi: 10.3389/fenvs.2023.1131897

Evaluation and revision of long-range single-site lightning location accuracy considering the time delay of ground wave

Jiahao Zhou, Qilin Zhang*, Junchao Zhang, Bingzhe Dai, Jie Li, Yao Wang and Jiaying Gu

Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters (CIC-FEMD)/ Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET)/Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology, Nanjing, China

Detecting the distance and orientation of long-distance thunderstorms has very important practical significance. The multi-station lightning location system relies on a high-precision time module and good network communication capabilities, but in some cases these conditions cannot be met, but there is still a need for lightning activity monitoring, and it is very important to establish a single-site lightning location system. In this paper, we have established a long-distance single-site lightning location station, and in order to improve the accuracy of distance estimation, a numerical algorithm is used to obtain the relationship between the ground wave arrival time delay and the propagation distance, and it is used to revise the time difference between the peak value of the skywave and ground wave. Moreover, we used multi-station lightning location data to revise the site-error in magnetic direction finder method to improve the accuracy of the direction calculation. The results show that the effective detection range of the single-site we have been established is 200 km–2000 km, and the revised average direction deviation dropped from 12.3° to 8.6°. The verification results of thunderstorms within the effective detection range show that the relative error of single-site lightning location is 8.4%–18.6% after the revision.

KEYWORDS

single-site sensor, lightning location, ground wave arrival time delay, site-error, location accuracy

1 Introduction

Lightning location data has become an integral part of meteorological data observations, especially in the early warning of severe weather, which has led to continuous improvements in ground-based and space-based lightning location systems (DiGangi et al., 2022). Now, ground-based lightning location systems (LLSs) typically use multiple detection stations (≥4 stations) to detect lightning electromagnetic pulses (LEMP) from lightning radiation (Koshak et al., 2004; Pohjola and Mäkelä, 2013; Wu et al., 2018; Liu et al., 2020; Wang et al., 2020). Lightning discharges generate broadband electromagnetic radiation, mainly in the frequency band from 1 Hz to 300 MHz (Gu et al., 2022), among them, electromagnetic pulses in the very-low-frequency (VLF; 3–30 kHz) band are widely used in long-range lightning location. Such signals are also called sferics, which can propagate thousands of kilometers in the Earth-ionosphere waveguide (EIWG) with little attenuation (~2–3 dB/

1,000 km) (Ammar and Ghalila, 2020). The LLSs mainly use the time of arrival (TOA) method and the time difference of arrival (TDOA) method, which uses the time difference between the LEMP signal and the station to location, so LLSs can obtain high lightning location accuracy, and the average error is generally on the order of hundred meters (Zhang et al., 2010; Shi et al., 2017; Wang et al., 2021). This also requires that each station is equipped with a high-precision time module and good network communication so that the LEMP collected at each station can be aggregated to the data processing center to give real-time lightning location results (Wu et al., 2018).

Due to the constraints of the site environment, some places do not have good network transmission capabilities or have many restrictions on transmission with external networks, a single-site can be used to determine the lightning location and also help save operation and maintenance (O&M) costs. Because of the good mobility and flexibility of single-site, it has unique superiority in lightning location in civil aviation, military activities and other fields. Single-site lightning location technology is a combination of propagation distance estimation and direction-finding technology, what matters is how to improve the accuracy of single-site lightning location. In contrast, the magnetic direction finder (MDF) method can detect sferics thousands of kilometers away, a typical magnetic sensor usually consists of two orthogonal magnetic antennas. However, the method suffers from an angular uncertainty of 180°, and an electric field antenna is usually added to determine the lightning polarity, thus eliminating the angular uncertainty (Herrman et al., 1976; Rakov, 2013; Nag et al., 2014). Ramachandran et al. (Ramachandran et al., 2007) used the period and delay extracted from the quasi-periodic waveform of the electric field received at the station to estimate distance, the error of the distance estimate reported in the text is 8.8%. Nagano et al. (Nagano et al., 2007) used sferics pulses to locate lightning (direction and distance) at close range with an error of about 10% with less interference from noise. Mostajabi et al. (Mostajabi et al., 2019) combines the Electromagnetic Time Reversal (EMTR) and Machine Learning (ML) to propose a new method for single-site lightning location, which is more suitable for application in mountainous areas, the model gives an average error of 253 m for six return strikes (RS) occurring at a distance of 14.7 km. Wang et al. (Wang et al., 2022) proposed a single-site location method, The method is based on deep learning and predicts the LEMP propagation distance by learning the characteristic points of sferics waveforms with different propagation distances. Within the detection range of 1,000 km, the relative error of this method is 4.91%–15.26%. Zhang et al. (Antunes de Sa and Marshall, 2020) used a single-station magnetic orientation method to locate narrow bipolar events (NBEs), they found that the positive NBEs produced between 7 km and 15 km and negative NBEs produced above 14 km. Andre et al. (Zhang et al., 2016) used the time difference between ground wave and the first skywave combined with ML to realize the lightning distance estimation by a single-site, 68% of the data error is within 32 km.

In single-site lightning location that relies on MDF to obtain lightning directions, there are unavoidable system errors that arise from two main factors: random errors (from non-vertical lightning channels, background noise and electronics differences) and site errors (from the surrounding terrain and secondary radiation from surrounding conductive structures), with the former typically having an effect of 1–2° on the orientation results, while the latter has a much larger effect, possibly in the range of 10–30° (Mach et al., 1986; Chen et al., 2013; Lu et al., 2017). Due to the curvature of the earth and the finite conductivity of the ground, the electromagnetic waves generated by lightning will be attenuated when propagating along the ground at long distances, and there will be some deviation when estimating the lightning distance (Honma et al., 1998; Shao and Jacobson, 2009). To obtain the direction and distance of lightning relative to a single-site, we use a combination of the ionosphere reflection model and MDF, and also use an analytical algorithm to simulate the relationship between ground wave arrival time delay and distance and use LLSs established in China to obtain the site-error in single-site lightning location, revise the estimation results of distance and direction respectively, so as to obtain more accurate single-site lightning location results.

2 Data analysis method

2.1 Description of experimental instrument and data

The multi-station lightning location data used in this paper are from a VLF LLSs established in China (Li et al., 2022; Zhang et al., 2022), the single-site equipment installed in Nanjing, Jiangsu Province, China. The single-site receives lightning signals in the range of about 100 Hz to 80 kHz, records LEMP waveform data by continuous acquisition, and stores them in a hard disk drive (HDD). The sensors used for detection include a fast electric field antenna and two orthogonal magnetic antennas, induce the electric field signal and the magnetic field signal in the east-west (EW) and north-south (NS) directions, respectively. The sampling frequency of the equipment is 1MS/s, and the GPS receiver provides an output of one-pulse-per-second output (1 PPS) as a reference source for tagging data sample times, with an accuracy of ±50 ns. A trigger threshold will be set according to the background noise, and the original waveform data greater than the threshold will be extracted. The record length of the extracted waveform is 1000 μs, and the pre-trigger time is 300 μs, while the extracted signal is de-noised by the Modified Empirical Wavelet Transform (MEWT) method based on the Empirical Wavelet Transform (EWT) (Dai et al., 2022).

The data used in this paper are from August 7, and August 11-15 August 2022, during which frequent thunderstorm activities occurred, the location data from August 11-August 15 are used to revise the site-error by comparing it with the multi-station lightning location results, while some sustained all-day thunderstorm activity recorded by multiple stations during these days will also be used to calculate the ionosphere equivalent reflection height, and the data from August 7 are used to verify the effect of the revision. Figure 1 shows the lightning that occurs in different areas on different days. These data will be used to calculate the trend of the ionosphere equivalent reflection height over a 24-h period. Since the data in the same area on August 12 is not continuous for most of the time, this will affect the reliability of

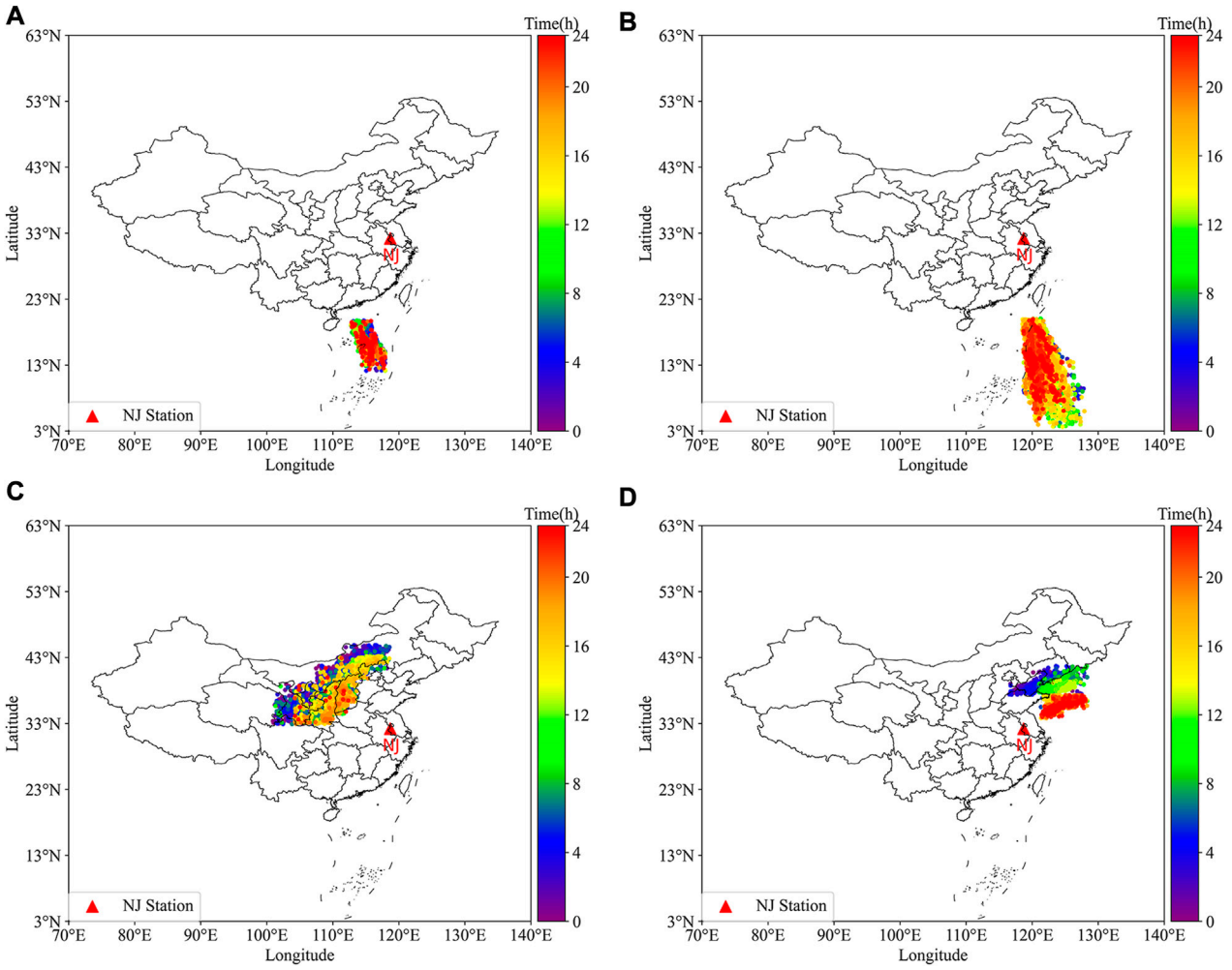


FIGURE 1
The distribution of lightning events in selected regions **(A)** From 00:00 to 23:59 on 11 August2022 (local time); **(B)** From 00:00 to 23:59 on 13 August2022 (local time); **(C)** From 00:00 to 23:59 on 14 August2022 (local time); **(D)** From 00:00 to 23:59 on 15 August2022 (local time).

the results, so the dates of this day are not used to calculate the ionosphere equivalent reflection height. Since the ionosphere reflection regions calculated in Figures 1A–D are relatively close to each other. Therefore, in this paper, the ionosphere reflection region calculated by Figures 1A, B is referred to as region A, and the ionospheric reflection region calculated by Figures 1C, D is referred to as region B.

2.2 Direction calculation

The MDF method has been widely used in the study of lightning location, assuming that lightning is a dipole discharge approximately perpendicular to the ground, and two orthogonal magnetic antennas to measure the horizontal magnetic field generated by lightning. According to the ratio of the induced voltage of the lightning magnetic field in the two vertical directions, the direction of the lightning relative to the detection station can be determined. For long-distance lightning location, the scale of the discharge channel is smaller relative to the detection distance, and the hypothesis of dipole is reasonable. According to the above-mentioned method, the value of the azimuth φ of the lightning occurrence position relative to the single-site is:

$$\varphi = arctan\frac{B_{NS}}{B_{EW}} \tag{1}$$

In Eq. 1, B_{NS} and B_{EW} are the induced electric potentials generated on the NS and EW magnetic antennas, respectively. The azimuth angle φ is defined as the angle rotated by clockwise rotation starting from north direction. Also, in combination with the received LEMP electric field waveform, the polarity of the lightning can be determined, thus eliminating the problem of 180° angle ambiguity. The schematic diagram of the site error is shown in Figure 2.

The accurate azimuth angle φ' can be obtained from the LLSs, and the angle difference between φ' and φ is the site error γ. The site error calculation formula is:

$$\gamma = \varphi - \varphi' \tag{2}$$

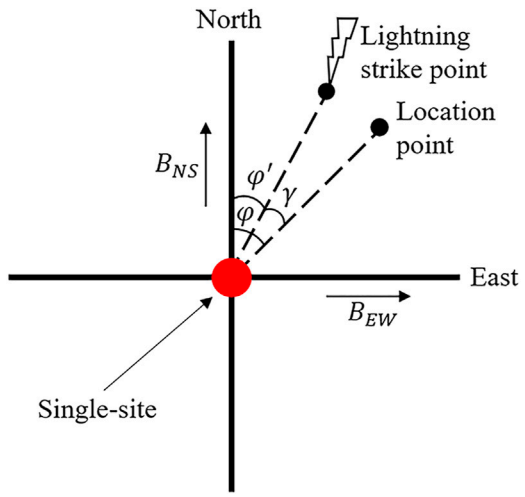


FIGURE 2
Effect of site error on direction calculation.

2.3 Time delay revision

2.3.1 Ground wave time delay revision

However, when electromagnetic waves radiated by lightning discharges propagate along the surface of the earth, they will be affected by irregular ground conductivity distribution and complex terrain on the propagation path, especially during long-distance propagation. The propagation speed of electromagnetic waves will be less than the speed of light, and the arrival time will lag behind the ideal arrival time, this phenomenon has appeared in both simulation and actual observations (Shao and Jacobson, 2009; Hou et al., 2018). In this article, t_d is defined as the ground wave arrival time delay due to the increase in the propagation distance.

The numerical algorithm used in this paper is based on the propagation theory of ground waves in finite conductivity proposed by Hill and Wait (Hill and Wait, 1980) Considering the propagation effect the limited ground conductivity and the propagation effect due to the curved surface of the earth, the attenuation coefficient in the frequency domain is calculated as (Wait, 1974):

$$W = e^{-j\pi/4}\sqrt{\pi x}\sum_{s=1}^{\infty}\frac{e^{-jxt_s}}{t_s - q^2} \tag{3}$$

$$x = (k_0R/2)^{\frac{1}{3}}(d/R) \tag{4}$$

$$q = -j(k_0R/2)^{\frac{1}{3}}\Delta \tag{5}$$

$$\Delta = k_0/k\sqrt{1-(k_0/k)^2} \tag{6}$$

$$k = \omega\sqrt{\varepsilon_r\varepsilon_0\mu_0 - j\sigma\mu_0/\omega} \tag{7}$$

In Eqs 3–7: t is the normalized ground surface impedance, k and k_0 are the wave numbers of electromagnetic waves propagating in soil and vacuum, respectively, d is propagation distance, R is the radius of the earth, ω is the angular frequency, ε_0, μ_0 are the dielectric constant and magnetic permeability in the vacuum, ε_r, σ are the relative dielectric constant and conductivity of the ground respectively, t_s is the roots of the complex equation, the complex equation is:

$$w_1'(t) - qw_1(t) = 0 \tag{8}$$

$w_1(t)$ is expressed as:

$$w_1(t) = \sqrt{\pi}\left(Bi(t) - jAi(t)\right) \tag{9}$$

$Ai(t)$ and $Bi(t)$ are the Airy functions.

In this paper, three different lightning current sources are considered as the typical first return stroke (RS), subsequent RS, and dipole source. The model used is a modified transmission line model with exponential current attenuation with height (MTLE) mode (Nucci, 1988). Among them, the current source of the first RS and the subsequent RS assumes that as the lightning current travels on the lightning channel, the amplitude decreases exponentially with the increase of height, the waveform is in the form of a double Heidler function (Heidler et al., 1999). While the current waveform of the dipole source is assumed to be uniform along the lightning discharge channel.

Table 1 shows the typical lightning-based current waveforms of the First RS and Subsequent RS commonly used in engineering calculations (Rachidi et al., 2001).

i_{01} and i_{02} are the peak current of the breakdown current and corona current respectively, τ_{11} and τ_{12} are the rising and falling edge times of the breakdown current, τ_{21} and τ_{22} are the rising and falling edge times of the corona current.

The return stroke current waveform of the dipole source is assumed to be uniform along the lightning discharge channel and given as:

$$I(t) = I_0\frac{\nu_0}{\gamma}\left[e^{-at} - e^{-bt}\right]\left[1 - e^{-\gamma t}\right]\Big/H \tag{10}$$

Where $I_0 = 20$ kA, $\nu_0 = 8 \times 10^7$ m/s, $\gamma = 3 \times 10^4 s^{-1}$, $a = 2 \times 10^4 s^{-1}$, $b = 2 \times 10^5 s^{-1}$, and H is the lightning discharge channel length (Dennis and Pierce, 1964; Hu and Cummer, 2006).

The ground conductivity used in this paper is taken to be a typical value of 0.01 S/m, which corresponds to the wet ground condition. Figure 3A shows the current waveforms of the typical first RS and subsequent RS, Figure 3B shows the current waveform of the dipole source. The rising edge of the subsequent RS is steeper than the other two current sources and contains more high-frequency components. Figure 3C shows the ground wave arrival time delay at different distances, compared with the arrival time propagated at the speed of light. That is, the difference between the ideal ground wave arrival time and the actual ground wave arrival time, which is t_d as defined above. It can be seen from Figure 3C, the delay time increases approximately linearly with distance, and the difference between the three current sources is small. Due to the different frequency bands of the three current sources, the high-frequency component will arrive earlier than the low-frequency component during the long-distance propagation due to the propagation effect, and the subsequent RS contains more high-frequency components compared with the other two current sources, so the t_d of the subsequent RS at the same propagation distance is relatively small compared with the other two current sources. In order to revise the ground wave arrival time delay, the average curve is used to represent the relationship between the ground wave time delay and the propagation distance. Further from the data, it can be obtained that the peak arrival time of lightning electromagnetic waves is delayed by 0.9 μs on average for every 100 km increase in

TABLE 1 Typical lightning current waveform parameters of the First RS and Subsequent RS.

Type	i_{01} (kA)	τ_{11} (μs)	τ_{12} (μs)	i_{02} (kA)	τ_{21} (μs)	τ_{22} (μs)
First RS	28	1.8	95	-	-	-
Subsequent RS	10.7	0.25	2.5	6.5	2	230

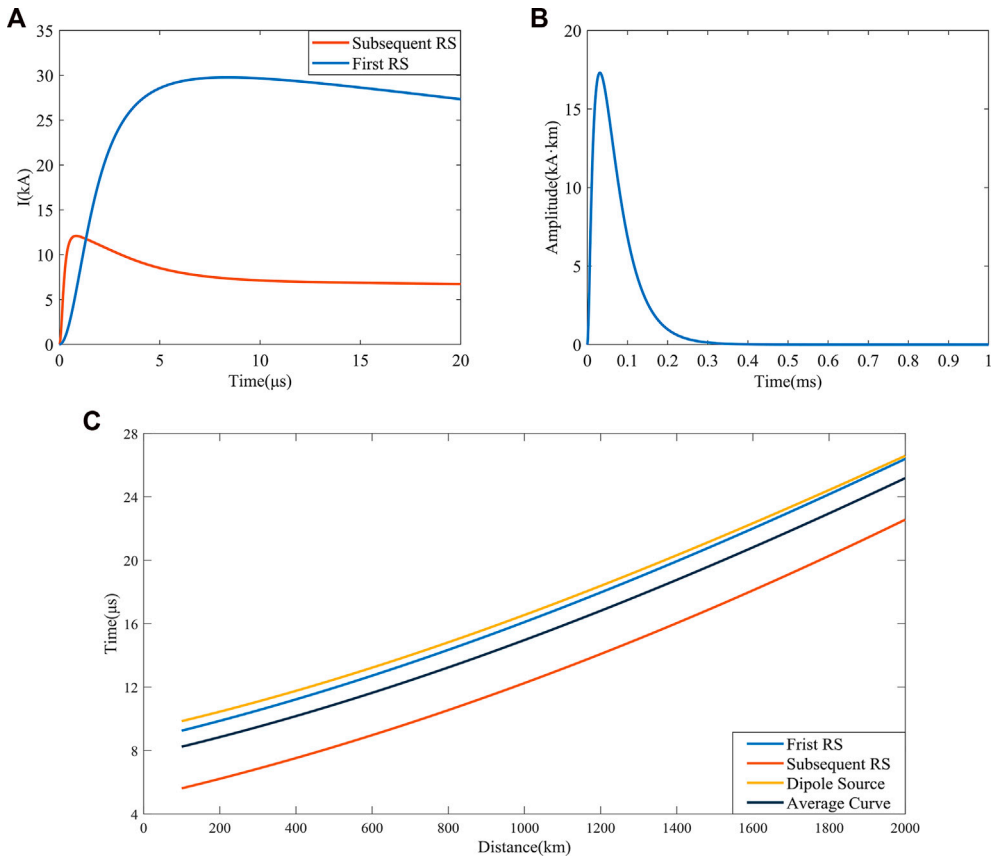


FIGURE 3
The lightning current waveforms and ground wave arrival time delay. **(A)** Current waveforms of the typical first RS and subsequent RS; **(B)** current waveform of the dipole source; **(C)** Delay of ground wave peak arrival time compared to *d*/*c*.

propagation distance, which indicates that for long-range LEMP, the ground wave arrival delay time brings non-negligible impact on the accuracy of distance estimation. According to the research results of other scholars, it demonstrates that propagation over the land for the distance of about 130 km with a conductivity of 3 mS/m, the peak of the RS pulse was delayed by an average of 1.8 μs (Han and Cummer, 2010a). The simulation results of Shao et al. (Shao and Jacobson, 2009) show the leading edge and the peak are delayed by 5 μs and 13 μs, respectively, at a distance of 1,000 km. These results are consistent with the simulation results given in this paper, so the average curve can be used to revise the propagation effect.

2.3.2 Ionospheric reflection model

Figure 4A shows the geometric model of sferics propagating in EIWG. The signal propagates through multiple specular reflections of the earth's surface and the ionospheric D layer in the EIWG, the signal received by the station that arrives directly along the surface is called ground wave, and the signal that arrives after reflection from the ionosphere is called skywave (Li et al., 2022). Figure 4B shows a set of electric and magnetic field waveforms received by the single-site. A black asterisk is used to indicate the peak points of the ground wave and the first skywave. The propagation speed of electromagnetic waves is assumed to be the speed of light in the model, and the ground is a good conductor. With this model, it is possible to calculate the ionosphere equivalent reflection heights for different directions and regions based on the lightning location results obtained from LLS, and compare if significant differences exist. Using the ionosphere equivalent reflection heights, the lightning distance occurred can also be estimated.

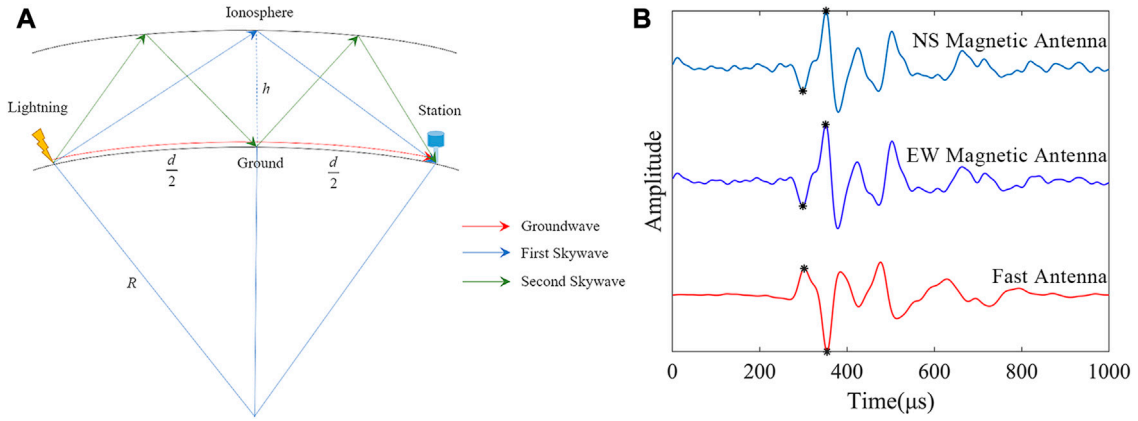


FIGURE 4
Lightning signal propagation in EIWG. **(A)** Schematic diagram of the first and second skywave reflections in EIWG; **(B)** Lightning waveform.

The reflection height (H_1) of the first sky wave can be derived in the following way (Somu et al., 2015).

$$H_1 = R\left[cos\left(\frac{d}{2R}\right) - 1\right] + \sqrt{\left\{R^2\left[cos^2\left(\frac{d}{2R}\right) - 1\right] + \left(\frac{ct_1 + d}{2}\right)^2\right\}} \tag{11}$$

In Eq. 11, R is the radius of the earth, and d is the spherical distance between the lightning and the station. c is the speed of light in free space, and t_1 is the arrival time difference between the ground wave and the first skywave obtained the LEMP waveform received by the detection station.

However, when propagating over long distances, using the above equation, the estimation results are often biased due to the ground wave peak time delay. In general, electromagnetic wave propagation over terrain with lower ground conductivity brings a larger arrival time delay. The ground wave arrival time delay caused by the propagation effect is considered in this work, and the theoretical time difference between the arrival time of the skywave and the ground wave is assumed to be T. The time difference between the skywave and the theoretical ground wave should be expressed as Eq. 12:

$$T_1 = t_1 + t_d \tag{12}$$

t_d is ground wave peak time delay at a specific distance obtained from Figure 3.

We substitute the distance (d) and the revised time difference between the first skywave and ground wave (T_1) of the lightning event given in Figure 1 into Eq. 11 to calculate the continuous 24-h variation of the ionosphere equivalent reflection height. Since lightning occurs far away from the station, the altitudes of the source and single-site are ignored in the calculations, and the accuracy of the calculation results will not be affected.

The model can also be used to estimate the lightning distance by bringing the time of receiving the LEMP waveform and the ionosphere equivalent reflection height corresponding to that time into Eq. 11, and the distance between the lightning strike point and the single-site is obtained by iterative solution.

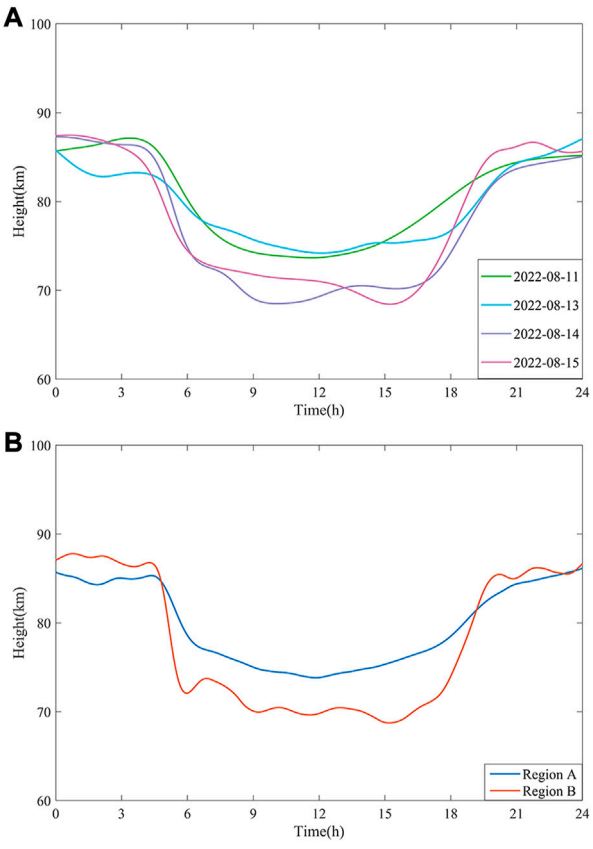


FIGURE 5
Ionosphere equivalent reflection height variations of continuous 24 h. **(A)** On different dates; **(B)** On different regions.

In Section 2.2, and Section 2.3, we obtained the direction and distance of the lightning relative to the single-site, respectively. While the location of the single-site is known, we use the Bessel geodesic problem algorithm to first establish an auxiliary sphere

centered on the ellipsoid with any length as the radius, and then project ellipsoid elements onto the spherical surface according to certain conditions, solve the earth problem on the spherical surface, and finally convert the obtained spherical elements into ellipsoid elements according to the corresponding projection relationship. The reference system used in this paper is the WGS84 coordinate system, from which we can finally rely on a single-site to obtain the location of the lightning strike point.

3 Results and analysis

3.1 The ionosphere equivalent reflection height changes over 24-h

Figures 5A, B shows the line graphs of ionosphere equivalent reflection height with time delay considered. From Figure 5A, it can be seen that even in different regions, the difference in the ionosphere equivalent reflection height is not very large, especially at night (00:00–05:00). The reason for the fluctuations during the daytime period (06:00–18:00) on August 14 and August 15 was that the amount of data during that period was very small, which also brought unexpected fluctuations. The trend of changes in the H_1 calculated in different regions is similar. Due to the small amount of data in the time period from 14:30 to 17:00 on August 11, the curve has increased in advance, which is somewhat different from the situation in other days. In order to reduce the deviation that occurs when there is less data, we further fit the two curves obtained in the same area into one, and the result is shown in Figure 5B. Among them, Region A is mainly the southwestern part of Fujian Province, China, while Region B is mainly the southwestern part of Shandong Province, China, and the border with Hebei Province.

In Figure 5B, during the night time from 00:00 to 05:00, the H_1 in region A and region B fluctuated in the range of 84 km–85 km and 86km–87km, respectively. At 05:00, the height was rapidly dropping, the drop lasted about one hour. Between 06:00 and 18:00, the height was in a relatively stable state. Region A and region B fluctuated in the range of 73 km–77 km and 68 km–73 km, the H_1 in region A is slightly higher. At the nighttime (from 20:00 to 23:59), region A and region B fluctuate in the range of 84 km–85 km and 84 km–86 km, respectively. The Consultative Committee on International Radio (CCIR) recommends the ionosphere reflection height at night and noon is about 87 km and 70 km, respectively (Zhou et al., 2021). At night, Han et al. (Han and Cummer, 2010b) revealed an average ionospheric height of 84.9 km in the D region, ranging from 82.0 km to 87.2 km which is close to the results obtained in this paper. In the subsequent distance estimation, according to the time of lightning occurrence, the ionosphere equivalent reflection height corresponding to region A or region B will be selected according to the direction of the lightning strike point relative to the Nanjing station and brought into Eq. 11 for iterative calculation.

3.2 Distance estimation

In this section, we bring the revised arrival time difference between the ground wave and the first skywave (T_1) and the ionosphere equivalent reflection height (H_1) into Eq. 11 and solve the distance between the lightning strike point and the single-site by an iterative method.

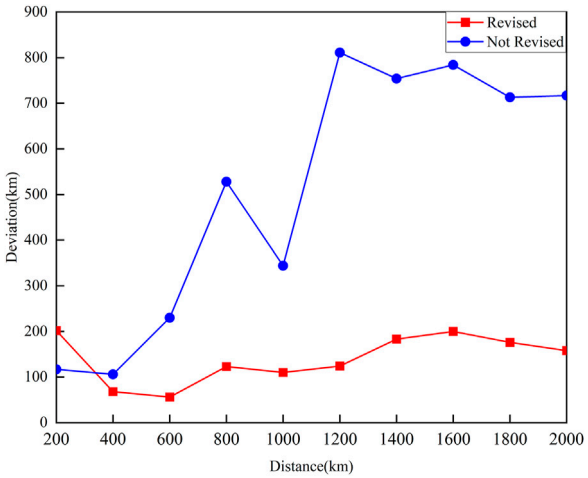


FIGURE 6
Comparison of the mean deviations of the revised and not revised methods at different distances.

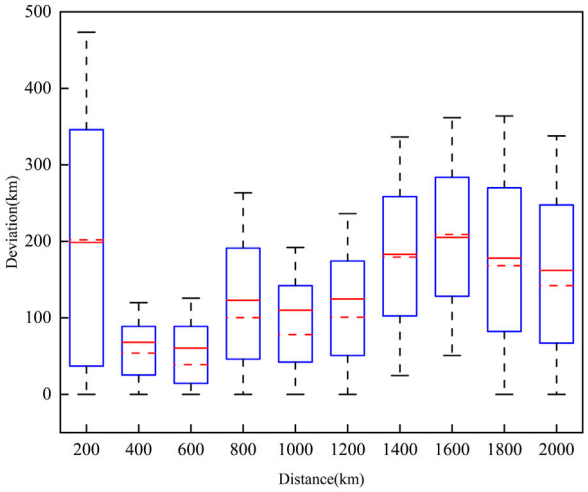


FIGURE 7
The changing trend of the deviation with distance.

Figure 6 shows the effect of the revised ground wave arrival time delay on the accuracy of distance estimation, it can be seen that when the lightning strike point is close to the station, the arrival time delay of the ground wave is small, the revised method brings greater errors. And as the propagation distance increases, the impact of the propagation effect becomes more pronounced, the time difference between the ground wave and the skywave in the actual received waveform is less than the ideal situation, resulting in the estimated distance is smaller than the actual propagation distance, thus bringing a greater deviation. As the propagation distance increases, the distance deviation of the revised method decreases

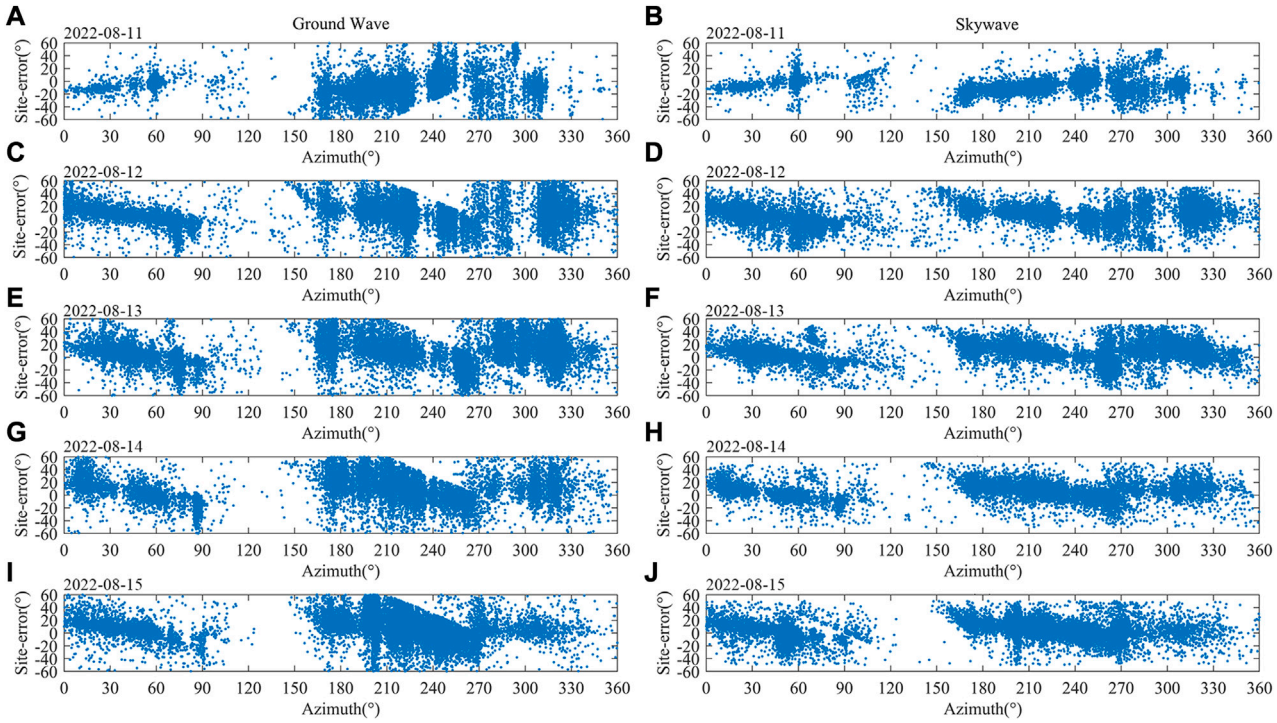


FIGURE 8
Scatter graph of site error distribution. **(A)** On August 11, based on the ground wave peak; **(B)** On August 11, based on the skywave peak; **(C)** On August 12, based on the ground wave peak; **(D)** On August 12, based on the skywave peak; **(E)** On August 13, based on the ground wave peak; **(F)** On August 13, based on the skywave peak; **(G)** On August 14, based on the ground wave peak; **(H)** On August 14, based on the skywave peak; **(I)** On August 15, based on the ground wave peak; **(J)** On August 15, based on the sky wave peak.

rapidly, and reaches the lowest point at 600km, then rises, and remains relatively stable at 800 km–1200 km. When the propagation distance is greater than 1200 km, the distance deviation of the revised method has a small increase, and the distance estimation deviation has increased by 59 km. When the propagation distance is greater than 1600 km, the distance estimation deviation has decreased slightly, which is 46 km lower than the maximum average deviation. Overall, the distance deviation of the revised method remains within a relatively stable and credible range, and there is no failure as the propagation distance increases.

Figure 7 gives the changing trend of the deviation with distance from lightning to the station, where the red dashed line represents the median deviation and the red solid line represents the mean deviation. When lightning occurs within 400 km from the station, the deviation is large. This is because when the propagation distance is short, the arrival time delay of the ground wave is not obvious, and the revised arrival time difference between the ground wave and skywave is greater than the not revised, resulting in a larger deviation. With the increase of the propagation distance, affected by the propagation effect, the arrival time of the ground wave lags behind the arrival time of the lightspeed. In this way, the revised method mentioned in this paper can be used to compensate for the ground wave arrival time well, so that it is closer to the ideal situation. Therefore, in the range of 400 km–2000 km, compared with the multi-station lightning location dates, the mean deviation of the estimated distance obtained by the single-site is 7.9%–17.0%.

3.3 Site-error revision

Figure 8 shows the scatter plot of the site error distribution obtained from August 11 to August 15. The five graphs in the left column are the calculated results based on the ground wave peak in the signal received by the two orthogonal magnetic antennas, while five graphs on the right column are the results calculated based on the skywave peak. The azimuth is defined as the angle rotated by clockwise rotation starting from north, the due north direction of Nanjing Single station is 0°. It can be seen from the figure that the number of lightning that occurred in the direction of 120°–150° in these 5 days is very small, while other times the scatter shows almost the same distribution. At the same time, it can be seen from the distribution of scattered points that the site errors obtained by the skywave peak are more concentrated, which also means that if the results obtained by the skywave peak are used to revise the site errors, the effect will be better.

The fitting results of the scatter plot are given in Figure 9, which shows more intuitively that the site-error obtained from the skywave have a smaller fluctuation range and are closer to 0° compared to the results obtained from the ground wave. It can be seen from the figure that the trend of changes in the results obtained from different dates is close. Among them, there was a reverse result appears on August 10 compared with other dates. Which may be due to the fact that there are fewer data in this range, resulting in fewer scattered points affecting the fitting result, which is not statistically significant. It may also be due to changes in the surrounding background noise or

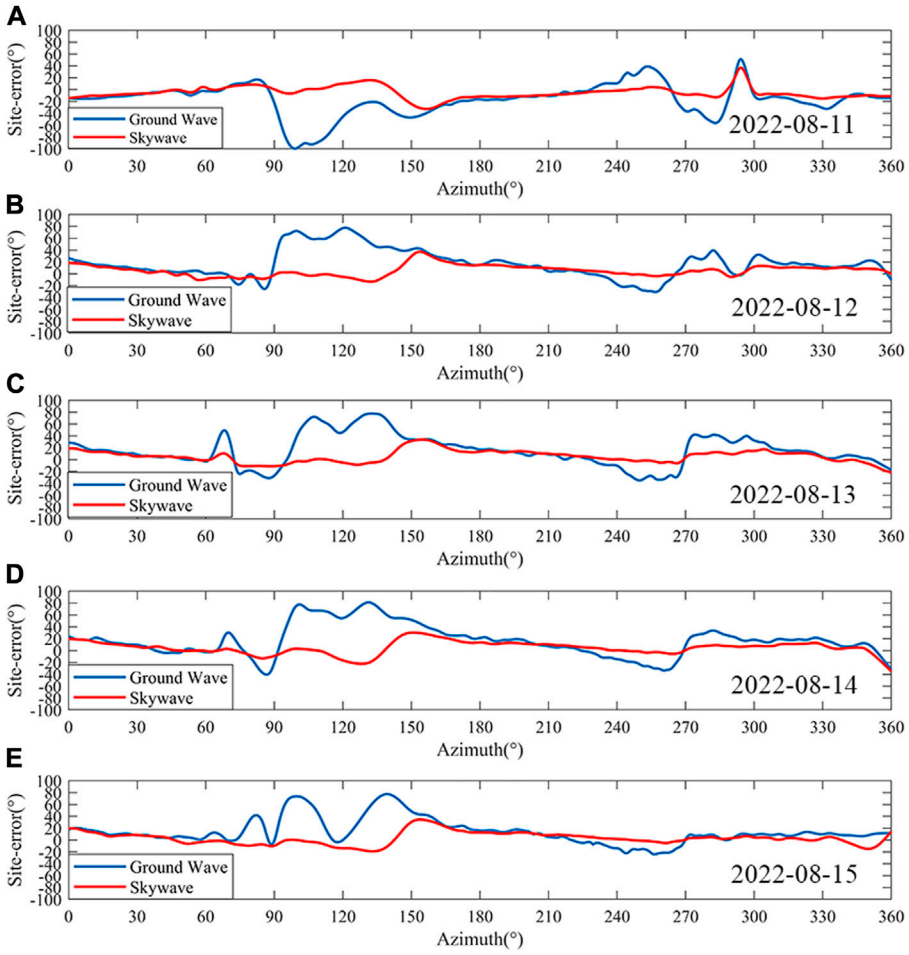


FIGURE 9
Fitting results for site error on different dates. **(A)** On August 11; **(B)** On August 12; **(C)** On August 13; **(D)** On August 14; **(E)** On August 15.

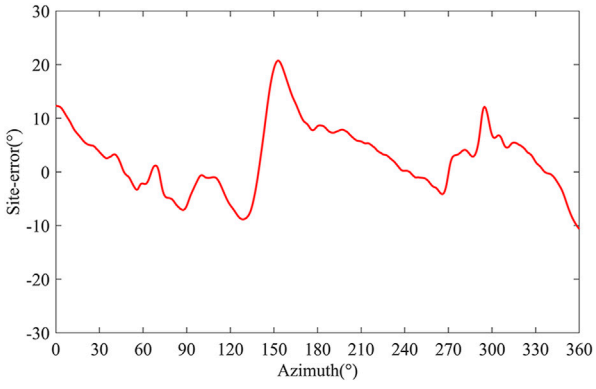


FIGURE 10
Site-error revision curve.

interference from other equipment. Due to data limitations, it is impossible to determine whether this interference is periodic. After obtaining more data, it can be well explained whether this situation occurs by accident or is periodic, which also has practical significance for the siting of the station. In this article, in order to minimize its impact on the revised result, the five-day curve is further fitted to obtain the final revised curve shown in Figure 10. And this paper will use the site-error revision curve obtained by skywave to revise the azimuth calculated by the single-site.

As can be seen from Figure 10, the peak of the positive deviation of the site-error appears near the three angles of 0°, 150°, and 300°, and the peak of the reverse deviation is reached near 130° and 270°, respectively. The largest site-error is 20°, which occurs when the azimuth is 152°. The change of site error shows a trend of slow decline and fast rise. The occurrence of larger site errors has a certain periodicity, which may be related to the background noise of the site. In the results given by Lu et al. (Lu et al., 2017), this phenomenon also occurs at some sites.

3.4 Single-site lightning location

In this section, we selected lightning activities at different distances from the single-site which occurred on 7 August 2022 (local time) to evaluate the accuracy of single-site lightning location. The location results of the single-site will be

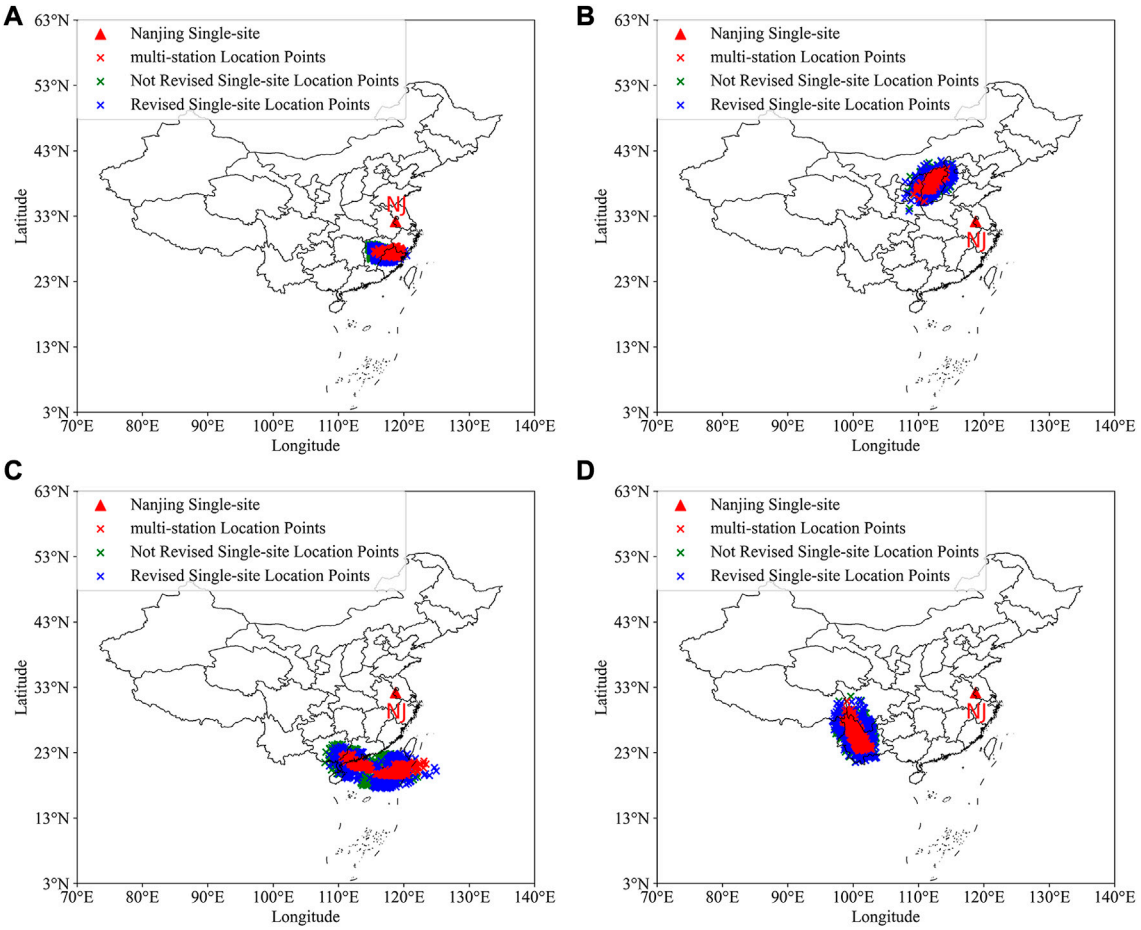


FIGURE 11
Lightning location results of single-site and multi-station. **(A)** In the range of 400–600 km; **(B)** In the range of 800–1000 km; **(C)** In the range of 1200–1400 km; **(D)** In the range of 1800–2000 km.

compared to the location results obtained by the multi-station lightning location network we built. The accuracy of multi-station location results has been compared with Arrival Time Difference Thunderstorm Detection system (ADTD) in the previous work. The average positioning error is 4.32 km, and the standard deviation is 2.46 km (Zhang et al., 2022). From Figure 11, we can see that the single-site lightning location results are more dispersed compared to the multi-station results. However, compared with the not revised location results, the revised location results are closer to multi-station results. It can be seen that as the distance increases, the multi-station lightning location results are more concentrated, while the single-station lightning location results are more dispersed, this is more likely due to errors in distance estimation.

We give the location error of the single-site at different distances (multi-station results as a reference). Figure 12 shows the relationship between the trend of deviation and the distance from lightning to the single-site, where the red dashed line represents the median deviation and the red solid line represents the mean deviation. It can be seen that as the distance increases, the average error and the median deviation both increases. When the distance is in the range of 600–800 km,

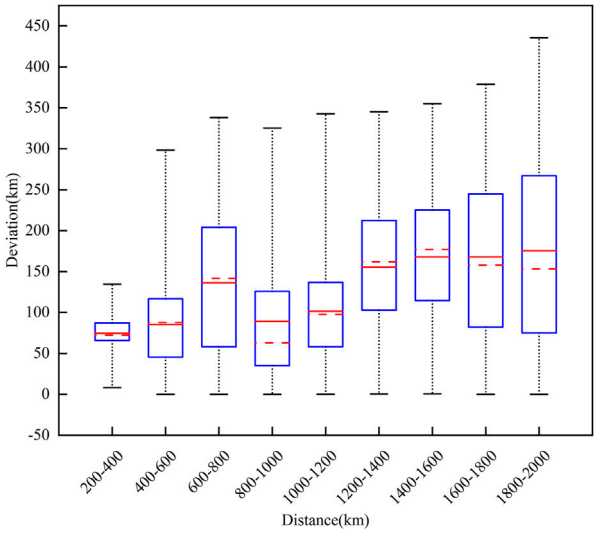


FIGURE 12
The changing trend of the single-site lightning location deviation with distance.

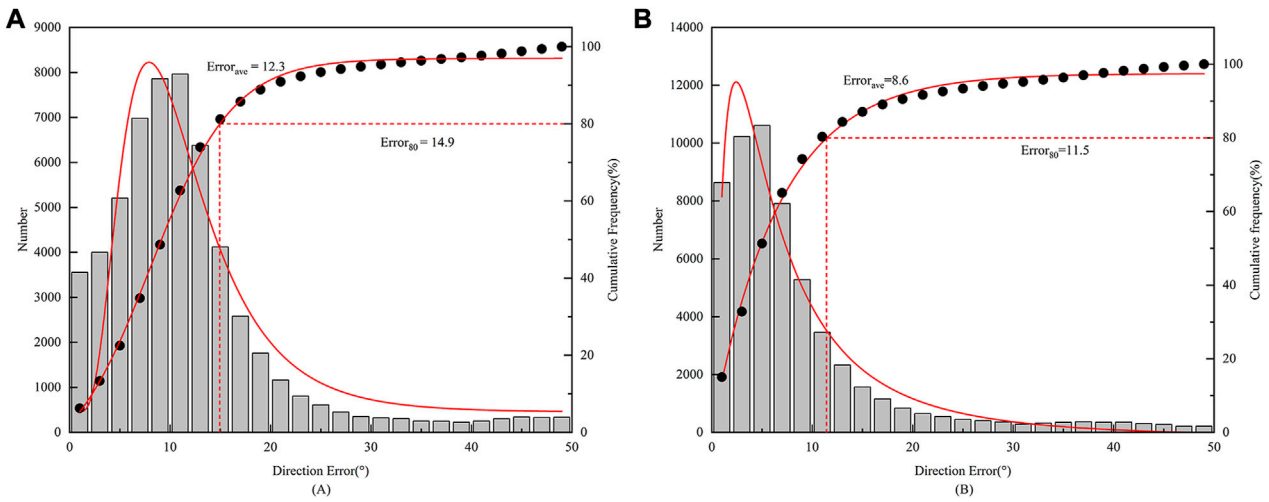


FIGURE 13
Compared with the multi-station results, the distribution of direction error of the single-site. **(A)** Not revised the site-error; **(B)** Revised the site-error.

TABLE 2 Comparison between existing methods and methods in this way on single-site lightning location.

Method	Distance (km)	Error (%)
Nagano et al. (Nagano et al., 2007)	200 km	12.5%
Chen et al. (Chen et al., 2015)	<130 km	15.5%–20%
Wang et al. (Wang et al., 2022)	<1000 km	4.91%–15.26%
Method in this paper	200km–2000 km	8.4%–18.6%

the deviation has a big change, and in subsequent longer distances, the deviation has decreased again. In the two farthest distance groups, the average deviation decreased, which shows that the single-site lightning location method proposed in this paper has practical application value in long-distance lightning location, and the results are still credible. In the effective detection range (200km–2000 km), the deviation range of this single-site lightning location is 8.4%–18.6%.

4 Discussion

In order to better illustrate the effectiveness of using multi-station lightning location results to revise site-error, Figure 13 shows the azimuth results obtained by the single-site using the not revised method and revised method (multi-station results are for reference). It can be clearly seen that the direction error revised by using the site-error revision curve is smaller and the distribution is more concentrated. Smaller direction error account for most of the total number. The number of smaller direction error after the revision has been greatly improved compared to when it was not revised. After revision, the average direction error has been reduced from 12.3° to 8.6°, and 80% of the direction error is within 11.5°, which has a significant effect on the accuracy of long-distance single-site lightning location.

After revising the ground wave arrival time delay and site-error, the error of single-station lightning positioning proposed in this paper is 8.4%–18.6%. The comparison between method in this study and other methods is shown in Table 2. The outstanding advantages of this paper have two parts. Firstly, the universality of the method, because it is easy to obtain the peak points of ground wave and skywave, so the method can be used to estimate lightning occurring at different distances from any station. The statistics of site-error are relatively simple, and a few days of data can be used to revise site-error, but more data revision will be better. Secondly, for lightning occurring far away from the station, the estimation error can still be controlled within a reasonable range.

By analyzing the waveform data, it is found that the data with large deviation values are mainly due to the fact that the sferics is affected by noise during propagation. And thus, when matching with the waveform bank, the wrong ground wave peak point or skywave peak point will be identified within the specified time window, which will make the time between the ground wave and the skywave becomes smaller or bigger, resulting in the deviation between the estimated distance and the actual distance becomes larger. At the same time, the noise will affect the ratio of the peak points of the skywave in the sferics waveform received by the two orthogonal magnetic antennas, which also leads to the deviation of the direction calculation, making the single-site lightning location results deviate greatly compared to multi-station lightning location results.

5 Conclusion

In this paper, we propose two methods to improve the positioning accuracy of single station and the method can be used in engineering practice. Single-site lightning location does not depend on high precision timing system and internet, and the operating cost is low. The relationship curve between the

ground wave arrival time delay and the propagation distance is obtained by numerical calculation method, which is used to revise the ground wave arrival time delay caused by the propagation effect. This is used to improve the accuracy of the distance estimation; the site error of a single station is revised by using multi-station lightning location data to improve the accuracy of direction calculation. The accuracy of the revised single-site lightning location data was evaluated using multi-station data. The main conclusions of this paper are as follows.

(1) Through the statistical analysis of the distance estimation deviation, it is determined that the applicable detection range of the single-site lightning location system mentioned in this paper is 200 km–2000 km. Moreover, when the ground wave arrival time delay has been revised, the accuracy of the distance estimation is much smaller than the not revised results.
(2) It is found that the azimuth deviation calculated using the peak value of the skywave is more concentrated. The revised average azimuth deviation decreased from 12.3° to 8.6° compared to the non-revised results.
(3) The results compared with the multi-station results, the error range of single-station lightning location is 8.4%–18.6%, and it also has a good detection performance for thunderstorms that occur at a long distance from the single-site. At the same time, with the increase of the detection distance, the detection accuracy can be maintained within a reasonable range.

It should be noted that if there are enough data, we can establish a waveform bank composed of actual sferics waveforms, which can effectively improve the accuracy of single-site lightning location. At the same time, these waveforms can also be used to study the relationship between the ground wave arrival time delay and the propagation distance under the real terrain, which can improve the accuracy of short-range lightning location, it can also further expand the range of lightning detection and improve the accuracy of lightning location.

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

Conceptualization, JiZ (JiZ) and QZ; Data curation, JiZ (JiZ), JuZ (JuZ), BD, and YW; Formal analysis, JiZ (JiZ), JL, and JG; Funding acquisition, QZ; investigation, JiZ (JiZ) and QZ; methodology, JiZ (JiZ) and JuZ (JuZ); software, JiZ (JiZ); JuZ (JuZ) and JL; supervision, QZ; visualization, JiZ (JiZ); writing—original draft, JiZ (JiZ); writing—review and editing, QZ. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the National Key R&D Program of China under Grant 2017YFC1501505, in part by the National Natural Science Foundation of China under Grant 41775006.

Acknowledgments

Thanks to the National Satellite Meteorological Center (NSMC) of the China Meteorological Administration for providing the FY4A-AGRI dataset. The author would like to thank all the people and departments involved in the construction of the network. The authors would also like to thank the reviewers for their helpful feedback, which significantly improved the manuscript.

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

Ammar, A., and Ghalila, H. (2020). Estimation of nighttime ionospheric D-region parameters using tweek atmospherics observed for the first time in the north african region. *Adv. Space Res.* 66, 2528–2536. doi:10.1016/j.asr.2020.08.025

Antunes de Sa, A. L., and Marshall, R. A. (2020). Lightning distance estimation using LF lightning Radio signals via analytical and machine-learned models. *IEEE Trans. Geosci. Remote Sens.* 58, 5892–5907. doi:10.1109/TGRS.2020.2972153

Chen, M., Lu, T., and Du, Y. (2015). An improved wave impedance approach for locating close lightning stroke from single station observation and its validation. *J. Atmos. Sol.-Terr. Phys.* 122, 1–8. doi:10.1016/j.jastp.2014.11.001

Chen, M., Lu, T., and Du, Y. (2013). Properties of "site error" of lightning direction-finder (DF) and its modeling. *Atmos. Res.* 129–130, 97–109. doi:10.1016/j.atmosres.2012.09.003

Dai, B., Li, J., Zhou, J., Zeng, Y., Hou, W., Zhang, J., et al. (2022). Application of a modified empirical wavelet Transform method in VLF/LF lightning electric field signals. *Remote Sens.* 14, 1308. doi:10.3390/rs14061308

Dennis, A. S., and Pierce, E. T. (1964). The return stroke of the lightning flash to earth as a source of VLF atmospherics. *J. Res. Natl. Bur. Stand. Sect. Radio Sci.* 68D, 777. doi:10.6028/jres.068D.075

DiGangi, E., Lapierre, J., Stock, M., Hoekzema, M., and Cunha, B. (2022). Analyzing lightning characteristics in central and southern south America. *Electr. Power Syst. Res.* 213, 108704. doi:10.1016/j.epsr.2022.108704

Gu, J., Zhang, Q., Li, J., Zhang, J., Zhou, J., Dai, B., et al. (2022). Effect of number and configuration of participating stations on lightning location outside the network. *Remote Sens.* 14, 4242. doi:10.3390/rs14174242

Han, F., and Cummer, S. A. (2010). Midlatitude nighttime D region ionosphere variability on hourly to monthly time scales: D region measurement. *J. Geophys. Res. Space Phys.* 115. doi:10.1029/2010JA015437

Han, F., and Cummer, S. A. (2010). Midlatitude nighttime D region ionosphere variability on hourly to monthly time scales: D region measurement. *J. Geophys. Res. Space Phys.* 115. doi:10.1029/2010JA015437

Heidler, F., Cvetic, J. M., and Stanic, B. V. (1999). Calculation of lightning current parameters. *IEEE Trans. Power Deliv.* 14, 399–404. doi:10.1109/61.754080

Herrman, B., Uman, M., Brantley, R., and Krider, E. (1976). Test of the principle of operation of a wideband magnetic direction finder for lightning return strokes. *J. Appl. Meteorol.* 15, 402–405. doi:10.1175/1520-0450(1976)015

Hill, D. A., and Wait, J. R. (1980). Ground wave attenuation function for a spherical earth with arbitrary surface impedance. *Radio Sci.* 15, 637–643. doi:10.1029/RS015i003p00637

Honma, N., Suzuki, F., Miyake, Y., Ishii, M., and Hidayat, S. (1998). Propagation effect on field waveforms in relation to time-of-arrival technique in lightning location. *J. Geophys. Res. Atmos.* 103, 14141–14145. doi:10.1029/97JD02625

Hou, W., Zhang, Q., Zhang, J., Wang, L., and Shen, Y. (2018). A new approximate method for lightning-radiated ELF/VLF ground wave propagation over intermediate ranges. *Int. J. Antennas Propag.* 2018, 1–10. doi:10.1155/2018/9353294

Hu, W., and Cummer, S. A. (2006). An FDTD model for low and high altitude lightning-generated EM fields. *IEEE Trans. Antennas Propag.* 54, 1513–1522. doi:10.1109/TAP.2006.874336

Koshak, W. J., Solakiewicz, R. J., Blakeslee, R. J., Goodman, S. J., Christian, H. J., Hall, J. M., et al. (2004). North Alabama lightning mapping array (LMA): VHF source retrieval algorithm and error analyses. *J. Atmos. Ocean. Technol.* 21, 543–558. doi:10.1175/15200426

Li, J., Dai, B., Zhou, J., Zhang, J., Zhang, Q., Yang, J., et al. (2022). Preliminary application of long-range lightning location network with equivalent propagation velocity in China. *Remote Sens.* 14, 560. doi:10.3390/rs14030560

Liu, B., Shi, L., Qiu, S., Liu, H., Dong, W., Li, Y., et al. (2020). Fine three-dimensional VHF lightning mapping using waveform cross-correlation TOA method. *Earth Space Sci.* 7. doi:10.1029/2019EA000832

Lu, T., Chen, M., Du, Y., and Qiu, Z. (2017). A statistical approach for site error correction in lightning location networks with DF/TOA technique and its application results. *Atmos. Res.* 184, 103–111. doi:10.1016/j.atmosres.2016.10.009

Mach, D. M., MacGorman, D. R., David Rust, W., and Arnold, R. T. (1986). Site errors and detection efficiency in a magnetic direction-finder network for locating lightning strikes to ground. *J. Atmos. Ocean. Technol.* 3, 67–74. doi:10.1175/1520-0426(1986)003

Mostajabi, A., Karami, H., Azadifar, M., Ghasemi, A., Rubinstein, M., and Rachidi, F. (2019). Single-sensor source localization using electromagnetic time reversal and deep transfer learning: Application to lightning. *Sci. Rep.* 9, 17372. doi:10.1038/s41598-019-53934-4

Nag, A., Murphy, M. J., Schulz, W., and Cummins, K. L. (2014). "Lightning locating systems: Characteristics and validation techniques," in Proceedings of the 2014 International Conference on Lightning Protection (ICLP), Shanghai, China, October 2014, 1070–1082.

Nagano, I., Yagitani, S., Ozaki, M., Nakamura, Y., and Miyamura, K. (2007). Estimation of lightning location from single station observations of sferics. *Electron. Commun. Jpn. Part Commun.* 90, 25–34. doi:10.1002/ecja.20329

Nucci, C. A. (1988). "On lightning return stroke models for LEMP calculations," in Proceedings of the 19th International Conference on Lightning Protection, Graz, Austria, April 1988.

Pohjola, H., and Mäkelä, A. (2013). The comparison of GLD360 and EUCLID lightning location systems in europe. *Atmos. Res.* 123, 117–128. doi:10.1016/j.atmosres.2012.10.019

Rachidi, F., Janischewskyj, W., Hussein, A. M., Nucci, C. A., Guerrieri, S., Kordi, B., et al. (2001). Current and electromagnetic field associated with lightning-return strokes to tall towers. *IEEE Trans. Electromagn. Compat.* 43, 356–367. doi:10.1109/15.942607

Rakov, V. A. (2013). Electromagnetic methods of lightning detection. *Surv. Geophys.* 34, 731–753. doi:10.1007/s10712-013-9251-1

Ramachandran, V., Prakash, J. N., Deo, A., and Kumar, S. (2007). Lightning stroke distance estimation from single station observation and validation with WWLLN data. *Ann. Geophys.* 25, 1509–1517. doi:10.5194/angeo-25-1509-2007

Shao, X.-M., and Jacobson, A. R. (2009). Model simulation of very low-frequency and low-frequency lightning signal propagation over intermediate ranges. *IEEE Trans. Electromagn. Compat.* 51, 519–525. doi:10.1109/TEMC.2009.2022171

Shi, D., Zheng, D., Zhang, Y., Zhang, Y., Huang, Z., Lu, W., et al. (2017). Low-frequency E-field detection array (LFEDA)—construction and preliminary results. *Sci. China Earth Sci.* 60, 1896–1908. doi:10.1007/s11430-016-9093-9

Somu, V. B., Rakov, V. A., Haddad, M. A., and Cummer, S. A. (2015). A study of changes in apparent ionospheric reflection height within individual lightning flashes. *J. Atmos. Sol.-Terr. Phys.* 136, 66–79. doi:10.1016/j.jastp.2015.09.007

Wait, J. R. (1974). Recent analytical investigations of electromagnetic ground wave propagation over inhomogeneous earth models. *Proc. IEEE* 62, 1061–1072. doi:10.1109/PROC.1974.9570

Wang, J., Ma, Q., Zhou, X., Xiao, F., Yuan, S., Chang, S., et al. (2020). Asia-pacific lightning location network (APLLN) and preliminary performance assessment. *Remote Sens.* 12, 1537. doi:10.3390/rs12101537

Wang, J., Xiao, F., Yuan, S., Song, J., Ma, Q., and Zhou, X. (2022). A novel method for ground-based VLF/LF single-site lightning location. *Measurement* 196, 111208. doi:10.1016/j.measurement.2022.111208

Wang, Y., Min, Y., Liu, Y., and Zhao, G. (2021). A new approach of 3D lightning location based on pearson correlation combined with empirical mode decomposition. *Remote Sens.* 13, 3883. doi:10.3390/rs13193883

Wu, T., Wang, D., and Takagi, N. (2018). Lightning mapping with an array of fast antennas. *Geophys. Res. Lett.* 45, 3698–3705. doi:10.1002/2018GL077628

Zhang, G., Wang, Y., Qie, X., Zhang, T., Zhao, Y., Li, Y., et al. (2010). Using lightning locating system based on time-of-arrival technique to study three-dimensional lightning discharge processes. *Sci. China Earth Sci.* 53, 591–602. doi:10.1007/s11430-009-0116-x

Zhang, H., Lu, G., Qie, X., Jiang, R., Fan, Y., Tian, Y., et al. (2016). Locating narrow bipolar events with single-station measurement of low-frequency magnetic fields. *J. Atmos. Sol.-Terr. Phys.* 143–144, 88–101. doi:10.1016/j.jastp.2016.03.009

Zhang, J., Zhou, J., Li, J., Gu, J., Zhang, Q., Dai, B., et al. (2022). Location accuracy improvement of long-range lightning detection network in China by compensating ground wave propagation delay. *Remote Sens.* 14, 3397. doi:10.3390/rs14143397

Zhou, X., Wang, J., Ma, Q., Huang, Q., and Xiao, F. (2021). A method for determining D region ionosphere reflection height from lightning skywaves. *J. Atmos. Sol.-Terr. Phys.* 221, 105692. doi:10.1016/j.jastp.2021.105692

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 17 April 2023
DOI 10.3389/fphy.2023.1185103

OPEN ACCESS

EDITED BY
Kaijie Xu,
University of Alberta, Canada

REVIEWED BY
Peng Nie,
Xidian University, China
Yihui Hu,
Xi'an University of Posts and Telecommunications, China
Jiazhong Zhou,
Huaqiao University, China

*CORRESPONDENCE
Qi Zhang,
zhangqi@ecut.edu.cn

SPECIALTY SECTION
This article was submitted to Optics and Photonics, a section of the journal Frontiers in Physics

RECEIVED 13 March 2023
ACCEPTED 31 March 2023
PUBLISHED 17 April 2023

CITATION
Zhang Q (2023), Robust predictability in discrete event systems under sensor attacks.
Front. Phys. 11:1185103.
doi: 10.3389/fphy.2023.1185103

Robust predictability in discrete event systems under sensor attacks

Qi Zhang*

School of Information Engineering, East China University of Technology, Nanchang, China

The problem of robust predictability against sensor attacks is investigated. The objective of a diagnoser is to predict the occurrence of a critical event of a discrete event system (DES) under partial observation. An attacker may rewrite the diagnoser observation by inserting fake events or erasing real events. Two novel structures, namely, real diagnoser and the fake diagnoser, are constructed based on the diagnoser of the system. We compute the hybrid diagnoser as the parallel composition of the real diagnoser and the fake diagnoser. The hybrid diagnoser can be used to verify if a critical event of the system is robustly predictable when an attacker tampers with the diagnoser observation.

KEYWORDS

discrete event system, automaton, predictability, diagnoser, sensor attack

1 Introduction

Suppose that a plant is modeled by a discrete event system (DES) under partial observation, *predictability* is a property that describes if a diagnoser can predict the occurrence of a critical event (either observable or unobservable) according to its observation of the system. As the system and the diagnoser are connected via a network, a malicious attacker may corrupt such a communication channel with the insertion of fake events and the deletion of real events that have happened in the system. Therefore, the problem of robust predictability against sensor attacks is addressed. It characterizes the ability of a diagnoser to predict the occurrence of a critical event, even if an attacker may tamper with its observation.

Genc and Lafortune [1] proposed the problem of predictability in the centralized case, and Kumar and Takai [2] considered this problem in the decentralized case. From this point, many studies have focused on this topic in different contexts and problem settings. Takai and Kumar [3, 4] considered the problem of failure prognosis with communication delays. In [5–7], the problem of predictability is studied in the context of stochastic DESs. Benmessahel et al. [8] investigated the problem of predictability in fuzzy DESs. Yin and Li [9] studied the problem of reliable decentralized fault predictability. They supposed that only partial local prognostic decisions are accessible to the coordinator. In [10], the authors showed how to use one prognoser to predict the occurrence of any failure for a set of models. Xiao and Liu [11] considered the problem of robust fault prognosis against loss of observations, where some observable events may become unobservable because of sensor failures. Finally, the problem of predictability is investigated in [12–14] in the framework of Petri nets.

The notion of diagnosability was first proposed in [15]. We assume that a DES contains an unobservable fault event. A fault event is said to be diagnosable if we can determine its occurrence within a limited delay. We point out that if the property of predictability is

stronger than that of diagnosability, i.e., if an event is predictable, then this event is also diagnosable.

The problem of robust codiagnosability against Denial-of-Service and deception attacks has been considered in [16]. The authors assume that an attacker can insert fake packages into the network that transmits the sensor readings such that delays and loss of observations may occur. They construct a new diagnoser to verify the property of robust codiagnosability. In [17], the problem of robust codiagnosability against sensor attacks under cost constraint is proposed. The considered attacks include symbol insertion, symbol erasure, and symbol replacement attacks. They assumed that each attack action consumes a certain amount of cost. They developed a strategy to verify the robust codiagnosability against an attacker with a bounded total cost.

Mainly inspired by [16, 17] that considered the problem of robust diagnosability in DESs subject to cyberattacks, we propose the problem of robust predictability in DESs subject to sensor attacks. To the best of the author's knowledge, this problem has not been considered in the framework of DESs. We finally mention that in [18], a structure named joint estimator is addressed to solve the problem of joint state estimation under attacks. This is a general structure that can be used to consider a set of problems in DESs subject to sensor attacks. In this work, we extend such a structure to solve the problem of robust predictability against sensor attacks.

In Section 2, the automata model and the notions of predictability and diagnoser are given. In Section 3, the problem considered in this study is presented. In Section 4, the real diagnoser is computed. It characterizes the real evolution of the diagnoser subject to sensor attacks. In Section 5, the fake diagnoser is constructed. It characterizes the fake evolution of the diagnoser subject to sensor attacks. In Section 6, the hybrid diagnoser is computed. It allows us to test if a critical event is robustly predictable. Section 7 summarizes the main results of this work, and the possible future work is also pointed out.

2 Preliminaries

Let E be an alphabet and L a language defined over E^*. The prefix closure of L is defined by $\bar{L} = \{\sigma \in E^* \mid (\exists \sigma' \in E^*)\ \sigma\sigma' \in L\}$. The post language of L after $\sigma \in L$ is defined as $L/\sigma = \{\sigma' \in E^* \mid \sigma\sigma' \in L\}$. A language L is live if for all $\sigma \in L$, there always exists $e \in E$ such that $\sigma e \in L$. The set of words in L that end with event f is defined by $\Psi(f, L) = \{\sigma f \in L \mid \sigma \in E^*, f \in E\}$.

A deterministic finite-state automaton (DFA), denoted by G, is a four tuple $G = \{X, E, \delta, x_0\}$, where X is a set of states; E is a finite set of events; δ: $X \times E \to X$ is the transition function and can be extended from the domain $X \times E$ to the domain $X \times E^*$, that is, $\delta(x, \varepsilon) := x$, and $\delta^*(x, \sigma e) := \delta(\delta^*(x, \sigma), e)$, where $e \in E$, $\sigma \in E^*$, and x_0 is the initial state. The generated language of G is defined by $L(G) = \{\sigma \in E^* \mid \delta^*(x, \sigma)$ is defined$\}$. The set of active events at state x of G is defined by $\Gamma_G(x) = \{e \in E \mid \delta(x, e)$ is defined$\}$.

A set of states $\{x_1, x_2, \ldots, x_n\} \subseteq X$ and a word $\sigma = e_1 e_2 \ldots e_n \in E^*$ form a cycle if $\delta(x_i, e_i) = x_{i+1}$, $i = 1, 2, \ldots, n-1$, and $\delta(x_n, e_n) = x_1$. The accessible part of G with respect to state x is defined as $Ac(G, x) = (X_{ac}, E, \delta_{ac}, x_0)$, where $X_{ac} = \{x' \in X \mid (\exists \sigma \in E^*)\ \delta^*(x, \sigma) = x'\}$, $\delta_{ac} = \delta|_{X_{ac} \times E \to X_{ac}}$.

Due to the lack of observability in the system, E is divided into the set of observable events E_o and the set of unobservable events E_{uo}. The natural projection on E_o is denoted as P: $E^* \to E_o^*$. Considering a word $\sigma \in E^*$, $P(\sigma)$ simply removes the unobservable events from σ, that is, $P(\varepsilon) := \varepsilon$ and $P(\sigma e) := P(\sigma)e$ if $e \in E_o$ and $P(\sigma e) := P(\sigma)$ if $e \in E \setminus E_o$.

Definition 1. [1] Consider a prefix-closed and live language L on alphabet E. An event f is said to be predictable with respect to P if $(\exists n \in \mathbb{N})\ \forall \sigma \in \Psi(f, L), \exists t \in \bar{\sigma}$ such that $f \notin t \wedge \mathcal{P}$, where condition $\mathcal{P}$:

$\forall u \in L$ such that $P(u) = P(t)$, $f \notin u$. $\forall v \in L/u$ such that $|v| \geq n \Rightarrow f \in v$.

In plain words, an event f is predictable if it holds that once the observation $P(t)$ is produced, f will necessarily occur within n steps, where t is a normal prefix of a word σ that ends with f.

Definition 2. [1] Let $G = (X, E, \delta, x_0)$ be a plant and f an event that needs to be predicted. The diagnoser is a DFA, denoted as $D_g = (B, E_o, \delta_d, b_0)$, where

- $B \subseteq 2^{X \times \{N,F\}}$, for example, $b = \{(x_1, l_1), \ldots, (x_n, l_n)\}$, and $x_1, x_2 \ldots, x_n \in X$;
- δ_d: $B \times E_o \to B$, for example, if $\exists e \in E_o$ such that $\delta_d(b, e) = b'$, where $b = \{(x_1, l_1), \ldots, (x_m, l_m)\}$ and $b' = \{(x_1', l_1'), \ldots, (x_n', l_n')\}$, then $\exists i \in \{1, \ldots, m\}$, $\exists j \in \{1, \ldots, n\}$, and $\exists \sigma = te$: $t \in E_{uo}^*$ such that $\delta^*(x_i, \sigma) = x_j'$, where

$$l_j' = \begin{cases} N & if \quad l_i = N \wedge f \notin \sigma, \\ F & if \quad l_i = F \vee f \in \sigma. \end{cases}$$

If a state of the diagnoser is labeled N, it indicates that event f has not happened when the current state is reached. If a state of the diagnoser is labeled F, it implies that event f has happened when the current state is reached. By convention, the unobservable reach is not included in a diagnoser state.

Definition 3. [1] In the diagnoser $D_g = (B, E_o, \delta_d, b_0)$,

- We define $B_n = \{b = \{(x_1, l_1), \ldots, (x_n, l_n)\} \in B \mid \forall\, l_i \in \{l_1, \ldots, l_n\}, l_i = N\}$ as the set of normal states of D_g.
- We define $B_c = \{b = \{(x_1, l_1), \ldots, (x_n, l_n)\} \in B \mid \forall\, l_i \in \{l_1, \ldots, l_n\}, l_i = F\}$ as the set of certain states of D_g.
- We define $B_{uc} = \{b = \{(x_1, l_1), \ldots, (x_n, l_n)\} \in B \mid \exists\, l_i, l_j \in \{l_1, \ldots, l_n\}, l_i = N, l_j = F\}$ as the set of uncertain states of D_g.
- We denote by B_d the set of normal states with an instantaneous continuator, which is not normal, that is, $B_d = \{b \in B_n \mid (\exists e \in E_o)\ \delta_d(b, e) \notin B_n\}$.

In other words, a state $b \in B$ is normal if all the labels within it are N; a state $b \in B$ is certain if all the labels within it are F; and a state $b \in B$ is uncertain if there exist labels N and F within it.

Theorem 4. [1] Let G be a plant and $D_g = (B, E_o, \delta_d, b_0)$ its diagnoser. An event f is predictable if and only if for all $b_d \in B_d$, in the accessible part of the diagnoser $Ac(D_g, b)$, all cycles are cycles of certain states.

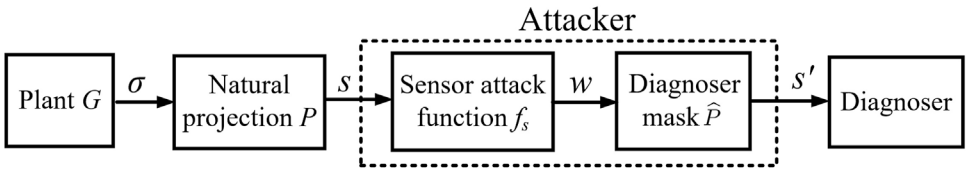


FIGURE 1
System under attack.

3 Problem formulation

Let $G = (X, E, \delta, x_0)$ be a plant modeled by a DFA. As shown in Figure 1, if the word $\sigma \in E^\star$ is generated by G, the observation $s = P(\sigma)$ may be corrupted by an attacker. Then, the diagnoser predicts the occurrence of a critical event in accordance with the corrupted observation s'. It should be noted that the internal structure of the attacker within the dotted lines will be discussed later.

Suppose that an attacker can only tamper with a subset of events of G, we call this subset the set of compromised events E_{com}. We divide E_{com} into two subsets, that is, $E_{com} = E_{ins} \cup E_{era}$, where E_{ins} is the set of events that may be inserted into the diagnoser observation, and E_{era} is the set of events that may be deleted from the diagnoser observation. The sets E_{ins} and E_{era} may contain common events.

To make a distinction between the attacker's action from the original behavior of G, we define two new sets of events. We denote by E_+ the set of inserted events, defined as $E_+ = \{e_+ \mid e \in E_{ins}\}$ [19]. We denote by E_- the set of erased events, defined as $E_- = \{e_- \mid e \in E_{era}\}$ [19]. If $e_+ \in E_+$ happens, it indicates that an attacker inserts the fake symbol $e \in E_{ins}$ into the diagnoser observation. If $e_- \in E_-$ happens, it indicates that an attacker erases the real symbol $e \in E_{era}$ from the diagnoser observation. Finally, we denote by E_a the attack alphabet, defined as $E_a = E_o \cup E_+ \cup E_-$. We point out that the three subsets E_o, E_+, and E_- are disjoint.

Definition 5. Let G be a plant and $E_{com} = E_{ins} \cup E_{era}$ the set of compromised events. An attacker is defined by a sensor attack function f_s: $P[L(G)] \to E_a^\star$:

(1) $f_s(\varepsilon) \in E_+^\star$,
(2) $\forall se \in P[L(G)]$:

$$\begin{cases} f_s(se) \in f_s(s)\{e_-, e\}E_+^\star & \text{if} \quad e \in E_{era}, \\ f_s(se) \in f_s(s)eE_+^\star & \text{if} \quad e \in E_o \backslash E_{era}. \end{cases} \tag{1}$$

Condition (1) means that a word in $E_+^\star$ can be inserted by the attacker before an observable event occurs in G. Condition (2) means that when an event that can be erased by the attacker occurs, the attacker either erases it or not; then, it inserts any word defined over $E_+^\star$. Finally, when an event that cannot be erased by the attacker happens, the attacker can insert a word defined over $E_+^\star$ after it.

Let G be a plant. We denote by $L(f_s, G)$ the attack language, defined by $L(f_s, G) = f_s(P[L(G)]) \subseteq E_a^\star$. We call $w \in L(f_s, G)$ an attack word. We denote by $\mathcal{F}_s$ the set of sensor attack functions. We denote by $L(\mathcal{F}_s, G)$ the union of all the attack languages, defined by $L(\mathcal{F}_s, G) = \bigcup_{f_s \in \mathcal{F}_s} f_s(P[L(G)])$.

Definition 6. The real mask $\tilde{P}$: $E_a^\star \to E_o^\star$ is defined as follows:

$$\tilde{P}(\varepsilon) = \varepsilon, \ \tilde{P}(we') = \begin{cases} \tilde{P}(w)e \ \text{if} \ e' = e \in E_o \vee e' = e_- \in E_-, \\ \tilde{P}(w) \ \ \text{if} \ e' = e_+ \in E_+. \end{cases} \tag{2}$$

In plain words, the real mask transforms events in E_a into real events that have happened in the system. As e_- means an erased event that has happened in the system, e_- is transformed into the corresponding event $e \in E_o$. e_+ is neglected because it is a fake event.

Definition 7. The diagnoser mask $\hat{P}$: $E_a^\star \to E_o^\star$ is defined as follows:

$$\hat{P}(\varepsilon) = \varepsilon, \ \hat{P}(we') = \begin{cases} \hat{P}(w)e \ \text{if} \ e' = e \in E_o \vee e' = e_+ \in E_+, \\ \hat{P}(w) \ \ \text{if} \ e' = e_- \in E_-. \end{cases} \tag{3}$$

In simple words, the diagnoser mask characterizes how the diagnoser observes events in E_a. Namely, the diagnoser cannot distinguish the real event $e \in E_o$ from the inserted event $e_+ \in E_+$, and it cannot observe erased events in E_-.

As shown in Figure 1 within the dotted lines, the observation $s \in E_o$ is corrupted into the attack word $w \in E_a^\star$ by the sensor attack function f_s; then, w is transformed into the corrupt observation $s' = \hat{P}(w)$. Therefore, the diagnoser actually observes s'.

In this study, let G be a plant. The following two assumptions are made:

1) The generated language $L(G)$ is live.
2) In G, there does not exist a cycle that consists of unobservable events only.

Assumption 1) is made for the sake of simplicity. Assumption 2) guarantees that plant G does not generate unobservable words with infinite length.

Definition 8. Let G be a plant that satisfies Assumption 1) and Assumption 2). An event f is robustly predictable with respect to P if $(\exists n \in \mathbb{N})\ \forall \sigma \in \Psi(f, L),\ \exists t \in \bar{\sigma}$ such that $f \notin t \wedge \mathcal{P}_r$, where condition $\mathcal{P}_r$:

$\forall w \in L(\mathcal{F}_s, G)$ such that $\tilde{P}(w) = P(t) \vee \hat{P}(w) = P(t)$. $\forall u \in L(G)$ such that $P(u) = \tilde{P}(w) \vee P(u) = \hat{P}(w)$, $f \notin u$. $\forall v \in L(G)/u$ such that $|v| \geq n \Rightarrow f \in v$.

In Definition 8, let t be a normal prefix of a word σ that ends with f. We use t to find all the attack words $w \in E_a^\star$ such that $\tilde{P}(w) = P(t) \vee \hat{P}(w) = P(t)$. Then, we use these attack words w to find all the word $u \in E^\star$ such that $P(u) = \tilde{P}(w) \vee P(u) = \hat{P}(w)$.

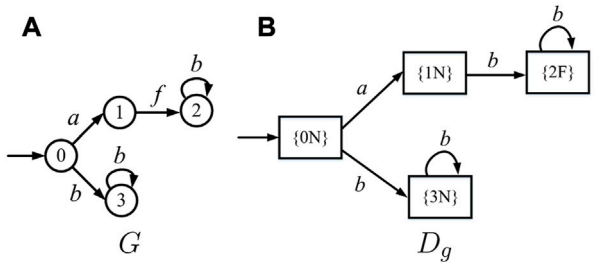


FIGURE 2
(A) Plant G and **(B)** diagnoser D_g in Example 10.

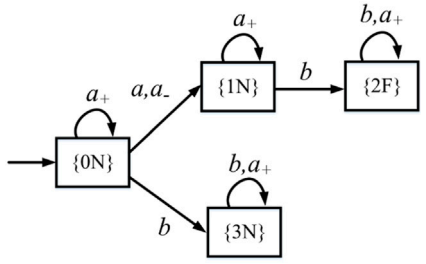


FIGURE 3
Real diagnoser D_r in Example 10.

An event f is robustly predictable if it holds that once the observation $P(u)$ is produced, then f will necessarily occur within n steps.

We point out that, for each attack word w, we distinguish the observations $P(u) = \tilde{P}(w)$ and $P(u) = \hat{P}(w)$ because the attacker can make these two observations look alike for the diagnoser.

4 Real diagnoser

The real diagnoser D_r describes the real evolution of the diagnoser in accordance with the attack alphabet E_a. Namely, the real diagnoser changes its states the same way in terms of $e \in E_{era}$ and the corresponding events e_-; the real diagnoser does not change its states when the fake event $e_+ \in E_+$ happens.

Definition 9. Let $G = (X, E, \delta, x_0)$ be a plant and $D_g = (B, E_o, \delta_d, b_0)$ the diagnoser. The real diagnoser is a DFA $D_r = (B, E_a, \delta_r, b_0)$, and its transition function δ_r satisfies the following:

$$\begin{cases} for\ all\ b \in B,\ for\ all\ e \in E_o: & \delta_r(b,e) := \delta_d(b,e), \\ for\ all\ b \in B,\ for\ all\ e \in E_{era}: & \delta_r(b,e_-) := \delta_r(b,e), \\ for\ all\ b \in B,\ for\ all\ e \in E_{ins}: & \delta_r(b,e_+) := b. \end{cases} \tag{4}$$

The construction of the real diagnoser can be explained as follows: first, we set the transition function of the real diagnoser D_r equal to the transition function of the diagnoser D_g. Then, each time there is a transition labeled $e \in E_{era}$, we add a transition labeled e_-. Finally, for each event in E_{ins}, for each state of D_r, we add a self-loop labeled e_+.

We point out that the real diagnoser D_r is similar to the attacker observer constructed by Algorithm 1 in [18]. Although the input of Algorithm 1 is the observer of G, here we replace it with the diagnoser of G.

Example 10. As sketched in Figure 2A, let G be the plant, $E_o = \{a, b\}$, and $E_{uo} = \{f\}$. Assume that f is the event that needs to be predicted. The diagnoser $D_g = (B, E_o, \delta_d, b_0)$ is sketched in Figure 2B.

Let $E_{ins} = E_{era} = \{a\}$. The real diagnoser is shown in Figure 3. We add a transition $\delta_r(\{0N\}, a_-) = \{1N\}$ in D_r because there exists a transition $\delta_d(\{0N\}, a) = \{1N\}$ such that $e \in E_{era}$ in D_g. Self-loops labeled a_+ are added at all the states of D_r because $a \in E_{ins}$.

Proposition 11. Let G be the plant, $D_g = (B, E_a, \delta, b_0)$ its diagnoser, and $D_r = (B, E_a, \delta_r, b_0)$ the real diagnoser.

(i) $L(D_r) = L(\mathcal{F}_s, G)$;

(ii) $\forall s\ L(D_g)$, $\forall f_s \in \mathcal{F}_s$ with $w = f_s(s) \in E_a^*$: $\delta_r^*(b_0, w) = \delta_d^*(b_0, s)$.

Proof. The proof is neglected because it is the same as the proof of Proposition 1 in [18]. In simple words, item 1) means that the real diagnoser generates the union of all the attack languages. Item 2) indicates that the state arrived in D_r by implementing $w = f_s(s) \in E_a^*$ equal to the state arrived in D_g by implementing $s \in E_o^*$.

5 Fake diagnoser

The fake diagnoser D_f describes the fake evolution of the diagnoser in accordance with the attack alphabet E_a. Namely, the fake diagnoser changes its states the same way in terms of $e \in E_{ins}$ and the corresponding events e_+ because it cannot distinguish the real event of the plant e from the fake event e_+. The fake diagnoser does not change its states in case of the occurrence of $e_- \in E_-$ because it cannot observe the erased event e_-. We add a new state $b_\varnothing$ in D_r. The diagnoser knows that the plant is under attack when this state is reached.

Definition 12. Let $G = (X, E, \delta, x_0)$ be a plant and $D_g = (B, E_o, \delta_d, b_0)$ the diagnoser. The fake diagnoser is a DFA $D_f = (B_f, E_a, \delta_f, b_0)$ such that $B_f = B \cup b_\varnothing$, and its transition function δ_f satisfies the following:

$$\begin{cases} for\ all\ b \in B,\ for\ all\ e \in E_o: & \delta_f(b,e) := \delta_d(b,e), \\ for\ all\ b \in B,\ for\ all\ e \in E_{ins}: & \delta_f(b,e_+) := \delta_f(b,e), \\ for\ all\ b \in B,\ for\ all\ e \in E_{era}: & \delta_f(b,e_-) := b, \\ for\ all\ b \in B,\ for\ all\ e \in E_a: & if\ \delta_f(b,e)\ is\ undefined,\ then\ \delta_f(b,e) := b_\varnothing. \end{cases} \tag{5}$$

The construction of the fake diagnoser can be explained as follows: first, we set the transition function of D_f equal to the transition function of the diagnoser D_g. Then, each time there is a transition labeled $e \in E_{ins}$, we add a transition labeled $e_+ \in E_+$. Self-loop labeled events in E_- are added at all the states of D_f. Finally, for each event in E_a and each state in B, we set $\delta_f(b, e_a) = b_\varnothing$ for all the undefined transitions. Note that state $b_\varnothing$ has no input and output arcs.

We point out that the fake diagnoser D_f is similar to the operator observer computed by Algorithm 2 of [18]. Although the input of Algorithm 2 is the observer of G, here we replace it with the diagnoser of G.

Example 13. Recall plant G with its diagnoser D_g in Example 10. Suppose that $E_{ins} = E_{era} = \{a\}$. Figure 4 shows the fake diagnoser.

First, we add a transition $\delta_r(\{0N\}, a_+) = \{1N\}$ in D_f as there is a transition $\delta_d(\{0N\}, a) = \{1N\}$ such that $e \in E_{ins}$ in D_g. Then, for all the states of D_f, self-loops labeled a_- are added because $a \in E_{era}$. Finally, all the undefined transitions lead to the state $b_\varnothing$.

The following definitions are given to formalize the generated language of the fake diagnoser D_f.

Definition 14. Consider a plant G with the fake diagnoser D_f.

- A sensor attack function f_s is stealthy if $\hat{P}[L(f_s, G)] \subseteq P[L(G)]$.
- The set of stealthy words is defined as $W_s = \{w \in E_a^* \mid \hat{P}(w) \in P[L(G)]\}$.
- The set of exposing words is defined as $W_e = \{we_a \in E_a^* \mid w \in W_s, e_a \in E_a, we_a \notin W_s\}$.

According to Definition 14, f_s is stealthy if the attack words in $L(f_s, G) \subseteq E_a^*$ can be transformed into words in $P[L(G)] \subseteq E_o^*$ via the diagnoser mask $\hat{P}$; that is, the diagnoser cannot discover the presence of an attacker. Set W_s includes all the words that keep the attacker stealthy. Each word in W_e is the concatenation of a stealthy word and an event in E_a, and the resulting word is no more stealthy.

Proposition 15. Let G be the plant, $D_g = (B, E_o, \delta_d, b_0)$ the diagnoser, and $D_f = (B, E_a, \delta_f, b_0)$ the fake diagnoser.

(i) $L(D_f) = W_s \cup W_e$;
(ii) $\forall w \in L(D_f)$: if $w \in W_s$, then $\delta_f^*(b_0, w) = \delta_d^*(b_0, \hat{P}(w))$; if $w \in W_e$, then $\delta_f^*(b_0, w) = b_\varnothing$.

Proof. The proof is ignored because it is the same as the proof of Proposition 2 in [18]. In plain words, item (i) implies that the language of the fake diagnoser equals the union of W_s and W_e. Item (ii) means that the state arrived in D_f by implementing $w \in E_a^*$ equal to the state arrived in D_g by implementing $\hat{P}(w) \in E_o^*$, and all the exposing words lead to state $b_\varnothing$.

6 Hybrid diagnoser

The notion of the hybrid diagnoser is given on the basis of the real diagnoser and fake diagnoser.

Definition 16. Let $G = (X, E, \delta, x_0)$ be a plant, $D_r = (B, E_a, \delta_r, b_0)$ the real diagnoser, and $D_f = (B_f, E_a, \delta_f, b_0)$ the fake diagnoser. The hybrid diagnoser $D_h = (R, E_a, \delta_h, r_0)$ is defined as the parallel composition of D_r and D_h, that is, $D_h = D_r \| D_f$, where

- $R = (b, b_f) \subseteq 2^{X\times\{N,F\}} \times 2^{X\times\{N,F\}}$;
- $\delta_h[(b, b_f), e] = [\delta_r(b, e), \delta_f(b_f, e)]$ if $e \in \Gamma_{D_r}(b) \cap \Gamma_{D_f}(b_f)$, where $\Gamma_{D_r}(b)$ $(\Gamma_{D_f}(b_f))$ denotes the set of active events at state b (b_f) of D_r (D_f);
- the initial state is $r_0 = (b_0, b_0)$.

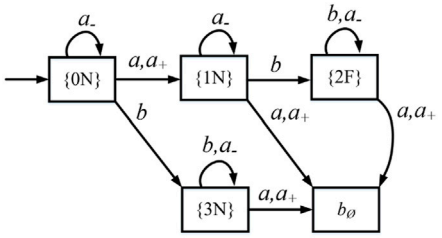


FIGURE 4
Fake diagnoser D_f in Example 13.

Now, we investigate the complexity of building the hybrid diagnoser D_h. Let $G = (X, E, \delta, x_0)$ be a plant. Its diagnoser D_g is built in $2^{|X|}$ steps. In accordance with Definition 9, the real diagnoser D_r contains at most $2^{|X|}$ states. In accordance with Definition 12, the fake diagnoser D_f contains at most $2^{|X|} + 1$ states. As $D_h = D_r \| D_f$, the computational complexity to build D_h is $O(2^{|X|} \cdot 2^{|X|})$.

Example 17. Recall plant G in Example 10. The hybrid diagnoser $D_h = D_r \| D_f$ is sketched in Figure 5, where D_r (D_f) is sketched in Figure 3 (Figure 4).

Definition 18. Let G be the plant, and $D_h = (R, E_a, \delta_h, r_0)$ be the hybrid diagnoser:

- We define $R_n = \{r = (b, b_f) \in R \mid b = \{(x_1, l_1), \ldots, (x_m, l_m)\}, b_f = \{(x_1', l_1'), \ldots, (x_n', l_n')\}$ such that $\forall\, l_i \in \{l_1, \ldots, l_m\}, \forall\, l_j' \in \{l_1', \ldots, l_n'\}, l_i = N, l_j' = N\}$ the set of normal states of D_h.
- We define $R_c = \{r = (b, b_f) \in R \mid b = \{(x_1, l_1), \ldots, (x_m, l_m)\}, b_f = \{(x_1', l_1'), \ldots, (x_n', l_n')\}$ such that $\forall\, l_i \in \{l_1, \ldots, l_m\}, \forall\, l_j' \in \{l_1', \ldots, l_n'\}, l_i = F, l_j' = F\}$ the set of certain states of D_h.
- We define $R_{uc} = \{r = (b, b_f) \in R \mid b = \{(x_1, l_1), \ldots, (x_m, l_m)\}, b_f = \{(x_1', l_1'), \ldots, (x_n', l_n')\}$ such that $\exists\, l_i \in \{l_1, \ldots, l_m\}, \exists l_j' \in \{l_1', \ldots, l_n'\}, l_i = N$ $(resp.,\ F), l_j' = F$ $(resp.,\ N)\}$ the set of uncertain states of D_h.
- We denote by R_d the set of normal states with an instantaneous continuator, which is not normal, that is, $R_d = \{r \in R_n \mid (\exists e_a \in E_a)\ \delta_h(r, e_a) \notin R_n\}$.

We point out that Definition 18, defined in hybrid diagnoser D_h, is the counterpart of Definition 3, defined in the diagnoser D_g.

Theorem 19. Let G be a plant, $D_g = (B, E_o, \delta_d, b_0)$ the diagnoser, and $D_h = (R, E_a, \delta_h, r_0)$ the hybrid diagnoser.

(a) $L(D_h) = L(\mathcal{F}_s, G) \cap (W_s \cup W_e)$;
(b) $\forall s \in P[L(G)], \forall f_s \in \mathcal{F}_s$ with $w = f_s(s) \in E_a^*$;
  (i) If $w \in W_s$, then $\delta_h^*(r_0, w) = (b, b_f) \Leftrightarrow \delta_d^*(b_0, s) = b,\ \delta_d^*[b_0, \hat{P}(w)] = b_f$;
  (ii) If $w \in W_e$, then $\delta_h^*(r_0, w) = (b, b_\varnothing) \Leftrightarrow \delta_d^*(b_0, s) = b,$ $\delta_d^*[b_0, \hat{P}(w)]$ is undefined.

Proof. The proof is neglected because it is the same as the proof of Theorem 1 in [18]. In other words, item (a) implies that the language

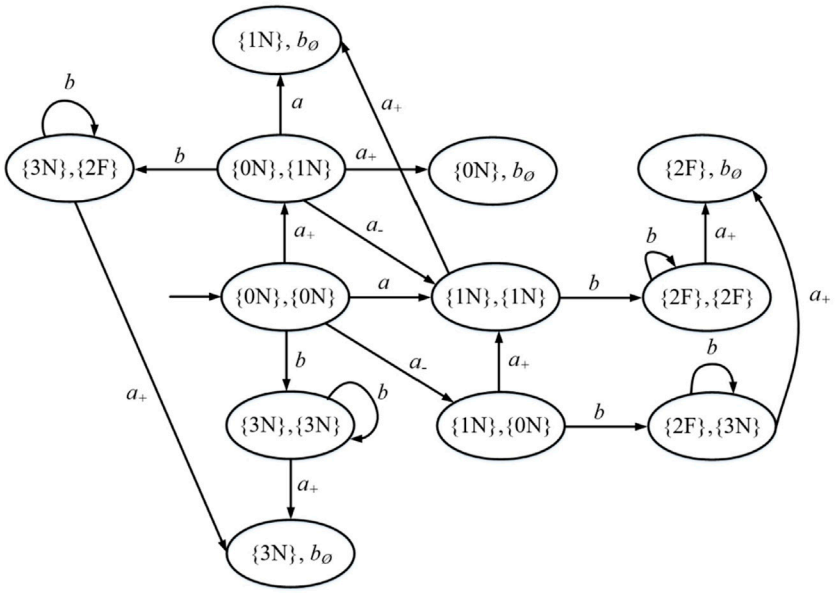


FIGURE 5
Hybrid diagnoser D_h in Example 17.

of the hybrid diagnoser D_h equals the intersection of the language of the real diagnoser and the language of the fake diagnoser.

Item (b) means that (i) if $w \in W_s$ and the state (b, b_f) is arrived in D_h by implementing $w = f_s(s)$, then the first element of this state equals the state arrived in the diagnoser D_g by implementing $s \in E_o^*$. The second element of this state equals the state arrived in D_g by implementing $\hat{P}(w)$. (ii) If $w \in W_e$, then $\delta_d^*(b_0, \hat{P}(w))$ is undefined.

Proposition 20. Let G be a plant and $D_h = (R, E_a, \delta_h, r_0)$ the hybrid diagnoser. In D_h, we suppose that a set of states $\{r_1, r_2, \ldots, r_n\} \subseteq R$ and a word $w = e_{a1}e_{a2}\ldots e_{an} \in E_a^*$ form a cycle. If $\exists r_i \in R_c$, then $\forall r_j \in R_c$, where $i, j \in \{1, 2, \ldots, n\}$ and R_c are the set of certain states.

Proof. Proposition 20 means that in a cycle of D_h, if a certain state exists, then all the other states in this cycle are certain. The proof follows from the fact that the label F propagates; once a state is labeled as a certain state, all the states that are reachable from this state are also certain.

Proposition 21. Let G be a plant, $D_g = (B, E_o, \delta_d, b_0)$ the diagnoser, and $D_h = (R, E_a, \delta_h, r_0)$ the hybrid diagnoser. In D_h, if a set of states $\{(b_1, b_{f1}), (b_2, b_{f2}), \ldots, (b_n, b_{fn})\} \subseteq R$ and a word $w = e_{a1}e_{a2}\ldots e_{an} \in E_a^*$ form a cycle, where $\forall i \in \{1, 2, \ldots, n\}$, $(b_i, b_{fi}) \in \{R_n \cup R_{uc}\}$. Then, in G, there exists a set of states $\{x_1, x_2, \ldots, x_n\} \subseteq X$ and a word $\sigma = e_1e_2\ldots e_n \in E^*$ forming a cycle such that $\forall i \in \{1, 2, \ldots, n\}$, $(x_i, l_i) \in b_i$, $l_i = N$, $w = f_s[P(\sigma)]$ or $\forall i \in \{1, 2, \ldots, n\}$, $(x_i, l_i) \in b_{fi}$, $l_i = N$, $P(\sigma) = \hat{P}(w)$, where f_s is the sensor attack function, and $\hat{P}$ is the diagnoser mask.

Proof. Assume that, in the hybrid diagnoser D_h, a set of states $\{(b_1, b_{f1}), (b_2, b_{f2}), \ldots, (b_n, b_{fn})\} \subseteq R$ and a word $w = e_{a1}e_{a2}\ldots e_{an} \in E_a^*$ form a cycle, where $\forall i \in \{1, 2, \ldots, n\}$, $(b_i, b_{fi}) \in \{R_n \cup R_{uc}\}$.

As $D_h = D_r \| D_f$, a set of states $\{b_1, , b_2, \ldots, b_n\} \subseteq B$ and the word $w = e_{a1}e_{a2}\ldots e_{an} \in E_a^*$ form a cycle in the real diagnoser D_r, and a set of states $\{b_{f1}, , b_{f2}, \ldots, b_{fn}\} \subseteq B_f$ and the word $w = e_{a1}e_{a2}\ldots e_{an} \in E_a^*$ form a cycle in the fake diagnoser D_f.

In accordance with Theorem 19, if $w \in W_s$, then $\delta_h^*(r_0, w) = (b, b_f) \Leftrightarrow \delta_d^*(b_0, s) = b$, $\delta_d^*[b_0, \hat{P}(w)] = b_f$, where $w = f_s(s)$, $s = P(\sigma) \in E_o^*$, and $\sigma = e_1e_2\ldots e_n \in E^*$. As $\forall i \in \{1, 2, \ldots, n\}$, $(b_i, b_{fi}) \in \{R_n \cup R_{uc}\}$, we distinguish two cases: 1) If $\forall i \in \{1, 2, \ldots, n\}$, $(x_i, l_i) \in b_i$, $l_i = N$, then in G, a set of states $\{x_1, , x_2, \ldots, x_n\} \subseteq X$ and a word $\sigma = e_1e_2\ldots e_n \in E^*$ form a cycle, where $w = f_s[P(\sigma)]$. 2) If $\forall i \in \{1, 2, \ldots, n\}$, $(x_i, l_i) \in b_{fi}$ and $l_i = N$, then in G, a set of states $\{x_1, , x_2, \ldots, x_n\} \subseteq X$ and a word $\sigma = e_1e_2\ldots e_n \in E^*$ form a cycle, where $P(\sigma) = \hat{P}(w)$.

Note that as state $b_\varnothing$ has no output arcs in the fake diagnoser D_f, then in D_h, the cycle does not contain the state whose second element is $b_\varnothing$. Therefore, the case of $w \in W_e$ is not considered when we use the results of Theorem 19. For the same reason, we exclude this case in the proof of Theorem 22.

Theorem 22. Let $G = (X, E, \delta, x_0)$ be a plant and $D_h = (R, E_a, \delta_h, r_0)$ the hybrid diagnoser. An event f is robustly predictable if and only if, for all $r_d \in R_d$, in the accessible part of the hybrid diagnoser $Ac(D_h, r_d)$, all cycles are cycles of certain states.

Proof. (If) Assume that for all $r_d \in R_d$, in $Ac(D_h, r_d)$, all cycles are cycles of certain states. Consider a word $\sigma \in \Psi(f, L(G))$ such that $\delta^*(x_0, \sigma) = x$. Let $\sigma_{uo}e_o \in L/\sigma$ such that $e_o \in E_o$ and $\delta^*(x, \sigma_{uo}e_o) = x'$.

Consider a word w such that $\tilde{P}(w) = P(\sigma)$ or $\hat{P}(w) = P(\sigma)$. Let $\delta_h^*(r_0, w) = r = (b, f_f)$ and $\delta_h(r, e_o) = r' = (b', b_f')$. According to

Theorem 19, $\delta_h^\star(r_0, we_o) = (b', b_f') = r' \Leftrightarrow \delta_d^\natural(b_0, s) = b'$, $\delta_d^\natural[b_0, \hat{P}(we_o)] = b_f'$. We consider the following two cases:

a) If $\tilde{P}(w) = P(\sigma)$, then $s = \tilde{P}(we_o) = P(\sigma e_o)$. It can be concluded that there exists $(x, l) \in b'$ such that $l = F$.
b) If $\hat{P}(w) = P(\sigma)$, then $\hat{P}(we_o) = P(\sigma e_o)$. It can be concluded that there exists $(x, l) \in b_f'$ such that $l = F$.

In any case, we can conclude that $r' \in R_{uc} \cup R_c$. As $\delta_h(r, e_o) = r'$, the following two cases are possible:

1) If $r \in R_n$, it means that $r \in R_d$ because $\delta_h(r, e_o) = r' \in \{R_{uc} \cup R_c\}$. Let $\sigma = tf$, where $t \in E^\star$. $\forall u \in L(G)$ such that $P(u) = \tilde{P}(w)$ or $P(u) = \hat{P}(w)$. As $\forall r_d \in R_d$, in $Ac(D_h, r_d)$, all cycles are cycles of certain states; then $\forall v \in L(G)/u$, $|v| \geq n$, and v contains f.
2) If $r \in R_{uc} \cup R_c$, then we can always find a state $r'' \in R_d$ such that state r is reachable from state r''. As a result, the proof for case 2) is reduced to the proof for case 1) by replacing r with r''.

(Only if) Assume that event f is robustly predictable, and there exists $r_d \in R_d$ such that $Ac(D_h, r_d)$ has a cycle that contains a state that is uncertain.

According to Proposition 20, in $Ac(D_h, r_d)$, as there exists a state that is uncertain in the cycle, then none of the states is certain in this cycle. In accordance with Proposition 21, as there exists a cycle where all the states are uncertain in $Ac(D_h, r_d)$, there exists a cycle where all the states are labeled N in plant G.

Suppose that, in D_h, $\delta_h^\star(r_0, w) = r_d = (b, b_f) \in R_d$. By Theorem 19, $\delta_h^\star(r_0, w) = (b, b_f) \Leftrightarrow \delta_d^\natural(b_0, s) = b$, $\delta_d^\natural[b_0, \hat{P}(w)] = b_f$. As $r_d \in R_d$, then there exists a word $\sigma \in \Psi(f, L(G))$ such that $\sigma = tf$, $t \in E^\star$, $\tilde{P}(w) = P(t)$ or $\hat{P}(w) = P(t)$. Let $r_1 = (b, b_f) \in R$ be a state of the cycle of $Ac(D_h, r_d)$ such that $\delta_h^\star(r_d, w') = r_1$. As $\delta_h^\star(r_0, w) = r_d$, then $\delta_h^\star(r_0, ww') = r_1$. Let x be a state of the cycle of G such that $\delta^\star(x_0, uv) = x$, and $\delta^\star(x, (e_1e_2 \ldots e_n)^m) = x$, where $u \in L(G)$, $v \in L(G)/u$ such that $P(u) = \tilde{P}(w)$ or $P(u) = \hat{P}(w)$. Then, $\delta^\star(x_0, uv(e_1e_2 \ldots e_n)^m) = x$. Because x is labeled by N in $Ac(D_h, r_d)$, then we can always find a word $v(e_1e_2 \ldots e_n)^m$ that does not contain f, and its length is greater than any $n \in \mathbb{N}$. As a result, the robustly predictable condition is violated, leading to a contradiction.

Example 23. Recall plant G in Example 10, where $E_o = \{a, b\}$ and $E_{uo} = \{f\}$. Assume that event f needs to be predicted. Let $E_{ins} = \{a\}$ and $E_{era} = \{a\}$.

In the diagnoser D_g in Figure 2B, state $\{1N\} \in B_d$. As $Ac(D, \{1N\})$ only contains one cycle (self-loop) labeled b at state $\{2F\}$, that is a certain state, according to Theorem 4, event f is predictable when no attack occurs.

In the hybrid diagnoser D_h visualized in Figure 5, states $(\{0N\}, \{1N\})$, $(\{1N\}, \{0N\})$, $(\{1N\}, \{1N\}) \in R_d$. As $Ac[D_h, (\{0N\}, \{1N\})]$ includes a cycle labeled b at state $(\{3N\}, \{2F\})$, that is not a certain state, and $Ac[D_h, (\{1N\}, \{0N\})]$ contains a cycle labeled b at state $(\{2F\}, \{3N\})$, that is not a certain state, in accordance with Theorem 22, event f is not robustly predictable when the attack occurs.

7 Conclusion

We consider the problem of robust predictability against sensor attacks. Based on a novel structure called hybrid diagnoser, an approach to test robust predictability is provided.

In the future, on one hand, as the construction of the diagnoser has exponential complexity, we intend to construct a verifier, which has polynomial complexity, to test robust predictability. On the other hand, we will try to extend the approach proposed in this work to the decentralized case.

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

QZ writes the manuscript. The author agrees to be accountable for the content of the work.

Funding

This work was supported by the Scientific Research Startup Fund of East China University of Technology.

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

References

1. Genc S, Lafortune S. Predictability of event occurrences in partially-observed discrete-event systems. *Automatica* (2009) 45:301–11. doi:10.1016/j.automatica.2008.06.022

2. Kumar R, Takai S. Decentralized prognosis of failures in discrete event systems. *IEEE Trans Autom Control* (2010) 55:48–59. doi:10.1109/TAC.2009.2034216

3. Takai S, Kumar R. Distributed failure prognosis of discrete event systems with bounded-delay communications. *IEEE Trans Autom Control* (2012) 57:1259–65. doi:10.1109/TAC.2011.2173419

4. Takai S, Kumar R. Distributed prognosis of discrete event systems under bounded-delay communications. In: Proc 48h IEEE Conf Decis Control, & 28th Chinese Control Conf; December 2009; Shanghai, China. IEEE (2009). p. 1235–40. doi:10.1109/CDC.2009.5399980

5. Chang M, Dong W, Ji Y, Tong L. On fault predictability in stochastic discrete event systems. *Asian J Control* (2013) 15:1458–67. doi:10.1002/asjc.748

6. Chen J, Kumar R. Stochastic failure prognosability of discrete event systems. *IEEE Trans Autom Control* (2015) 60:1570–81. doi:10.1109/TAC.2014.2381437

7. Liao H, Liu F, Wu N. Robust predictability of stochastic discrete-event systems and a polynomial-time verification. *Automatica* (2022) 144:110477. doi:10.1016/j.automatica.2022.110477

8. Benmessahel B, Touahria M, Nouioua F. Predictability of fuzzy discrete event systems. *Discrete Event Dyn Syst* (2017) 27:641–73. doi:10.1007/s10626-017-0256-7

9. Yin X, Li Z. Reliable decentralized fault prognosis of discrete-event systems. *IEEE Trans Syst Man Cybern: Syst* (2016) 46:1598–603. doi:10.1109/TSMC.2015.2499178

10. Takai S. Robust prognosability for a set of partially observed discrete event systems. *Automatica* (2015) 51:123–30. doi:10.1016/j.automatica.2014.10.104

11. Xiao C, Liu F. Robust fault prognosis of discrete-event systems against loss of observations. *IEEE Trans Autom Sci Eng* (2022) 19:1083–94. doi:10.1109/TASE.2021.3049400

12. Ammour R, Leclercq E, Sanlaville E, Lefebvre D. Fault prognosis of timed stochastic discrete event systems with bounded estimation error. *Automatica* (2017) 82:35–41. doi:10.1016/j.automatica.2017.04.028

13. Yin X. Verification of prognosability for labeled petri nets. *IEEE Trans Autom Control* (2018) 63:1828–34. doi:10.1109/TAC.2017.2756096

14. You D, Wang S, Seatzu C. Verification of fault-predictability in labeled petri nets using predictor graphs. *IEEE Trans Autom Control* (2019) 64:4353–60. doi:10.1109/TAC.2019.2897272

15. Sampath M, Sengupta R, Lafortune R, Sinnamohideen K, Teneketzis D. Diagnosability of discrete-event systems. *IEEE Trans Autom Control* (1995) 40:1555–75. doi:10.1109/9.412626

16. Alves MV, Barcelos RJ, Carvalho LK, Basilio JC. Robust decentralized diagnosability of networked discrete event systems against Dos and deception attacks. *Nonlinear Analysis: Hybrid Syst* (2022) 44:101162. doi:10.1016/j.nahs.2022.101162

17. Li Y, Hadjicostis CN, Wu N, Li Z. Error- and tamper-tolerant state estimation for discrete event systems under cost constraints. *IEEE Trans Autom Control* (2023) 1–8. doi:10.1109/TAC.2023.3239590

18. Zhang Q, Seatzu C, Li Z, Giua A. Joint state estimation under attack of discrete event systems. *IEEE Access* (2021) 9:168068–79. doi:10.1109/ACCESS.2021.3135870

19. Meira-Góes R, Kang E, Kwong RH, Lafortune S. Synthesis of sensor deception attacks at the supervisory layer of Cyber-Physical Systems. *Automatica* (2020) 121:109172. doi:10.1016/j.automatica.2020.109172

frontiers | Frontiers in Physics

TYPE Review
PUBLISHED 09 May 2023
DOI 10.3389/fphy.2023.1198457

OPEN ACCESS

EDITED BY
Huadan Zheng,
Jinan University, China

REVIEWED BY
Dekui Li,
Hefei University of Technology, China
Shengwei Cui,
Hebei University, China
Yuxiang Wu,
University of Electronic Science and Technology of China, China

*CORRESPONDENCE
Xiaopeng Shao,
xpshao@xidian.edu.cn

RECEIVED 01 April 2023
ACCEPTED 18 April 2023
PUBLISHED 09 May 2023

CITATION
Li X, Liu Z, Cai Y, Pan C, Song J, Wang J and Shao X (2023), Polarization 3D imaging technology: a review. *Front. Phys.* 11:1198457. doi: 10.3389/fphy.2023.1198457

Polarization 3D imaging technology: a review

Xuan Li[1,2], Zhiqiang Liu[1], Yudong Cai[1,2], Cunying Pan[1], Jiawei Song[1], Jinshou Wang[1] and Xiaopeng Shao[1,2]*

[1]Hangzhou Institute of Technology, Xidian University, Hangzhou, China, [2]School of Optoelectronic Engineering, Xidian University, Xi'an, China

Polarization three-dimensional (3D) imaging technology has received extensive attention in recent years because of its advantages of high accuracy, long detection distance, simplicity, and low cost. The ambiguity in the normal obtained by the polarization characteristics of the target's specular or diffuse reflected light limits the development of polarization 3D imaging technology. Over the past few decades, many shape from polarization techniques have been proposed to address the ambiguity issues, i.e., high-precision normal acquisition. Meanwhile, some polarization 3D imaging techniques attempt to extend experimental objects to complex specific targets and scenarios through a learning-based approach. Additionally, other problems and related solutions in polarization 3D imaging technology are also investigated. In this paper, the fundamental principles behind these technologies will be elucidated, experimental results will be presented to demonstrate the capabilities and limitations of these popular technologies, and finally, our perspectives on the remaining challenges of the polarization 3D imaging technology will be presented.

KEYWORDS

polarization 3D imaging, ambiguity normal, single view, passive imaging, depth fusion, deep learning, challenge analysis

1 Introduction

As an important approach for humans to record and perceive environmental information, traditional optoelectronic imaging techniques are increasingly ineffective due to the loss of high-dimensional information [1, 2]. With advancements in new sensors, data transmission, and storage, more information about the light field such as phase, polarization, and spectral information can be efficiently detected and recorded [3–6], which helps to construct a functional relationship between the reflected light information and the contour of the object surface to obtain 3D information. Currently, the 3D reconstruction technique has drawn widespread attention and achieved great progress in face recognition, industrial inspection, autonomous driving, and digital imaging [7–16].

Existing 3D imaging technologies can be divided into two major categories: active and passive techniques [17]. Generally, these methods vary in cost, hardware configuration, stability, running speed, and resolution. The active techniques use active illumination for 3D reconstruction such as time-of-flight (TOF), lidar 3D imaging, and structured light 3D imaging. Specifically, the TOF technique [18] employs an active emitter to modulate the light in the time domain and an optical sensor to collect the light scattered back by the object, and finally, it recovers depth information by calculating the time delay from the signal leaves the device and the signal returns to the device. The TOF technique has been widely used in commercial products like Kinect II, but the technology is susceptible to ambient light interference and limited by the temporal resolution of the signal system, so the achieved depth resolution is usually not high. The lidar 3D imaging [19]

adopts the laser ranging principle to acquire system-target micro-surface element distance information, and then the 3D information of the target surface is obtained by mechanical scanning or beam deflection. Therefore, lidar 3D imaging has poor real-time performance for 3D imaging of large targets, and it is difficult to achieve popularity due to its complex mechanical structure, which leads to large system size and high cost. The structured light 3D imaging technology utilizes a projection device to actively project structured patterns. For the structured light 3D imaging, a one-to-one correspondence is constructed between points in the camera plane and those in the projection plane by decoding the captured contour images [20, 21], and camera calibration parameters are combined to obtain 3D point cloud data. Despite the advantage of high-precision imaging, the structured light 3D imaging technology still suffers from the problem of poor resistance to ambient light interference and decreasing accuracy with increasing detection distance.

The passive techniques with no active illumination for 3D reconstruction mainly include stereo vision and light field cameras. Specifically, the stereo-vision system [22] captures images from at least two different viewpoints and finds corresponding points from these images for 3D coordinate calculations based on triangulation. Because its reconstruction accuracy is inversely proportional to the length of the camera baseline, the stereo vision method is difficult to obtain high-precision 3D surface information in long-distance detection. The light field camera 3D imaging technique [23] acquires light source direction by embedding a micro-lens array between the lens and detector, thus obtaining 3D information under passive conditions; however, similar to stereo vision, this technique is limited by the distance between the micro-lens arrays, so it cannot realize long-distance 3D imaging and suffers from a low imaging accuracy.

With the increasingly urgent demand for long-range, high-precision, and high-dimensional target information in many fields such as security surveillance, deep-space exploration, and target detection, 3D imaging with higher performance through deep mining and decoding the multi-dimensional physical information of the optical field has become the mainstream research direction. Since the 1970s, domestic and foreign researchers have investigated the utilization of polarization information for 3D shape recovery of target surfaces, and they have developed a series of polarization 3D imaging methods [24–27]. The core of these methods is to exploit Fresnel laws to establish the functional relationship between reflected light polarization characteristics and three-dimensional contour. Benefiting from the special reconstruction mechanism, polarization 3D imaging technology has the advantages of high reconstruction accuracy, simple detection equipment, and non-contact 3D reconstruction. Therefore, this paper will primarily focus on representative Shape from Polarization (SfP) techniques. This paper elucidates the principles of polarization 3D imaging based on specular reflection and diffuse reflection and presents some typical technical theories and their experimental results.

The rest of the paper is organized as follows: Section 2 introduces the basics of polarization 3D imaging based on specular reflection and diffuse reflection. Section 3 discusses the principles of SfP techniques along with some experimental results to demonstrate their performances; Section 4 presents our perspectives on the challenges of polarization 3D imaging technology; Section 5 summarizes this paper.

2 Basics of polarization 3D imaging

2.1 Principles of the polarization 3D imaging

Since the 3D contour of the target surface can be uniquely determined by the normal vector [28, 29], the polarization 3D imaging technology reconstructs the target 3D contour by obtaining the normal vector information of the surface micro-surface element. As shown in Figure 1A, the reflected light is detected by the detector through the polarizer when the incident light reaches the object surface and reflects [30]. The incident light, the object surface normal and the reflected light are in the same plane, ϕ is the angle of polarization (AoP), the transmission direction of the reflected light is z-axis positive direction under the assumption of orthographic projection. The object surface normal representation can be directly displayed in Figure 1B, where θ is the zenith angle of the object surface normal, and φ is the azimuth angle [31]. From the above analysis of the polarization 3D imaging process, the normal vector of the object surface based on the polarization characteristics of reflected light is expressed as $\vec{n} = [\tan\theta\cos\varphi, \tan\theta\sin\varphi, 1]$. In practice, the solution to the target normal vector is obtained by calculating polarization characteristics: the zenith angle (equal to the incident angle) θ and the azimuth angle φ. The zenith angle θ can be estimated from the degree of polarization (DoP) and the target surface refractive index (n). Meanwhile, the azimuth angle φ can be obtained from the AoP, corresponding to the polarizer rotation angle at which the detector acquires the maximum light intensity when rotating the polarizer. However, due to the difference between Fresnel reflection and transmission coefficients [30], 3D imaging based on the polarization characteristics of diffuse and specular reflection needs to be discussed separately, which will be detailed in Section 2.2.

2.2 Polarization 3D imaging model based on reflected lights

According to Fresnel Laws, light reflection and refraction occur when unpolarized light arrives at the target surface, which causes a change in the polarization state of the incident light [32]. Polarization state changes are significantly different for reflected and refracted light, and they are expressed as [29, 31]:

$$
\begin{aligned}
P_r &= \left|\frac{r_p^2 - r_s^2}{r_p^2 + r_s^2}\right| = \frac{\sqrt{\sin^4\theta\cos^2\theta(n^2 - \sin^2\theta)}}{[\sin^4\theta + \cos^2\theta(n^2 - \sin^2\theta)]/2} \\
P_t &= \left|\frac{\frac{n_2\cos\theta_2}{n_1\cos\theta_1}t_p^2 - \frac{n_2\cos\theta_2}{n_1\cos\theta_1}t_s^2}{\frac{n_2\cos\theta_2}{n_1\cos\theta_1}t_p^2 + \frac{n_2\cos\theta_2}{n_1\cos\theta_1}t_s^2}\right| \\
&= \frac{\left(n - \frac{1}{n}\right)^2\sin^2\theta}{2 + 2n^2 - \left(n + \frac{1}{n}\right)^2\sin^2\theta + 4\cos\theta\sqrt{n^2 - \sin^2\theta}}
\end{aligned}
\tag{1}
$$

where r_s, r_p, t_s, and t_p denote the reflection and transmission coefficients when decomposing a plane light wave into vertical

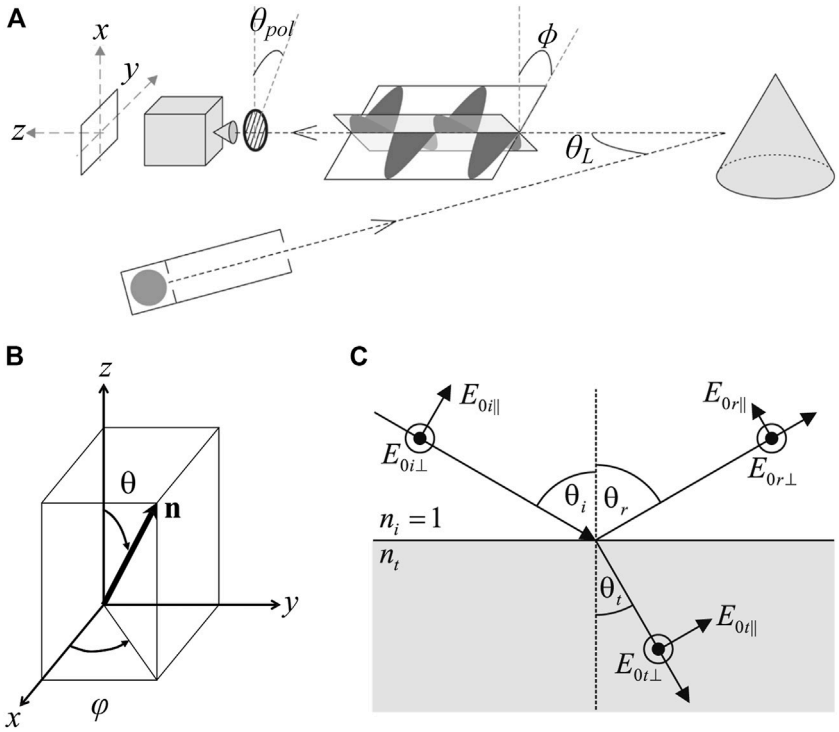


FIGURE 1
(A) The measurement process of polarization 3D imaging. Adapted from [30], with permission from IEEE. **(B)** The schematic of a normal vector. Reproduced/adapted from [31], with permission from IEEE. **(C)** The s- and p-components of reflected and refracted light. Reproduced from [30], with permission from IEEE.

and parallel components, i.e., s-component and p-component, respectively, as shown in Figure 1C. The reflection and transmission coefficients are defined as:

$$
\begin{aligned}
r_s &= \frac{E_{0rs}}{E_{0is}} = \frac{n_1 \cos\theta_1 - n_2 \cos\theta_2}{n_1 \cos\theta_1 + n_2 \cos\theta_2} \\
r_p &= \frac{E_{0rp}}{E_{0ip}} = \frac{n_2 \cos\theta_1 - n_1 \cos\theta_2}{n_2 \cos\theta_1 + n_1 \cos\theta_2} \\
t_s &= \frac{E_{0ts}}{E_{0is}} = \frac{2n_1 \cos\theta_1}{n_1 \cos\theta_1 + n_2 \cos\theta_2} \\
t_p &= \frac{E_{0ip}}{E_{0ip}} = \frac{2n_1 \cos\theta_1}{n_2 \cos\theta_1 + n_1 \cos\theta_2}
\end{aligned}
\tag{2}
$$

Since the polarization of the light emitted from the target surface differs in the reflected and refracted states, it is necessary to analyze the light state when it reaches the incident interface of different target surfaces, which provides a basis for selecting suitable polarization 3D reconstruction methods for different object surfaces. In 1991, Wolff provided a detailed classification and description of the types of light emitted from the target surface [28]. As shown in Figure 2A, Wolff classified the light into four categories: specular reflection light, diffuse reflection light, body reflection light (which can also be considered as special diffuse reflection light), and diffraction light. For body reflection light and body reflection light, they are generated with certain limitations and cannot reveal the true relationship between the target structure and the incident angle of the light, so the two types of light are usually not considered in recovering the 3D contours of the target and hence the existing polarization 3D imaging technology is mainly based on two different polarization characteristics of reflected light, namely, specular reflection and diffuse reflection. Specifically, as shown in Figure 2B, for smooth surfaces such as glass and metal, the reflected light is mainly displayed as specular reflection light with polarization information. For Lambertian objects like plaster, walls, and wood, it is usually assumed that the incident light enters the interior of the object, then scattered several times into unpolarized light, and finally transmitted into the air and received by the detector. Therefore, the selection of polarization 3D imaging methods based on different polarization characteristics is necessary for target surfaces of different materials, corresponding to two major methods based on specular reflection and diffuse reflection.

2.3 The ambiguity problem of polarization normal vector

As illustrated in Figure 1, the normal vectors of the target surface can be determined by the zenith angle θ and the azimuth angle φ. Therefore, to obtain accurate normal vector information in the study of polarization 3D imaging technology, it is necessary to accurately acquire the above two normal vector parameters. The following will analyze the problems of accurate normal vector acquisition based on specular reflection and diffuse reflection polarization characteristics, respectively.

10.3389/fphy.2023.1198457

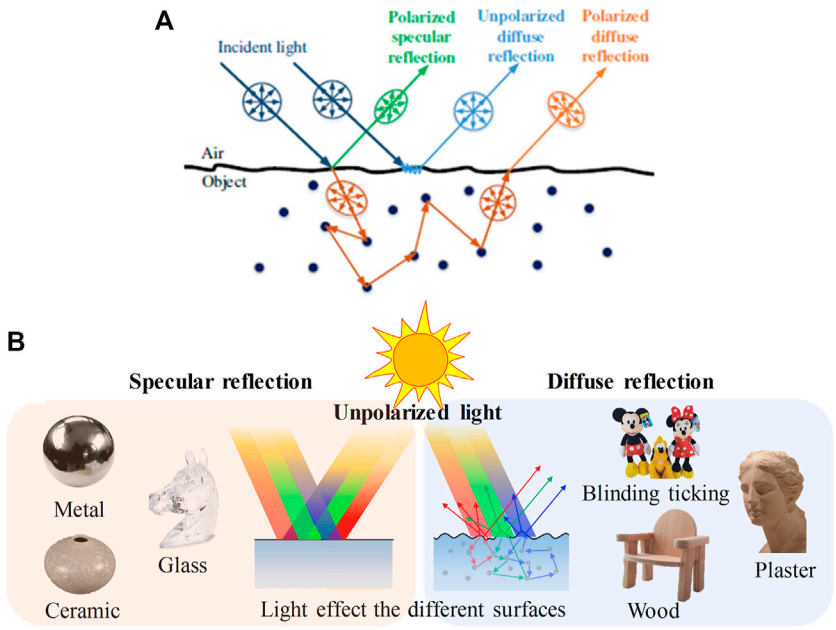


FIGURE 2
(A) Different types of reflected light from surfaces. Reproduced from [33], with permission from IEEE. **(B)** Polarization 3D imaging of different materials. Reproduced under CC-BY-4.0– [34] -http://journal.sitp.ac.cn/hwyhmb/hwyhmben/site/menu/20101220161647001.

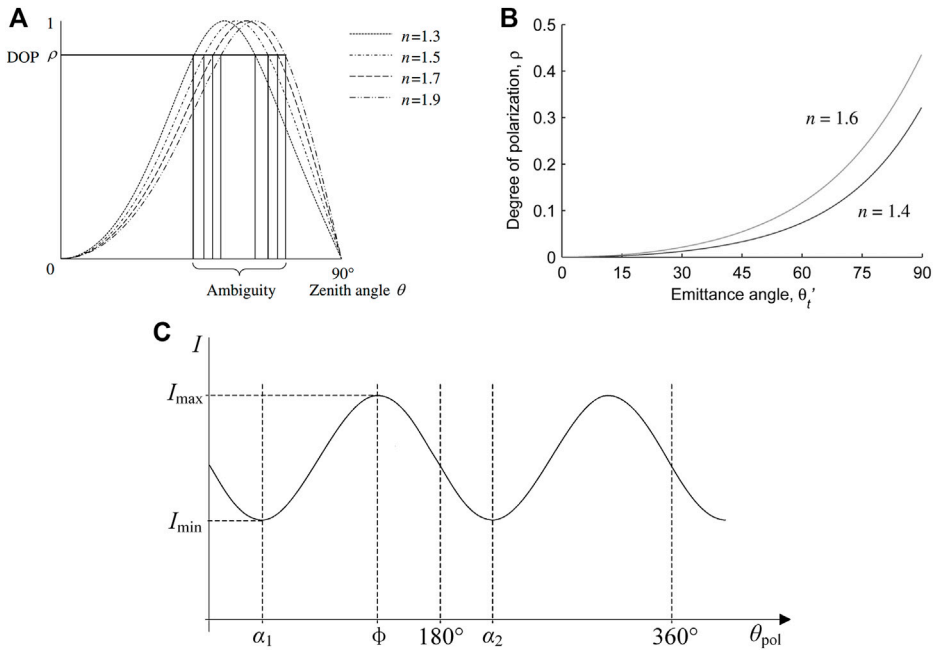


FIGURE 3
(A) The relationship between the degree of polarization and the incident angle with different refractive indices in polarization 3D imaging based on specular reflection. **(B)** The relationship between the degree of polarization and the incident angle with different refractive indices in polarization 3D imaging based on diffuse reflection. **(C)** The variation of light intensity information with the rotation angle of the polarizer. **(A–C)** Reproduced from [30], with permission from IEEE.

Since specular and diffuse reflection light emitted from the object surface follow the laws of reflection and transmission, the relationship between DoP and the zenith angle is shown in Eq. 2. Figure 3 presents the characteristic curves of polarization *versus* zenith angle θ for specular and diffuse reflections, respectively. It can be seen from Figures 3A,B that in the process of solving the zenith

angle based on specular reflection, a degree of polarization corresponds to two zenith angles θ, which are on both sides of the Brewster angle, causing the ambiguity of the zenith angle. In contrast, for the acquisition of the zenith angle based on diffuse reflection, the degree of polarization varies monotonically in the range of $\theta = [0^{\circ}, 90^{\circ}]$, indicating that the degree of polarization has a one-to-one correspondence with the zenith angle. Therefore, the solution to the zenith angle is a major challenge in polarization 3D imaging technology based on specular reflection.

Meanwhile, according to Malus laws [35], the relationship between the light intensity and the polarizer rotation angle can be represented as:

$$I = \frac{I_{\max} + I_{\min}}{2} + \frac{I_{\max} - I_{\min}}{2}\cos\left(2\theta_{\text{pol}} - 2\phi\right) \tag{3}$$

where $I_{\max}$ and $I_{\min}$ denote the maximum and minimum light intensity received by the detector during rotation, respectively, θ_{pol} denotes the rotation angle of the polarizer, and ϕ denotes the polarization angle. Figure 3C shows the actual measured variation curve of diffuse light intensity based on the change in the polarizer rotation angle [30], and combined with Eq. 3, it can be seen that the light intensity I reaches the maximum value when the polarizer rotation angle is equal to ϕ or ϕ+180°. In addition, since the curve fitting approach requires large numbers of polarization images, which is complex and need intensive calculations, Wolff used Stokes vectors to obtain the angle of polarization, which requires only three images with polarization direction at 0°, 45°, and 90°. Due to no one-to-one correspondence between $I_{\max}$ and the angle of polarization ϕ, there is a multi-valued problem of $\varphi = \phi$ or $\varphi = \phi \pm \pi$ in the solution of the azimuth angle, which leads to ambiguity in the normal vector information obtained from polarization. Meanwhile, the ambiguity problem of the azimuth angle still exists in the polarization 3D imaging technology based on specular reflection, which can be expressed as $\varphi = \phi \pm \pi/2$.

Solving the ambiguity problem of the zenith angle θ and azimuth angle φ is the major focus and difficulty for researchers to develop polarization 3D imaging technology. Meanwhile, other problems in polarization 3D imaging, such as the inability to obtain absolute depth information and the difficulty of specular-diffuse reflection separation, need to be solved. For different types of reflected light, various methods have been proposed to solve these problems. In this paper, typical techniques and methods for addressing the ambiguous normal vector and other challenges will be reviewed from the perspectives of polarization 3D imaging technology based on specular reflection and diffuse reflection. Here, the techniques for eliminating the ambiguous normal vector and their characteristics are outlined in Table 1 to help readers better grasp the core of this paper.

3 Polarization 3D imaging technology

3.1 Polarization 3D imaging based on specular reflection

Fresnel Laws indicate that the polarization characteristics of specular reflection light are easier to detect and more distinct than that of diffuse reflection light. Thus, researchers initially mainly used polarization characteristics of specular reflection information for 3D imaging of smooth surface materials like metals, transparent glass, and other specular targets. Many polarization 3D imaging techniques based on specular reflection have been proposed to solve the ambiguity problem of zenith and azimuth angles.

3.1.1 Rotational measurement method

Miyazaki et al. [31] developed a method to solve the ambiguity problem of the zenith angle by rotating the object. By assuming that the target surface is smooth, closed, and non-shaded, they divided the target surface into the Brewster-Equatorial region (B-E), the Brewster-South Pole region (B-N), and the Brewster-Brewster regions (B-B) based on different values of polarization, as shown in Figures 4B,C. It was assumed that the B-E region contained a region that obscures the boundary so that a boundary point zenith angle $\theta = 90^{\circ}$ existed in this region, and the zenith angle is determined by constraining the range of zenith angles of the points in the region $\theta_B < \theta < 90^{\circ}$. For the B-N region containing a region of pixel points with a zenith angle $\theta = 0^{\circ}$, the constraint on the zenith angle of the region is $0^{\circ} < \theta < \theta_B$. For the solution to the ambiguity problem in the B-B region, polarization images of the target before and after rotation are obtained, as shown in Figure 4A. Based on the relationship between the difference in the DoP of the corresponding point before and after rotation, the first-order derivative of the DoP and the rotation angle, i.e., $\rho(\theta+\Delta\theta)$-$\rho(\theta) \approx \rho'(\theta)\Delta\theta$, as well as the positive or negative sign of $\rho'(\theta)$ can determine whether the zenith angle θ of the pixel points in the B-B region belong to $0^{\circ} < \theta < \theta_B$ or $\theta_B < \theta < 90^{\circ}$, as shown in Figures 4D,E.

Since the rotational measurement method does not need to obtain the specific value of the rotation angle, calibration of the imaging system can be avoided. Even if there is an error in the reflected light polarization value, it does not affect the judgment of the Brewster angle in the first-order derivative of DoP, so the method has high robustness. However, multi-angle information acquisition and multiple measurements increase the complexity of the imaging system. Meanwhile, the existence of internal mutual reflection of transparent objects will result in a large error and poor reconstruction accuracy. Additionally, the method cannot image the moving target.

3.1.2 Polarization 3D imaging techniques with multi-spectrum

In 1995, Partridge et al. [36] analyzed the difference in DoP between reflected and transmitted light, and it was found that when the light was transmitted from the interior of the object, i.e., the emitted light was diffuse reflection light, its DoP corresponded to the zenith angle uniquely. Therefore, based on the target infrared radiation characteristics, Partridge adopted a far-infrared band detector for imaging to solve the zenith angle uniquely. However, random and systematic errors in the infrared detection system have a large impact on the imaging results.

Later, Miyazaki et al. [37] combined visible imaging with far-infrared imaging to obtain a unique solution of the zenith angle. To effectively detect the infrared information of the target, they used a hair dryer to heat the target surface and collected 36 infrared polarization images at different polarization angles by rotating the infrared polarizer to obtain the DoP of the target in the far-infrared band. By exploiting the unique correspondence between the

TABLE 1 Overview of SfP for eliminating the ambiguous normal vector.

Techniques	Problem solved	Illumination	Waveband	Objects	Complexity
Miyazaki et al. [31]	Zenith angle	Passive	Visible	Specular	High
Partridge et al. [36]	Zenith angle	Passive	Far-infrared	Specular	Middle
Miyazaki et al. [37]	Zenith angle	Passive	Visible and far-infrared	Specular	High
Stolz et al. [38]	Zenith angle	Passive	Multiple	Specular	Middle
Hao et al. [39, 40]	Zenith angle and refractive index	Passive	Multiple	Specular	Low
Morel et al. [41–43]	Azimuth angle	Active	Visible	Specular	High
Cui et al[33]	Azimuth angle	Passive	Visible	Specular or complex	Less high
Miyazaki et al. [44]	Normal vector	Passive	Visible	Black and specular	Less high
Atkinson et al. [45]	Azimuth angle	Active	Visible	Complex	High
Mahmoud et al. [46]	Azimuth angle	Passive	Visible	Diffuse	Low
Kadambi et al. [47].	Azimuth angle	Passive	Visible	Complex	Less high
Tian et al. [48]	Azimuth angle	Passive	Visible	Complex	Less high
Liu et al. [49]	Normal vector	Active	Visible	Complex	Less high
Deschaintre et al. [50]	Normal vector	Active and flash	Visible	Dielectric	Less high
Lei et al. [51]	Normal vector	Passive	Visible	Wild	Middle
Shao et al. [52]	Azimuth angle	Passive	Visible	Human face	Middle

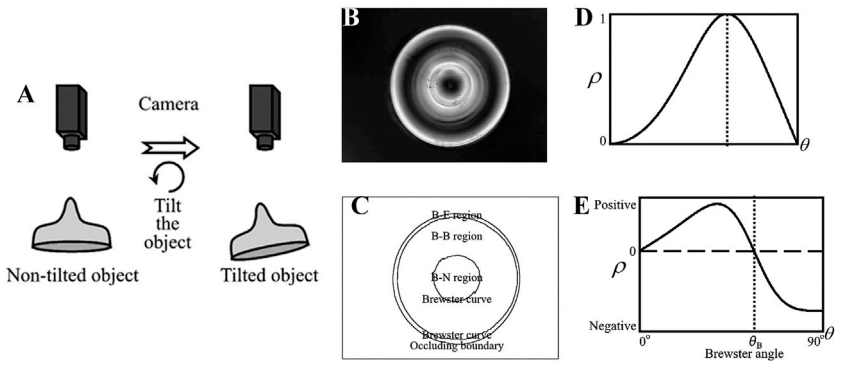


FIGURE 4
Schematic diagram of Miyazaki's experiment. Reproduced from [31], with permission from IEEE. **(A)** The target information acquisition process with the target rotating at a small angle. **(B)** Polarization degree. **(C)** Areas divided by Brewster's corner. **(D)** The relation curve between the polarization degree and the incident angle. **(E)** The derivative of polarization degree.

DoP and the zenith angle in the far-infrared band, they avoided the ambiguity problem of the zenith angle in visible imaging. However, due to different types of visible and infrared imaging, the imaging system of the method is complex and costly, and there are problems in practical applications such as the need to match between images of different wavelengths, increasing the complexity of the application.

Inspired by the above methods, in 2012, Stolz et al. [38] proposed a method based on multi-spectral polarization processing to solve the ambiguity problem of the zenith angle and the problem of complex and expensive detection systems when obtaining information from multi-band systems. According to Cauchy's dispersion formula, the refractive index decreases with the increase in the incident light wavelength in the visible wavelength [53]. Combining with Eq. 1 and Figure 3A, it can be seen that the refractive index is a parameter of the DoP and when the incident light wavelength increases, the corresponding zenith angle-polarization curve will also shift to the right, i.e., the Brewster angle shifts to a larger coordinate direction. By analyzing the variation curve of polarization with the zenith angle at different incident light wavelengths, the zenith angle ambiguity problem can be eliminated. Figure 5A illustrates the variation curves of DoP at different

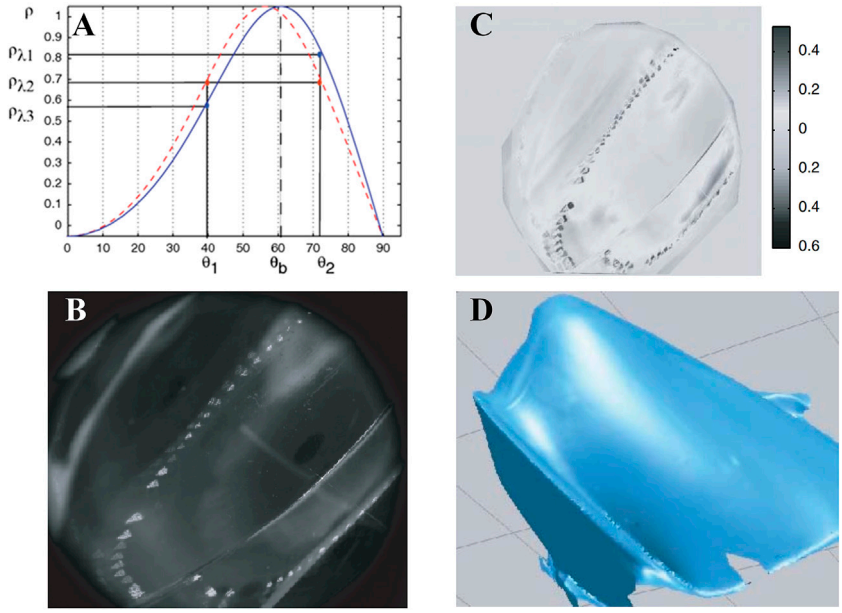


FIGURE 5
3D reconstruction of transparent targets with a partial high slope. Reproduced under CC-BY-4.0- [38] - https://opg.optica.org/ol/fulltext.cfm?uri=ol-37-20-4218&id=243204. **(A)** The DoP curve at different incident light wavelengths. **(B)** Intensity image. **(C)** Polarization degree. **(D)** 3D reconstruction result.

wavelengths, and the difference is exploited to solve the ambiguity of the zenith angle as follows: 1) Calculate the DoP at two different wavelengths (refractive index) $p_{\lambda 1}$ and $p_{\lambda 2}$ ($\lambda_1 > \lambda_2$), respectively; 2) Estimate the variability of polarization at different wavelengths $\Delta p = p_{\lambda 2}$-$p_{\lambda 1}$. 3) If Δp is larger than zero, the zenith angle $\theta = \theta_1$; otherwise, $\theta = \theta_2$. As shown in Figures 5B–D, Stolz et al. performed an experimental validation, and the result demonstrated that the method could realize undistorted 3D reconstruction of transparent targets. Additionally, the method can achieve accurate reconstruction acquisition for targets with local gradient mutations, providing an important reference for studying polarization 3D imaging techniques for complex target surfaces. However, because of the acquisition of intensity information in multiple bands, this method requires active illumination and cannot image in real time.

Hao and Zhao et al. [41, 39] also conducted an in-depth study on multi-spectral polarization information, and they proposed a method for solving the incidence angle and refractive index simultaneously by using multi-spectral polarization characteristics of the target, which realized polarization 3D reconstruction of highly reflective non-textured nonmetallic targets. Besides, they added wavelength information to the relationship between the zenith angle and DoP by using Cauchy's dispersion formula, so the problem of solving the zenith angle was transformed into a nonlinear least-squares problem. The utilization of spectral and polarization information effectively separates and suppresses stray light on the target surface, further improving the accuracy of target 3D reconstruction. The reconstruction results are illustrated in Figure 6.

In the study of polarization 3D imaging based on specular reflection light, the above-mentioned solutions to the ambiguity zenith angle problem have been introduced. After the accurate normal zenith angle θ is obtained, the last "obstacle" to realizing polarization 3D imaging based on specular reflection is to eliminate the ambiguity of the normal azimuth angle φ. In the early period of polarization 3D imaging, utilizing the ranking technique was the main method to solve the azimuthal ambiguity problem. This method determines the azimuth angle direction by assuming that the surface normal vector at the target boundary is perpendicular to the points at the boundary and there is no obvious "mutation" region on the target surface. Then, it eliminates surface azimuthal ambiguity by propagating the azimuth angle determined at the boundary to the interior of the target surface [31]. Later, Atkinson et al. conducted a related study on the ranking technique [30], but this method has weak applicability in complex target surfaces, and it requires high accuracy for the propagation algorithm, increasing the complexity of the polarization 3D imaging algorithm. Therefore, researchers have started to eliminate azimuthal angle ambiguity by changing the illumination conditions and combining priori information. Two representative types of azimuthal ambiguity elimination techniques are reviewed below.

3.1.3 Active illumination method

French scientists Morel et al. [41–43] proposed a method using active illumination to eliminate the ambiguity of azimuth angle. They constructed hemispherical diffuse dome light consisting of four mutual symmetric 1/8 spherical subsystems, as shown in Figure 7A. The system independently controlled four sub-sources to illuminate the target from different directions (east, south, west, and north) for four target hemisphere images, as

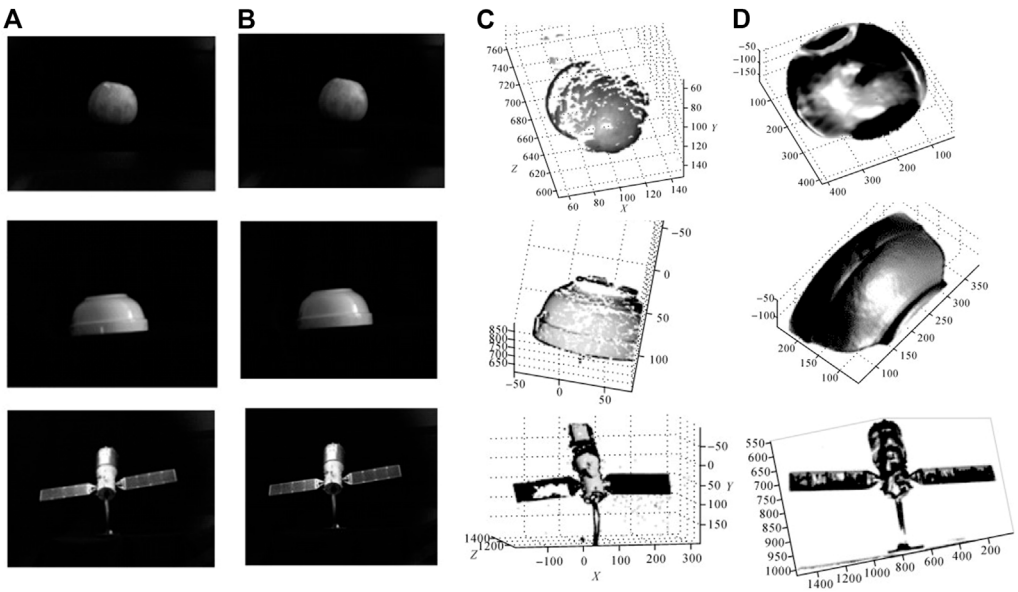


FIGURE 6
Surface reconstruction of the objects. Reproduced under CC-BY-4.0- [39] -http://xb.chinasmp.com/CN/10.11947/j.AGCS.2018.20170624. **(A)** Intensity images. **(B)** Results after removing the highlight. **(C)** Reconstruction results by stereo vision. **(D)** Reconstruction results by multi-spectral polarization.

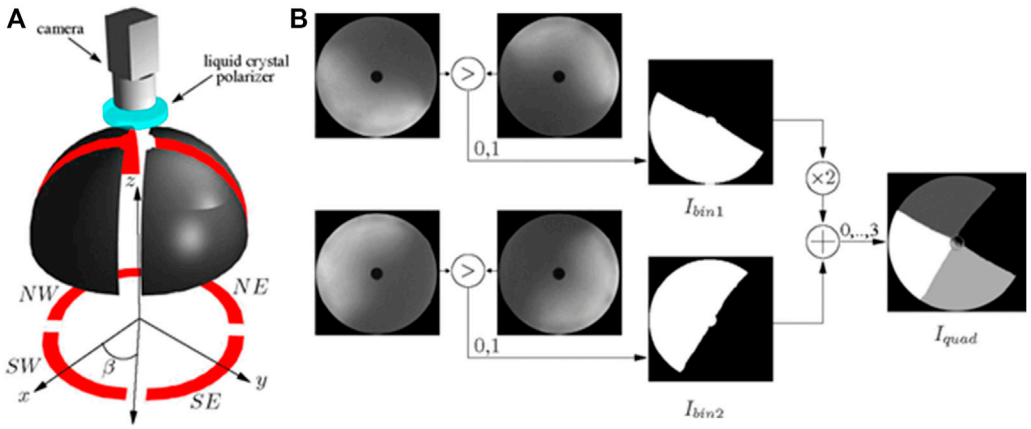


FIGURE 7
(A) The experimental diagram. **(B)** The acquisition principle of the segmented image. **(A,B)** Reproduced from [43], with permission from IEEE.

shown in Figure 7B. The binary image I_{bin1} that distinguishes between east and west directions was achieved by comparing the intensity images after illumination from the east and west directions. Similarly, the binary image I_{bin2}, which can distinguish between south and north directions, was obtained. The detailed procedures for solving the azimuthal ambiguity problem are: 1) φ (azimuthal angle) = ϕ (polarization angle) - $\pi/2$; 2) $I_{\text{quad}} = 2I_{\text{bin1}} + I_{\text{bin2}}$; 3) If $[(I_{\text{quad}} = 0) \wedge (\phi \leq 0)] \vee (I_{\text{quad}} = 1) \vee (I_{\text{quad}} = 3) \wedge (\phi \geq 0)$, $\varphi = \phi + \pi$.

However, in actual operations, the method requires multiple LED light sources and needs to regulate the light sources in different directions separately to solve the ambiguity of the azimuth angle, so it is complex and cannot be applied to moving targets. Meanwhile, it is difficult to implement in outdoor scenes.

3.1.4 Multi-view fusion method

In 2017 Cui et al. [33] provided a multi-view polarization 3D imaging technique for reconstructing smooth target surfaces and successfully realized complex target surface 3D reconstruction based on the study of the target local reflectivity. They obtained target intensity and polarization information from at least three viewpoints by setting up multiple polarization cameras at different spatial locations and recovered the camera position as well as an initial 3D shape through the methods of classical motion structure [54]

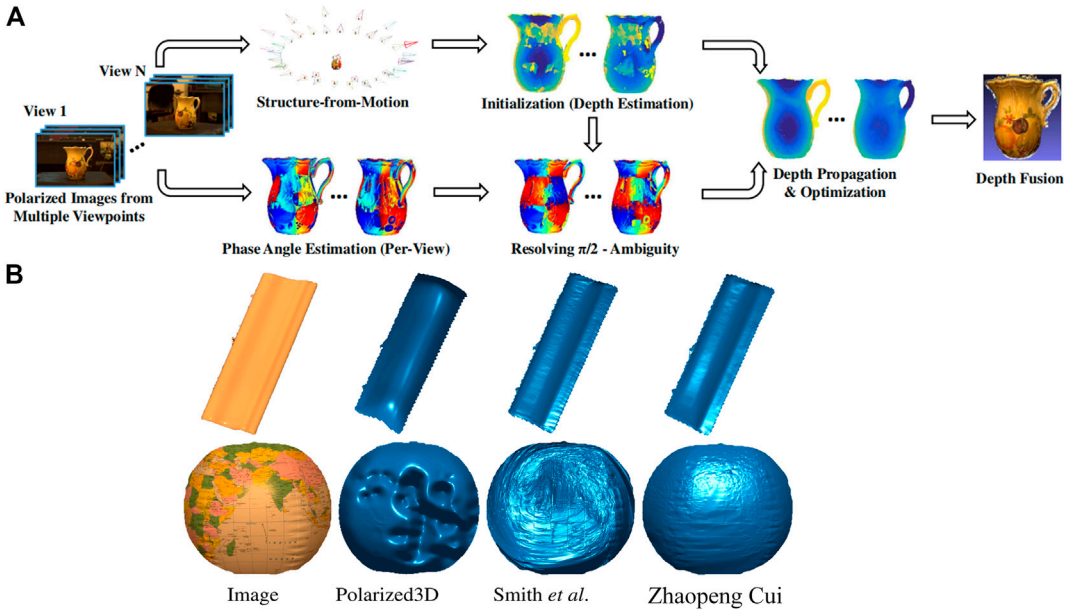


FIGURE 8
(A) The flowchart of the polarimetric multi-view stereo algorithm. **(B)** Comparison of depth estimation results of Cui [33], Polarized3D [57], and Smith [58]. **(A,B)** Reproduced from [33], with permission from IEEE.

and multi-view stereo [47, 55]. Then, the ambiguity zenith angle of complex object surfaces in high-frequency regions could be corrected by the acquired priori initial 3D shape information, as shown in Figure 8A. Additionally, drawing on the iso-depth contour tracking method [59] in photometric stereo, Cui et al. spreaded the azimuth angle information obtained from the recovery of high-frequency regions to low-frequency regions to eliminate the ambiguity azimuth angle in low-frequency regions. They compared the results with those of some other polarization 3D imaging methods [50, 57], as shown in Figure 8B, which illustrated the accuracy of the proposed method and the 3D imaging feasibility of target surfaces with different reflectivity. However, this method requires several standard points with reliable depth as the "seed" for depth propagation in the tracking process, so it cannot effectively eliminate the ambiguous azimuth angle problem when the feature of the target surface is not sufficient to provide multiple reliable depth reference points. Besides, the method cannot be applied to transparent objects for 3D reconstruction at present.

In the same year, Miyazaki et al. [44] also proposed a polarization 3D imaging technique based on the multi-view stereo to achieve 3D contours of black specular targets. They combined polarization 3D imaging with the space-carving technique. The corresponding points of each view angle image calculated from the camera pose obtained by camera calibration and the 3D shape obtained by spatial sculpting were exploited to analyze the phase angles at the same surface point. In this way, they acquired the surface normal of the entire object surface using the azimuth angle obtained from multiple viewpoints. The addition of polarization information compensates for the defects of the space carving technique in the 3D reconstruction such as the lack of detail texture.

3.2 Polarization 3D imaging based on diffuse reflection

Polarized 3D imaging based on specular reflection is sensitive to the direction of the light source when reconstructing metallic and transparent objects. However, it is not ideal for information extraction and 3D reconstruction of most targets in nature and needs to solve the zenith angle ambiguity problem. With the research and development in the field of materials science and new detectors, the ability to detect and analyze polarization information is improved, especially for the weak polarization properties in the optical field [60–62]. Therefore, an increasing number of researchers focus on the study of polarization 3D imaging based on diffuse reflection and have proposed many classical solutions to the azimuthal ambiguity problem.

3.2.1 Combined photometric stereo vision technique

Atkinson et al. [45] developed a SfP technique based on diffuse reflection light in 2007. They used photometric stereo vision with the illumination of multiple light sources to eliminate the azimuthal ambiguity in polarization 3D imaging, and their imaging system is shown in Figure 9A. They set up three light sources with fixed positions to collect target intensity images under different illumination conditions as shown in Figure 9A1 and then achieved the elimination of the ambiguity azimuth angle problem by comparing changes in light source intensity information received from different directions on different target areas. Figure 9A2 shows the schematic

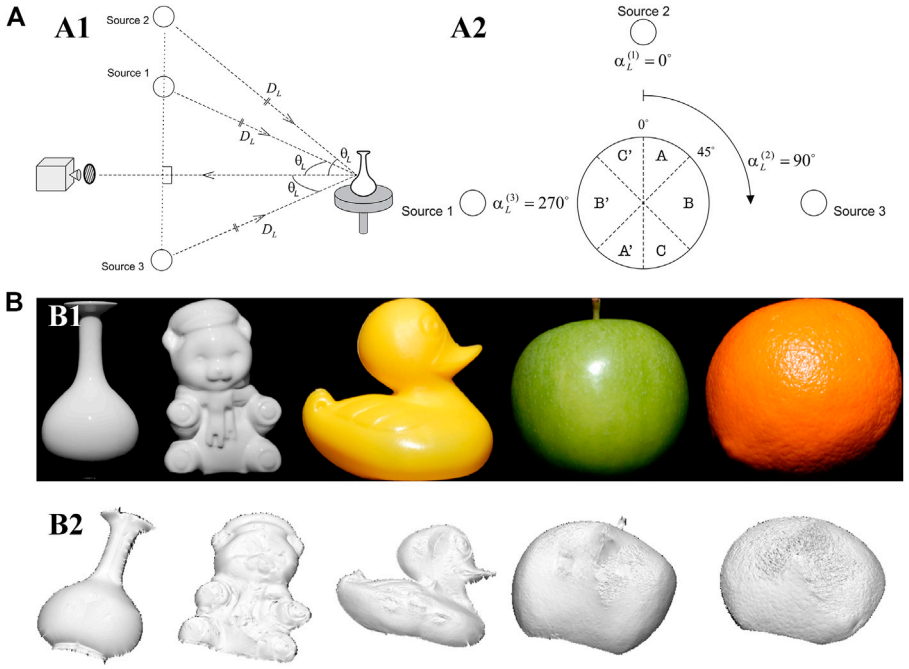


FIGURE 9
(A) The schematic diagram of the imaging system. Adapted from [45], with permission from Springer Nature. **(A1)** The geometric relationship of the imaging system. **(A2)** The view of a spherical target from the camera viewpoint. **(B)** Surface reconstruction of the objects. Adapted from [45], with permission from Springer Nature. **(B1)** The raw images of the test targets. **(B2)** The depth estimation of the targets.

diagram of the imaging system and the unique azimuth angle value is determined as follows:

$$\begin{aligned}
\text{if } \phi_k < 45^\circ \text{ then } \alpha_k &= \begin{cases} \phi_k \text{ if } I_k^{(2)} > I_k^{(1)} \\ \phi_k + 180^\circ \text{ otherwise} \end{cases} \\
\text{if } 45^\circ \leqslant \phi_k < 135^\circ \text{ then } \alpha_k &= \begin{cases} \phi_k \text{ if } I_k^{(3)} > I_k^{(1)} \\ \phi_k + 180^\circ \text{ otherwise} \end{cases} \\
\text{if } 135^\circ \leqslant \phi_k \text{ then } \alpha_k &= \begin{cases} \phi_k \text{ if } I_k^{(3)} > I_k^{(2)} \\ \phi_k + 180^\circ \text{ otherwise} \end{cases}
\end{aligned} \tag{4}$$

The reconstruction results of the method are presented in Figure 9B, which demonstrates that the 3D reconstruction of contour information can be achieved for targets with different surface materials. However, this technique is sensitive to the angle between multiple active light sources and the distance between the light sources and the target, which cannot be implemented easily in actual experiments. Meanwhile, the reconstructed 3D contours are smoother than the ground truth due to the unknown roughness, mutual reflection, and refractive index.

3.2.2 Combined shape from shading technique

In 2013, Mahmoud et al. [46] proposed a 3D imaging method that combines the polarization 3D imaging technique with the shape from shading (SFS) method. The 3D reconstruction results based on the SFS were exploited as priori deep information to solve the ambiguity problem of the azimuth angle in polarization 3D imaging technology. Since Mahmoud's polarization 3D imaging technique requires only one view and one imaging band, it is simple to operate, and the equipment required for imaging is easy to set up. The 3D imaging results of the technique are shown in Figure 10A. However, due to the application of the SFS, this technique assumes that the targets are all ideal Lambertian objects, resulting in limited applicable targets and sensitivity to stray light.

3.2.3 Depth map fusion technique

Kadambi et al. [47] proposed a method to fuse the depth map obtained by Kinect with polarization 3D imaging in 2017. Compared with polarization 3D imaging technology-based photometric stereo vision and SFS, this method avoided the estimation and assumption of scene information such as light sources and targets, and it extended the lighting conditions from special light sources to natural light, realizing high-precision polarization 3D reconstruction. The experimental setup included a Kinect, a normal SLR, and a linear polarizer. The "rough depth map" of the object surface with real depth information was obtained from Kinect, but due to the low resolution of Kinect, the details of the target surface could not be effectively recovered when reconstructing the 3D contour of the target. Therefore, Kadambi combined polarization 3D imaging results containing huge texture details of the target surface with the "rough depth map" to achieve high-precision 3D imaging in various scenes. The 3D imaging results and accuracy analysis for different scenes are shown in Figure 10B. However, due to the limitation of Kinect's effective detection distance, the polarization 3D imaging technique for depth map fusion cannot perform high-precision 3D imaging of targets at a long distance. Besides, as

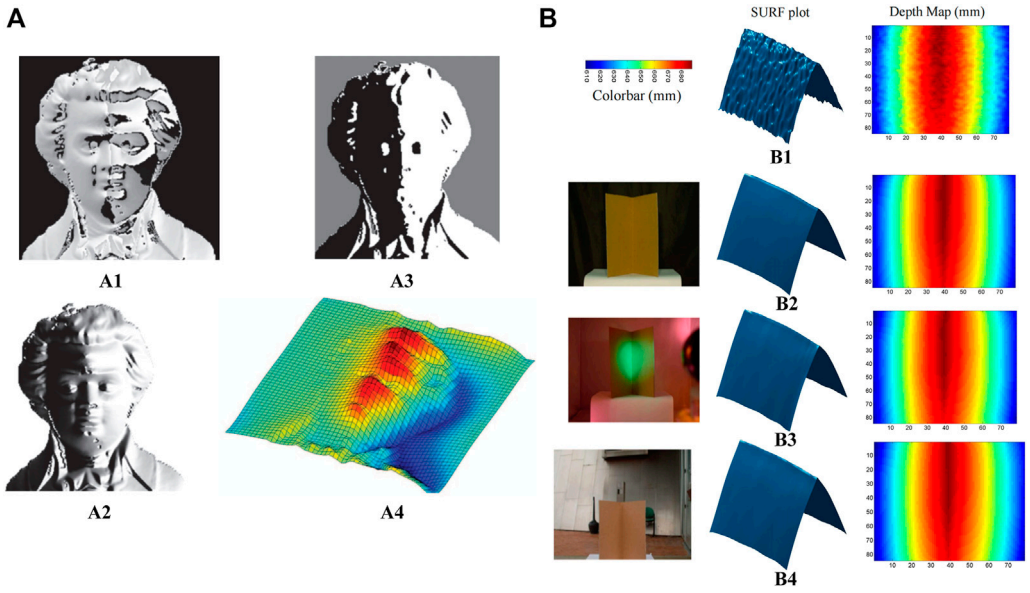


FIGURE 10
(A) Experimental results. Adapted from [46], with permission from IEEE. **(A1)** Polarization angle. **(A2)** Diffuse polarization degree. **(A3)** Intensity image. **(A4)** Reconstructed surface. **(B)** Polarization 3D imaging in various lighting conditions. Adapted from [47], with permission from Springer Nature. **(B1)** TOF Kinect. **(B2)** Polarization enhancement indoors. **(B3)** Polarization enhancement under disco lighting. **(B4)** Polarization enhancement outdoors on a partly sunny.

the "coarse depth map" is not consistent with the polarization 3D imaging in resolution and field of view, complex image processing techniques such as image scaling and registration are required in the actual reconstruction process.

3.2.4 Other fusion technique

Many polarization 3D imaging techniques fused with other imaging techniques have also been developed [48, 49, 63, 64], such as two excellent methods that have been proposed recently. Tian et al. [48] investigated a novel 3D reconstruction method based on the fusion of SfP and binocular stereo vision. They corrected the azimuth angle errors based on binocular depth; then, they proposed a joint 3D reconstruction model for depth fusion, including a data fitting term and a robust low-rank matrix factorization constraint, to achieve high-quality 3D reconstruction. A series of experiments on different types of objects were conducted to verify the efficiency of the proposed method in comparison with state-of-the-art methods. The reconstruction results presented in Figure 11A indicated that the proposed method can generate accurate 3D reconstruction results with fine texture details. However, it should be noted that due to the combination of binocular stereo vision, the increase of cameras, image scaling, and registration are unavoidable, increasing the cost and complexity of the imaging system. Liu et al. [49] proposed a 3D reconstruction method based on the fusion of SfP and polarization-modulated ranging (PMR), in which only a single image sensor was used to obtain both polarization images and depth data, thus avoiding the image registration problem. Since PMR can provide coarse but accurate and absolute depths and SfP can retrieve inaccurate 3D contours of objects with fine textures, they proposed two fusion models: a joint azimuth estimation model to obtain a fused azimuth angle with π-ambiguity corrected, and a joint zenith estimation model to estimate an accurate fused zenith angle, thus achieving high-quality reconstruction. The specific 3D imaging results are shown in Figure 11B. However, this technique needs further improvement in two aspects: 1) PMR requires active illumination, which makes the proposed technique difficult to be applied outdoors. 2) multiple images are required (three or more polarized images for SfP and two polarization-modulated images for PMR) and the light source needs to be switched.

3.3 Polarization 3D imaging based on deep learning

To overcome the limitations of scenes and objects in the polarization 3D imaging technology based on specular reflection or diffuse reflection, researchers have investigated polarization 3D imaging techniques that can be applied to complex scenes and objects. Recently, the rapid proliferation of deep learning, which has successful applications in other fields of 3D imaging [69–73], brings the possibility of breaking through the limitations in traditional polarization 3D imaging techniques. An increasing number of researchers have focused on solving the azimuthal ambiguity problem for complex targets in polarimetric 3D imaging through deep learning [51, 52, 74, 75]. Some successful polarization 3D imaging techniques combined with deep learning are outlined below.

Deschaintre et al. [75] combined polarization imaging with deep learning to achieve a high-quality estimate of 3D object

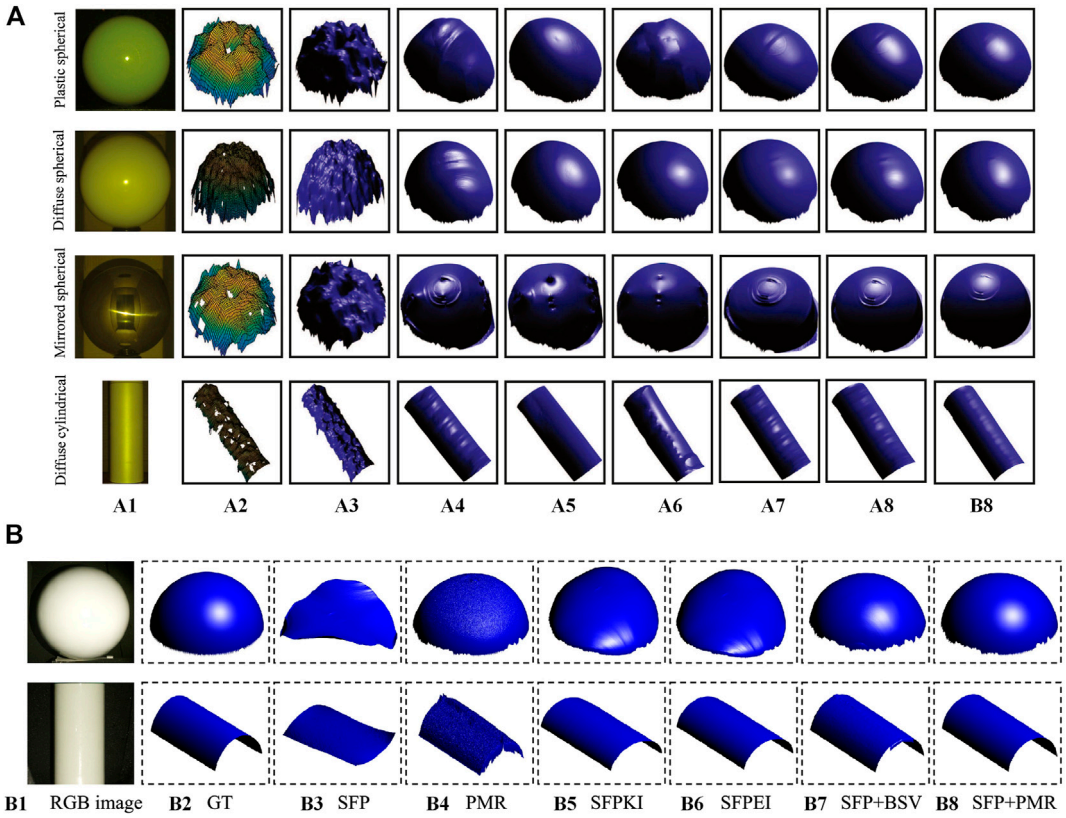


FIGURE 11
(A) Comparison of 3D reconstruction results on regular objects. Adapted from [48], with permission from Elsevier. **(A1)** RGB image. **(A2)** Binocular depth. **(A3)** MC [65]. **(A4)** DES [, 63, 57]. **(A5)** DRLPR [30]. **(A6)** SP [46]. **(A7)** SFPIK [58]. **(A8)** SFPIE [67]. **(A9)** Tian's method [48]. **(B)** Comparison of 3D reconstruction results on standard geometric targets. Adapted from [49], with permission from Elsevier. **(B1)** RGB image. **(B2)** GT. **(B3)** SFP. **(B4)** PMR [68]. **(B5)** SFPKI [58]. **(B6)** SFPEI [67]. **(B7)** SFP + BSV [48]. **(B8)** Liu's method [49].

shapes under frontal unpolarized flash illumination. To avoid the correction of ambiguity normal vectors, they used single-view polarization imaging to directly obtain surface normal vectors, depth, and other information like diffuse albedo, roughness, and specular albedo through an encoder-decoder architecture shown in Figure 12A. They trained their network on 512 × 512 images by using two losses: an L1 loss to regularize the training, compute an absolute difference between the output maps and the targets, and a novel polarized rendering loss. Figure 12B presents the comparison of reconstruction results with those of Li et al. [76]. It demonstrates that the technique can recover global 3D contours and other information about the object well. However, at present, the method can only be applied to flash illumination dielectric objects, and the utilization of flash illumination will bring a few specular highlights.

Lei et al. [51] applied polarization 3D imaging to complex scenes in the wild. They provided the first real-world scene-level SfP dataset with paired input polarization images and ground-truth normal maps to address the issue of lacking real-world SfP data in complex scenes. In addition to the application of multi-headed self-attentive convolutional neural networks (CNNs) for SfP, per-pixel viewing encoding was also applied to the neural network to handle non-orthographic projection for scene-level SfP. Then, they trained their network on 512 × 512 images by using a cosine similarity loss [50], and the reconstruction results are illustrated in Figure 13A, which reveals that this approach produces accurate surface normal maps. However, the datasets in this technique need to be collected and trained complicatedly, and they are only applicable to specific scenarios.

Shao et al. [52] proposed a learning-based method for passive 3D polarization face reconstruction. The method uses a CNN-based 3D morphable model (3DMM) to generate a rough depth map of the face from the directly captured polarization image. Then, the ambiguity surface normal obtained from polarization can be eliminated from the rough depth map. The construction results of the proposed method in both indoor and outdoor scenarios are shown in Figure 13B. Although the 3D faces are well reconstructed, the dataset requirements still exist in this technique.

3.4 Other polarization 3D imaging methods

The above-mentioned techniques are solutions to the major challenge (the ambiguity problem of normal vectors) in polarimetric 3D imaging. However, many other factors still affect the reconstruction accuracy throughout the polarization 3D imaging process.

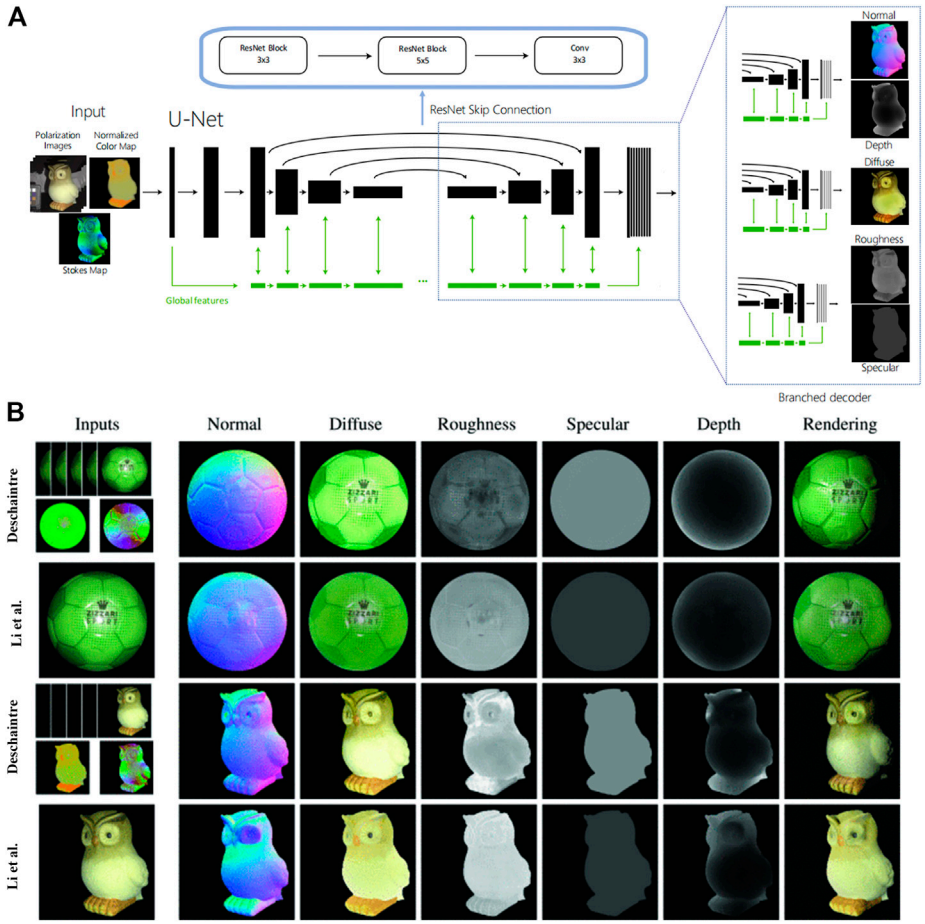


FIGURE 12
(A) Deschaintre's network architecture based on a general U-Net. **(B)** A comparison of Deschaintre's results and the results of Li et al. [76] on real objects. **(A,B)** Adapted from [75], with permission from IEEE.

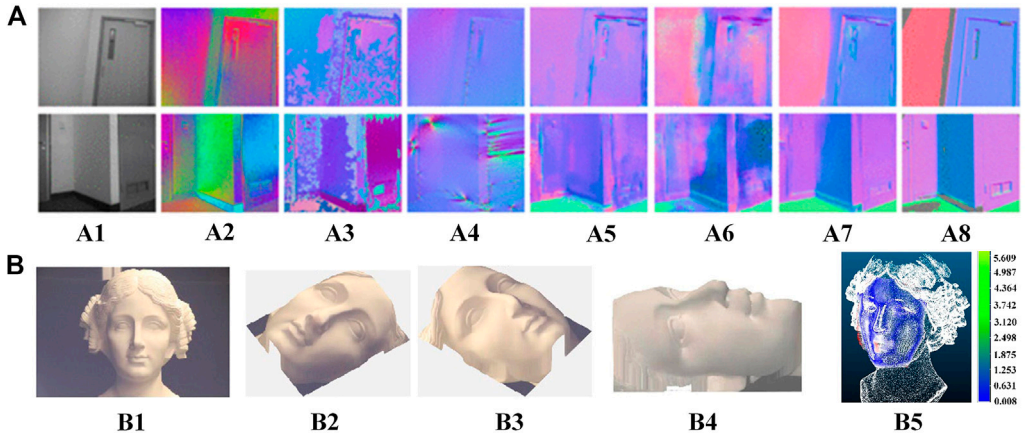


FIGURE 13
(A) Qualitative comparison between Lei's approach and other polarization methods. Reproduced from [51], with permission from IEEE. **(A1)** Unpolarized image input. **(A2)** polarization angle input. **(A3)** Miyazaki [77]. **(A4)** Smith [67]. **(A5)** DeepSfP [50]. **(A6)** Kondo [78]. **(A7)** Lei [51]. **(A8)** ground truth. **(B)** Reconstructed 3D faces using CNN-based 3DMM. Reproduced under CC-BY-4.0- [52] -https://www.mdpi.com/2304-6732/9/12/924. **(B1)** A plaster statue with indoor lighting. **(B2-B4)** three different views of the recovered 3D face of **(B1)**. **(B5)** point cloud comparison between the laser scanner and Shao's method.

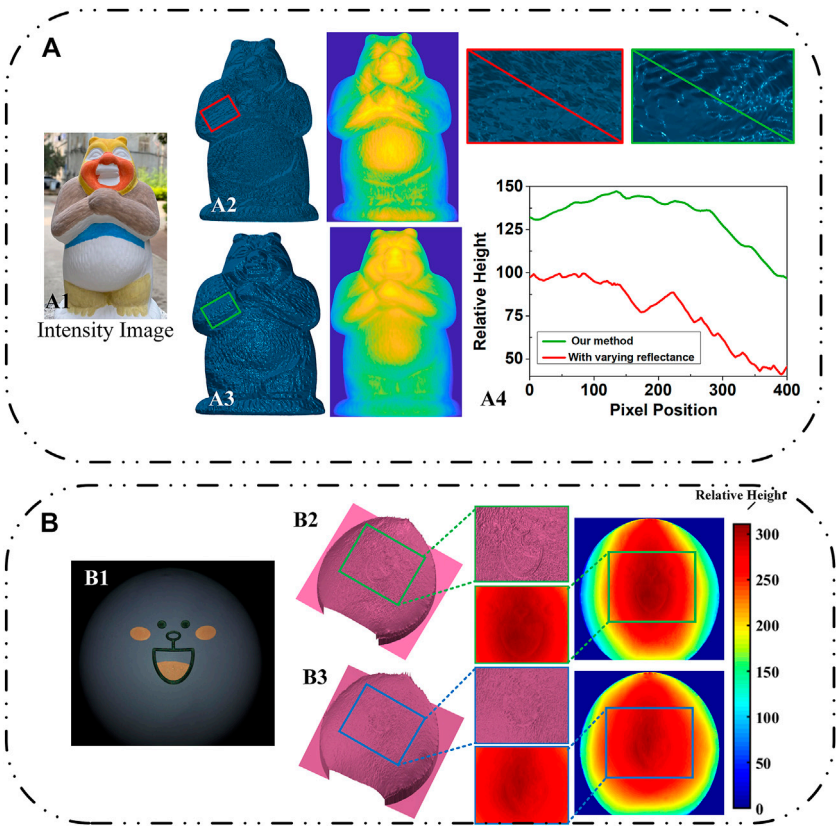


FIGURE 14
(A) Li's 3D reconstruction results of a colored cartoon plaster target. Reproduced under CC-BY-4.0- [79] - https://opg.optica.org/oe/fulltext.cfm?uri=oe-29-10-15616&id=450808. **(A1)** Intensity image. **(A2)** 3D-recovered result without correction for the reflectance. **(A3)** 3D-recovered result with varying reflectance. (A4) the height variations in the pixels of **(A2)** and**(A3)**. **(B)** Experimental results of Cai et al. Reproduced under CC-BY-4.0- [80] - https://opg.optica.org/ao/fulltext.cfm?uri=ao-61-21-6228&id=479184. **(B1)** The RGB polarization sub-image captured at a direction of 0°. **(B2)** and **(B3)** ultimate results of proposed polarization 3D imaging without and with color removal theory, respectively.

To solve the problem of reconstruction distortion in polarization 3D imaging for non-uniform reflective targets, Li et al. [79] presented a near-infrared (NIR) monocular 3D computational polarization imaging method. They addressed the issue of varying intensity between the color patches due to the nonuniform reflectance of the colored target by using the reflected light feature in the near-infrared band. A normalized model of the near-infrared intensity gradient field was established in the non-uniform reflectivity region under monocular to solve the azimuthal ambiguity problem in polarization 3D imaging, and the reconstruction results are shown in Figure 14A.

Similar to Li's work, Cai et al. [80] also proposed a novel polarization 3D imaging technique to restore the 3D contours of multi-colored Lambertian objects, but the difference was that they adopted the chromaticity-based color removal theory. Based on the recovered intrinsic intensity, they solved the azimuthal ambiguity problem in a similar approach to achieve high-precision 3D reconstruction, and the results are shown in Figure 14B.

In addition to the influence of color, there are other interference factors in polarization 3D imaging. However, due to the page constraint, other problems and solutions will be briefly introduced as follows.

- Mixture of specular reflection and diffuse reflection

Since the polarization 3D imaging technology is developed based on the polarization characteristics of the reflected light for 3D reconstruction, the accuracy of polarization acquisition is the primary factor affecting the reconstruction accuracy. The mixture of specular and diffuse reflections is the most common challenge affecting polarization accuracy. Umeyama et al. [81] addressed the issue of inaccurate interpretation of polarization field information caused by the mixture of specular and diffuse reflected light. They proposed to exploit the difference of polarization intensity between specular reflection and diffuse reflection to analyze polariton images in different directions through independent component analysis for separating the diffuse and specular components of surface reflection. Shen et al. [82] proposed a method to separate diffuse and specular reflection components from a single image. They constructed a pseudo-chromatic space to classify image pixels and achieved fast and accurate specular reflection light removal without any local operation on the specular reflection information. Wang et al. [83] presented a global energy minimization specular reflection light removal method based on polarization characteristics to eliminate color distortions based on color intensity information for separating

specular reflection from diffuse reflection and reducing the error in the interpretation of polarization information. Li et al. [84] derived an analytical expression of diffuse reflection light under mixed optical fields and proposed a method to remove specular reflection light based on the analysis of the intensity and polarization field distribution characteristics, thus accurately interpreting polarization information.

- Gradient field integration

Gradient field integration directly affects the accuracy of reconstructing results, and it is another essential process of reconstructing the surface from the obtained normal vector of micro-plane elements. The Frankot-Chellappa method [85] is commonly used in the SfP method, but it takes the finite center difference as the differentiation operator, which has a large truncation error and does not constrain adjacent heights, to establish the difference-slope relationship, increasing reconstruction errors. To improve the integration accuracy, Ren et al. [86] proposed an improved higher-order finite-difference least-squares product method for circular regions and incomplete gradient data, which can handle incomplete gradient data more directly and efficiently. Qiao et al. [87] formulated Fourier-based exact integration square break error to increase the height and slope of the operator for obtaining higher 3D reconstruction accuracy. Smith et al. [88] presented an integration method for segmental fitting of gradient data using spline curves, which can obtain accurate integration results at surface boundaries.

- Absolute depth recovery

According to the analysis in Section 2, the depth information reconstructed by polarization 3D imaging is the result of integrating normal vectors of all pixel points, so it is the relative depth display in the pixel coordinate system. To recover the absolute depth of the 3D contour based on the polarization characteristics, Ping et al. [89] combined conventional polarization 3D imaging with binocular stereo vision. They calculated the coordinate transformation parameters between the image pixel coordinate system and the world coordinate system with the camera parameters obtained from binocular stereo calibration; then, the relative depth in the pixel coordinate system acquired by polarization was converted to the absolute depth in the world coordinate system by using the least-squares method. Some other methods [48, 49, 64] can also recover the absolute depth of the object 3D contour obtained from SfP.

4 Challenges

Although the polarization 3D imaging technology has been further developed and many SfP techniques have been proposed, its application to engineering, medical, industrial, and daily life fields is still a challenge. This section lists some challenging issues worth exploring to further advance the field of polarization 3D imaging.

4.1 Technical researches

As introduced in Section 3, many SfP techniques have been proposed, but there are many limitations in these techniques, preventing widespread commercial applications of the polarization 3D imaging technology. More challenges need to be overcome: 1) How to perform high-precision polarization 3D reconstruction under monocular conditions without the assistance of other detection methods? 2) Is it possible to realize real-time polarization 3D imaging? 3) How to apply SfP to more complex objects and scenes like multiple discrete objects? 4) How to solve the polarization dataset acquisition problem in deep learning? 5) How to improve polarization detection accuracy in visible, infrared, and other wavelengths? The above-mentioned problems are important research directions of future polarization 3D imaging technology, and some difficult points need to be addressed in practical applications of polarization 3D imaging.

4.2 Data storage

The record and storage follow the information acquisition process. However, the sizes of conventional 3D image files represented by OBJ, PLY, and STL are often one order of magnitude larger than those of 2D files represented by JPG, BMP, and TIF. How can one effectively store and deliver such a huge amount of 3D data is a key issue to applying polarization 3D imaging technology in practice. Although some efforts [90–94] have been made to compress 3D range data, none of these methods is proposed for efficient 3D data storage.

4.3 Hardware integration

Traditionally, the polarization characteristics of reflected light are obtained by rotating a polarizer placed in front of the camera, which is difficult to promote practically. Although some manufacturers like Sony has manufactured integrated devices such as polarization chips, the chips have shortcomings of weak extinction ratio, low resolution [95–97], etc. The integration of hardware facilities has gradually failed to keep up with the development of polarization 3D imaging technology. Therefore, efforts to promote hardware integration of polarization 3D imaging systems are highly needed to advance polarization 3D imaging technology.

4.4 Applications

The value of practical applications is a huge motivation for technological progress. However, though many SfP techniques have been developed, none of the existing techniques are effectively applied in practice. So far, many other 3D imaging techniques have been applied to practical fields. Binocular stereo vision is commonly used in fields of robotic vision [98–100]. Deep learning and lidar 3D imaging are widely used in

autonomous driving recently [101–105]. Structured light 3D imaging technology is mostly used in the detection of defects in precision industrial products because of its high accuracy [106–108]. Therefore, the specific application fields of polarization 3D imaging technology should be selected according to its technical advantages. For example, considering the long-distance detection advantage of polarization 3D imaging technology, 3D mapping of remote sensing may become a significant application field.

5 Summary

This paper presents a comprehensive and detailed review of the polarization 3D imaging mechanism and some classical SfP techniques. Especially, this paper focuses on the problems and solutions of normal ambiguity in polarization 3D imaging technology by explaining technical fundamentals, demonstrating experimental results, and analyzing capabilities/limitations. Besides, other problems and related techniques in polarization 3D imaging are also introduced. Finally, our perspectives on some remained challenges in polarization 3D imaging technology are summarized to inspire the readers.

Author contributions

XL, ZL, and YC prepared the references and data. JW was involved in the conception and design of the project. XL wrote the manuscript with input from all authors. XS reviewed and improved the writing. CP and JS supervised the project. All authors contributed to the manuscript revision, read, and approved the submitted version.

Funding

National Natural Science Foundation of China (NSFC) (62205256); China Postdoctoral Science Foundation (2022TQ0246); CAS Key Laboratory of Space Precision Measurement Technology (SPMT2023-02).

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

1. Sonka M, Hlavac V, Boyle R. *Image processing, analysis, and machine vision*. Boston, MA: Cengage Learning (2014).

2. Shao X-P, Liu F, Li W, Yang L, Yang S, Liu J. Latest progress in computational imaging technology and application. *Laser Optoelectronics Prog* (2020) 57(2):020001. doi:10.3788/lop57.020001

3. Shechtman Y, Eldar YC, Cohen O, Chapman HN, Miao J, Segev M. Phase retrieval with application to optical imaging: A contemporary Overview. *IEEE Signal Processing Magazine* (2015) 32(3):87–109. doi:10.1109/MSP.2014.2352673

4. Rybka T, Ludwig M, Schmalz MF, Knittel V, Brida D, Leitenstorfer A. Sub-cycle optical phase control of nanotunnelling in the single-electron regime. *Nat Photon* (2016) 10(10):667–70. doi:10.1038/nphoton.2016.174

5. Chen C, Gao L, Gao W, Ge C, Du X, Li Z, et al. Circularly polarized light detection using chiral hybrid perovskite. *Nat Commun* (2019) 10(1):1927. doi:10.1038/s41467-019-09942-z

6. Wang X, Cui Y, Li T, Lei M, Li J, Wei Z. Recent advances in the functional 2d photonic and optoelectronic devices. *Adv Opt Mater* (2019) 7(3):1801274. doi:10.1002/adom.201801274

7. Tahara T, Quan X, Otani R, Takaki Y, Matoba O. Digital holography and its multidimensional imaging applications: A review. *Microscopy* (2018) 67(2):55–67. doi:10.1093/jmicro/dfy007

8. Clark RA, Mentiplay BF, Hough E, Pua YH. Three-dimensional cameras and skeleton pose tracking for physical function assessment: A review of uses, validity, current developments and Kinect alternatives. *Gait & posture* (2019) 68:193–200. doi:10.1016/j.gaitpost.2018.11.029

9. Zhou S, Xiao S. 3d face recognition: A survey. *Human-centric Comput Inf Sci* (2018) 8(1):35–27. doi:10.1186/s13673-018-0157-2

10. Adjabi I, Ouahabi A, Benzaoui A, Taleb-Ahmed A. Past, present, and future of face recognition: A review. *Electronics* (2020) 9(8):1188. doi:10.3390/electronics9081188

11. Ben Abdallah H, Jovančević I, Orteu J-J, Brèthes L. Automatic inspection of aeronautical mechanical assemblies by matching the 3d cad model and real 2d images. *J Imaging* (2019) 5(10):81. doi:10.3390/jimaging5100081

12. Qian J, Feng S, Xu M, Tao T, Shang Y, Chen Q, et al. High-resolution real-time 360° 3d surface defect inspection with fringe projection profilometry. *Opt Lasers Eng* (2021) 137:106382. doi:10.1016/j.optlaseng.2020.106382

13. Pham QH, Sevestre P, Pahwa RS, Zhan H, Pang CH, Chen Y, et al. A 3d dataset: Towards autonomous driving in challenging environments. In: Proceeding of the 2020 IEEE International Conference on Robotics and Automation (ICRA); May 2020; Paris, France. IEEE (2020).

14. Li P, Zhao H, Liu P, Cao F. Rtm3d: Real-Time monocular 3d detection from object keypoints for autonomous driving. In: Computer Vision–ECCV 2020: 16th European Conference; August 23–28, 2020; Glasgow, UK. Springer (2020). Proceedings, Part III 16.

15. Xue Y, Cheng T, Xu X, Gao Z, Li Q, Liu X, et al. High-accuracy and real-time 3d positioning, tracking system for medical imaging applications based on 3d digital image correlation. *Opt Lasers Eng* (2017) 88:82–90. doi:10.1016/j.optlaseng.2016.07.002

16. Pan Z, Huang S, Su Y, Qiao M, Zhang Q. Strain field measurements over 3000 C using 3d-digital image correlation. *Opt Lasers Eng* (2020) 127:105942. doi:10.1016/j.optlaseng.2019.105942

17. Chen F, Brown GM, Song M. Overview of 3-D shape measurement using optical methods. *Opt Eng* (2000) 39(1):10–22. doi:10.1117/1.602438

18. Fang Y, Wang X, Sun Z, Zhang K, Su B. Study of the depth accuracy and entropy characteristics of a tof camera with coupled noise. *Opt Lasers Eng* (2020) 128:106001. doi:10.1016/j.optlaseng.2020.106001

19. Liu B, Yu Y, Jiang S. Review of advances in lidar detection and 3d imaging. *Opto-Electronic Eng* (2019) 46(7):190167. doi:10.12086/oee.2019.190167

20. Zuo C, Feng S, Huang L, Tao T, Yin W, Chen Q. Phase shifting algorithms for fringe projection profilometry: A review. *Opt lasers Eng* (2018) 109:23–59. doi:10.1016/j.optlaseng.2018.04.019

21. Zhu Z, Li M, Zhou F, You D. Stable 3d measurement method for high dynamic range surfaces based on fringe projection profilometry. *Opt Lasers Eng* (2023) 166:107542. doi:10.1016/j.optlaseng.2023.107542

22. Tippetts B, Lee DJ, Lillywhite K, Archibald J. Review of stereo vision algorithms and their suitability for resource-limited systems. *J Real-Time Image Process* (2016) 11: 5–25. doi:10.1007/s11554-012-0313-2

23. Shi S, Ding J, New TH, Soria J. Light-field camera-based 3d volumetric particle image velocimetry with dense ray tracing reconstruction technique. *Experiments in Fluids* (2017) 58:78–16. doi:10.1007/s00348-017-2365-3

24. Koshikawa K. A polarimetric approach to shape understanding of glossy objects. *Adv Robotics* (1979) 2(2):190.

25. Wallace AM, Liang B, Trucco E, Clark J. Improving depth image acquisition using polarized light. *Int J Comp Vis* (1999) 32(2):87–109. doi:10.1023/A:1008154415349

26. Saito M, Sato Y, Ikeuchi K, Kashiwagi H. Measurement of surface orientations of transparent objects by use of polarization in highlight. *JOSA A* (1999) 16(9):2286–93. doi:10.1364/JOSAA.16.002286

27. Müller V. Elimination of specular surface-reflectance using polarized and unpolarized light. In: Proceeding of the Computer Vision—ECCV'96: 4th European Conference on Computer Vision Cambridge; April 15–18, 1996; UK. Springer (1996). Proceedings Volume II 4.

28. Wolff LB, Boult TE. Constraining object features using a polarization reflectance model. *Phys Based Vis Princ Pract Radiom* (1993) 1:167.

29. Sato Y, Wheeler Ikeuchi, MD K. Object shape and reflectance modeling from observation. Proceedings of the 24th annual conference on Computer graphics and interactive techniques; 1997.

30. Atkinson GA, ER Hancock, . Recovery of surface orientation from diffuse polarization. *IEEE Trans image Process* (2006) 15(6):1653–64. doi:10.1109/TIP.2006.871114

31. Miyazaki, Kagesawa, Ikeuchi, Polarization-based transparent surface modeling from two views. *Proceedings Ninth IEEE International Conference on Computer Vision* (2003) 2: 1381–1386. doi:10.1109/ICCV.2003.1238651

32. Bass M. *Handbook of Optics: Volume I-geometrical and physical Optics, polarized light, components and instruments*. McGraw-Hill Education (2010).

33. Cui Z, Gu J, Shi B, Tan P, Kautz J. Polarimetric multi-view stereo. In: Proceedings of the IEEE conference on computer vision and pattern recognition; July 2017; Honolulu, HI, USA. IEEE (2017).

34. Li X, Liu F, Shao X-P. Research progress on polarization 3d imaging technology. *J Infrared Millimeter Waves* (2021) 40(2):248–62.

35. Monteiro M, Stari C, Cabeza C, Martí AC. The polarization of light and Malus' law using smartphones. *Phys Teach* (2017) 55(5):264–6. doi:10.1119/1.4981030

36. Partridge, M Saull, R editors. Three-dimensional surface reconstruction using emission polarization. *Image and signal processing for remote sensing II*. Paris, France: SPIE (1995).

37. Miyazaki D, Saito M, Sato Y, Ikeuchi K. Determining surface orientations of transparent objects based on polarization degrees in visible and infrared wavelengths. *JOSA A* (2002) 19(4):687–94. doi:10.1364/JOSAA.19.000687

38. Stolz C, Ferraton M, Meriaudeau F. Shape from polarization: A method for solving zenithal angle ambiguity. *Opt Lett* (2012) 37(20):4218–20. doi:10.1364/OL.37.004218

39. Jinglei H, Yongqiang Z, Haimeng Z, Brezany P, Jiayu S. 3d reconstruction of high-reflective and textureless targets based on multispectral polarization and machine vision. *Acta Geodaetica et Cartographica Sinica* (2018) 47(6):816. doi:10.11947/j.AGCS.2018.20170624

40. Zhao Y, Yi C, Kong SG, Pan Q, Cheng Y, Zhao Y, et al. *Multi-band polarization imaging*. Springer (2016).

41. Morel, O, Meriaudeau, F, Stolz, C, Gorria, P, editors. Polarization imaging applied to 3d reconstruction of specular metallic surfaces. *Machine vision applications in industrial inspection XIII*. San Jose, CA: SPIE (2005).

42. Morel O, Stolz C, Meriaudeau F, Gorria P. Active lighting applied to three-dimensional reconstruction of specular metallic surfaces by polarization imaging. *Appl Opt* (2006) 45(17):4062–8. doi:10.1364/AO.45.004062

43. Morel O, Ferraton M, Stolz C, Gorria P. Active lighting applied to shape from polarization. In: Proceeding of the 2006 International Conference on Image Processing; October 2006; Atlanta, GA, USA. IEEE (2006).

44. Miyazaki D, Shigetomi T, Baba M, Furukawa R, Hiura S, Asada N. Surface normal estimation of black specular objects from multiview polarization images. *Opt Eng* (2017) 56(4):041303. doi:10.1117/1.OE.56.4.041303

45. Atkinson GA, Hancock ER. Surface reconstruction using polarization and photometric stereo. In: Computer Analysis of Images and Patterns: 12th International Conference, CAIP 2007; August 27-29, 2007; Vienna, Austria, 12. Springer (2007).

46. Mahmoud AH, El-Melegy MT, Farag AA. Direct method for shape recovery from polarization and shading. In: Proceeding of the 2012 19th IEEE International Conference on Image Processing; September 2012; Orlando, FL, USA. IEEE (2012).

47. Kadambi A, Taamazyan V, Shi B, Raskar R. Depth sensing using geometrically constrained polarization normals. *Int J Comp Vis* (2017) 125:34–51. doi:10.1007/s11263-017-1025-7

48. Tian X, Liu R, Wang Z, Ma J. High quality 3d reconstruction based on fusion of polarization imaging and binocular stereo vision. *Inf Fusion* (2022) 77:19–28. doi:10.1016/j.inffus.2021.07.002

49. Liu R, Liang H, Wang Z, Ma J, Tian X. Fusion-based high-quality polarization 3d reconstruction. *Opt Lasers Eng* (2023) 162:107397. doi:10.1016/j.optlaseng.2022.107397

50. Ba Y, Gilbert A, Wang F, Yang J, Chen R, Wang Y, et al. Deep shape from polarization. In: Proceeding of the Computer Vision–ECCV 2020: 16th European Conference; August 23–28, 2020; Glasgow, UK. Springer (2020).*Proceedings, Part XXIV 16*

51. Lei C, Qi C, Xie J, Fan N, Koltun V, Chen V. Shape from polarization for complex scenes in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (2022).

52. Han, P, Li, X, Liu, F, Cai, Y, Yang, K, Yan, M, et al. *Accurate passive 3d polarization face reconstruction under complex conditions assisted with deep learning. Photonics*. Basel, Switzerland: Multidisciplinary Digital Publishing Institute (2022).

53. Born M, Wolf E. *principles of Optics, 7th (expanded) edition*. 461. United Kingdom: Press Syndicate of the University of Cambridge (1999). p. 401–24.

54. Wu C. Towards linear-time incremental structure from motion. In: Proceeding of the 2013 International Conference on 3D Vision-3DV 2013; June 2013; Seattle, WA, USA. IEEE (2013).

55. Fuhrmann, S, Langguth, F, Goesele, M, editors. *Mve-a multi-view reconstruction environment*. Eindhoven, Netherlands: GCH (2014).

56. Galliani S, Lasinger K, Schindler K. Massively parallel multiview stereopsis by surface normal diffusion. In: Proceedings of the IEEE International Conference on Computer Vision; December 2015; Santiago, Chile. IEEE (2015).

57. Kadambi A, Taamazyan V, Shi B, Raskar R. Polarized 3d: High-quality depth sensing with polarization cues. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE (2015).

58. Linear depth estimation from an uncalibrated, monocular polarisation image. In: Smith WA, Ramamoorthi R, Tozza S, editors. Proceedings of the Computer Vision–ECCV 2016: 14th European Conference; October 11-14, 2016; Amsterdam, The Netherlands. Springer (2016). Proceedings, Part VIII 14.

59. Zhou Z, Wu Z, Tan P. Multi-view photometric stereo with spatially varying isotropic materials. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; June 2013; Portland, OR, USA. IEEE (2013).

60. Wei ZM, Xia JB. Recent progress in polarization-sensitive photodetectors based on low-dimensional semiconductors. *Acta Physica Sinica -Chinese Edition* (2019) 68(16): 163201. doi:10.7498/aps.68.20191002

61. Zhang C, Hu J, Dong Y, Zeng A, Huang H, Wang C. High efficiency all-dielectric pixelated metasurface for near-infrared full-Stokes polarization detection. *Photon Res* (2021) 9(4):583–9. doi:10.1364/PRJ.415342

62. Shang X, Wan L, Wang L, Gao F, Li H. Emerging materials for circularly polarized light detection. *J Mater Chem C* (2022) 10(7):2400–10. doi:10.1039/D1TC04163K

63. Shakeri M, Loo SY, Zhang H, Hu K. Polarimetric monocular dense mapping using relative deep depth prior. *IEEE Robotics Automation Lett* (2021) 6(3):4512–9. doi:10.1109/LRA.2021.3068669

64. Tan Z, Zhao B, Ji Y, Xu X, Kong Z, Liu T, et al. A welding seam positioning method based on polarization 3d reconstruction and linear structured light imaging. *Opt Laser Tech* (2022) 151:108046. doi:10.1016/j.optlastec.2022.108046

65. Nguyen LT, Kim J, Shim B. Low-rank matrix completion: A contemporary survey. *IEEE Access* (2019) 7:94215–37. doi:10.1109/ACCESS.2019.2928130

66. Kovesi P. Shapelets correlated with surface normals produce surfaces. In: Proceeding of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1; October 2005; Beijing, China. IEEE (2005).

67. Smith WA, Ramamoorthi R, Tozza S. Height-from-Polarisation with unknown lighting or albedo. *IEEE Trans pattern Anal machine intelligence* (2018) 41(12):2875–88. doi:10.1109/TPAMI.2018.2868065

68. Liu R, Tian X, Li S. Polarisation-modulated photon-counting 3d imaging based on a negative parabolic pulse model. *Opt Express* (2021) 29(13):20577–89. doi:10.1364/OE.427997

69. Poggi M, Tosi F, Batsos K, Mordohai P, Mattoccia S. On the synergies between machine learning and binocular stereo for depth estimation from images: A survey. *IEEE Trans Pattern Anal Machine Intelligence* (2021) 44(9):5314–34. doi:10.1109/TPAMI.2021.3070917

70. Hu Y, Zhen W, Scherer S. Deep-learning assisted high-resolution binocular stereo depth reconstruction. In: Proceeding of the 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE (2020).

71. Chen G, Han K, Shi B, Matsushita Y, Wong K-YK. Deep photometric stereo for non-lambertian surfaces. *IEEE Trans Pattern Anal Machine Intelligence* (2020) 44(1): 129–42. doi:10.1109/TPAMI.2020.3005397

72. Van der Jeught S, Dirckx JJ. Deep neural networks for single shot structured light profilometry. *Opt express* (2019) 27(12):17091–101. doi:10.1364/OE.27.017091

73. Yang G, Wang Y. Three-dimensional measurement of precise shaft parts based on line structured light and deep learning. *Measurement* (2022) 191:110837. doi:10.1016/j.measurement.2022.110837

74. Zou S, Zuo X, Qian Y, Wang S, Xu C, Gong M, et al. 3d human shape reconstruction from a polarization image. In: Proceeding of the Computer Vision–ECCV 2020: 16th European Conference; August 23–28, 2020; Glasgow, UK. Springer (2020). Proceedings, Part XIV 16.

75. Deschaintre V, Lin Y, Ghosh A. Deep polarization imaging for 3d shape and svbrdf acquisition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (2021).

76. Li Z, Xu Z, Ramamoorthi R, Sunkavalli K, Chandraker M. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Trans Graphics (Tog)* (2018) 37(6):1–11. doi:10.1145/3272127.3275055

77. Miyazaki D, Tan RT, Hara K, Ikeuchi K. Polarization-based inverse rendering from a single view. In: Proceeding of the Computer Vision, IEEE International Conference on; October 2003; Nice, France. IEEE Computer Society (2003).

78. Kondo Y, Ono T, Sun L, Hirasawa Y, Murayama J. Accurate polarimetric brdf for real polarization scene rendering. In: Proceeding of the Computer Vision–ECCV 2020: 16th European Conference; August 23–28, 2020; Glasgow, UK. Springer (2020). Proceedings, Part XIX 16.

79. Li X, Liu F, Han P, Zhang S, Shao X. Near-infrared monocular 3d computational polarization imaging of surfaces exhibiting nonuniform reflectance. *Opt Express* (2021) 29(10):15616–30. doi:10.1364/OE.423790

80. Cai Y, Liu F, Shao X, Cai G. Impact of color on polarization-based 3d imaging and countermeasures. *Appl Opt* (2022) 61(21):6228–33. doi:10.1364/AO.462778

81. Umeyama S, Godin G. Separation of diffuse and specular components of surface reflection by use of polarization and statistical analysis of images. *IEEE Trans Pattern Anal Machine Intelligence* (2004) 26(5):639–47. doi:10.1109/TPAMI.2004.1273960

82. Shen H-L, Zheng Z-H. Real-time highlight removal using intensity ratio. *Appl Opt* (2013) 52(19):4483–93. doi:10.1364/AO.52.004483

83. Wang F, Ainouz S, Petitjean C, A Bensrhair,. Specularity removal: A global energy minimization approach based on polarization imaging. *Computer Vis Image Understanding* (2017) 158:31–9. doi:10.1016/j.cviu.2017.03.003

84. Li, X, Liu, F, Cai, Y, Huang, S, Han, P, Shao, X, editors. Polarization 3d imaging having highlighted areas. *Frontiers in Optics*. Washington D.C: Optica Publishing Group (2019).

85. Frankot RT, Chellappa R. A method for enforcing integrability in shape from shading algorithms. *IEEE Trans pattern Anal machine intelligence* (1988) 10(4):439–51. doi:10.1109/34.3909

86. Ren H, Gao F, Jiang X. Improvement of high-order least-squares integration method for stereo deflectometry. *Appl Opt* (2015) 54(34):10249–55. doi:10.1364/AO.54.010249

87. Sun Z, Qiao Y, Jiang Z, Xu X, Zhou J, Gong X. An accurate fourier-based method for three-dimensional reconstruction of transparent surfaces in the shape-from-polarization method. *IEEE Access* (2020) 8:42097–110. doi:10.1109/ACCESS.2020.2977097

88. Smith GA. 2d zonal integration with unordered data. *Appl Opt* (2021) 60(16):4662–7. doi:10.1364/AO.426162

89. Ping XX, Liu Y, Dong XM, Zhao YQ, Zhang-Yan,University NP. 3-D reconstruction of textureless and high-reflective target by polarization and binocular stereo vision. *J Infrared Millimeter Waves* (2017). doi:10.11972/j.issn.1001-9014.2017.04.009

90. Karpinsky N, Zhang S. 3d range geometry video compression with the H 264 codec. *Opt Lasers Eng* (2013) 51(5):620–5. doi:10.1016/j.optlaseng.2012.12.021

91. Wang Y, Zhang L, Yang S, Ji F. Two-Channel high-accuracy holoimage technique for three-dimensional data compression. *Opt Lasers Eng* (2016) 85:48–52. doi:10.1016/j.optlaseng.2016.04.020

92. Li S, Jiang T, Tian Y, Huang T. 3d human skeleton data compression for action recognition. In: Proceeding of the 2019 IEEE Visual Communications and Image Processing (VCIP); December 2019; Sydney, NSW, Australia. IEEE (2019).

93. Wang J, Ding D, Li Z, Ma Z. Multiscale point cloud geometry compression. In: Proceeding of the 2021 Data Compression Conference (DCC); March 2021; Snowbird, UT, USA. IEEE (2021).

94. Gu S, Hou J, Zeng H, Yuan H, Ma K-K. 3d point cloud attribute compression using geometry-guided sparse representation. *IEEE Trans Image Process* (2019) 29:796–808. doi:10.1109/TIP.2019.2936738

95. Maruyama Y, Terada T, Yamazaki T, Uesaka Y, Nakamura M, Matoba Y, et al. 3.2-MP Back-Illuminated Polarization Image Sensor with Four-Directional Air-Gap Wire Grid and 2.5-<inline-formula> <tex-math notation=LaTeX>μ </tex-math> </inline-formula>m Pixels. *IEEE Trans Electron Devices* (2018) 65:2544–51. doi:10.1109/TED.2018.2829190

96. Ren H, Yang J, Liu X, Huang P, Guo L. Sensor modeling and calibration method based on extinction ratio error for camera-based polarization navigation sensor. *Sensors* (2020) 20:3779. doi:10.3390/s20133779

97. Lane C, Rode D, Rösgen T. Calibration of a polarization image sensor and investigation of influencing factors. *Appl Opt* (2022) 61:C37–C45. doi:10.1364/AO.437391

98. Wang C, Zou X, Tang Y, Luo L, Feng W. Localisation of litchi in an unstructured environment using binocular stereo vision. *Biosyst Eng* (2016) 145:39–51. doi:10.1016/j.biosystemseng.2016.02.004

99. Zhai G, Zhang W, Hu W, Ji Z. Coal mine rescue robots based on binocular vision: A review of the state of the art. *IEEE Access* (2020) 8:130561–75. doi:10.1109/ACCESS.2020.3009387

100. Kim W-S, Lee D-H, Kim Y-J, Kim T, Lee W-S, Choi C-H. Stereo-vision-based crop height estimation for agricultural robots. *Comput Elect Agric* (2021) 181:105937. doi:10.1016/j.compag.2020.105937

101. Fujiyoshi H, Hirakawa T, Yamashita T. Deep learning-based image recognition for autonomous driving. *IATSS Res* (2019) 43(4):244–52. doi:10.1016/j.iatssr.2019.11.008

102. Grigorescu S, Trasnea B, Cocias T, Macesanu G. A survey of deep learning techniques for autonomous driving. *J Field Robotics* (2020) 37(3):362–86. doi:10.1002/rob.21918

103. Cui Y, Chen R, Chu W, Chen L, Tian D, Li Y, et al. Deep learning for image and point cloud fusion in autonomous driving: A review. *IEEE Trans Intell Transportation Syst* (2021) 23(2):722–39. doi:10.1109/TITS.2020.3023541

104. Royo S, Ballesta-Garcia M. An Overview of lidar imaging systems for autonomous vehicles. *Appl Sci* (2019) 9(19):4093. doi:10.3390/app9194093

105. Zhao X, Sun P, Xu Z, Min H, Yu H. Fusion of 3d lidar and camera data for object detection in autonomous vehicle applications. *IEEE Sensors J* (2020) 20(9):4901–13. doi:10.1109/JSEN.2020.2966034

106. Chu H-H, Wang Z-Y. A vision-based system for post-welding quality measurement and defect detection. *Int J Adv Manufacturing Tech* (2016) 86:3007–14. doi:10.1007/s00170-015-8334-1

107. Yang L, Li E, Long T, Fan J, Liang Z. A novel 3-D path extraction method for arc welding robot based on stereo structured light sensor. *IEEE Sensors J* (2018) 19(2):763–73. doi:10.1109/JSEN.2018.2877976

108. Ma Y, Fan J, Deng S, Luo Y, Ma X, Jing F, et al. Efficient and accurate start point guiding and seam tracking method for curve weld based on structure light. *IEEE Trans Instrumentation Meas* (2021) 70:1–10. doi:10.1109/TIM.2021.3072103

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 18 May 2023
DOI 10.3389/fphy.2023.1196067

OPEN ACCESS

EDITED BY
Xukun Yin,
Xidian University, China

REVIEWED BY
Baoquan Jin,
Taiyuan University of Technology, China
Xiang Zhong,
Hefei University of Technology, China
Sheng Liang,
Beijing Jiaotong University, China

*CORRESPONDENCE
Sufan Yang,
yangsufan@buaa.edu.cn

RECEIVED 29 March 2023
ACCEPTED 10 May 2023
PUBLISHED 18 May 2023

CITATION
Zhang C, Yang S and Wang X (2023), Dual pulse heterodyne distributed acoustic sensor system employing SOA-based fiber ring laser.
Front. Phys. 11:1196067.
doi: 10.3389/fphy.2023.1196067

Dual pulse heterodyne distributed acoustic sensor system employing SOA-based fiber ring laser

Chunxi Zhang[1,2], Sufan Yang[1,2,3]* and Xiaxiao Wang[1,2]

[1]School of Instrumentation Science and Opto-Electronics Engineering, Beihang University, Beijing, China, [2]Institute of Optics and Electronics Technologies, Beihang University, Beijing, China, [3]Shenyuan Honors College, Beihang University, Beijing, China

Distributed Acoustic Sensor (DAS) has potential in applications such as hydroacoustic detection. In this paper, a dual-pulse heterodyne distributed acoustic sensor (DAS) system using a semiconductor optical amplifier (SOA)-based fiber ring laser (FRL) is proposed. Unlike the previous DAS system configurations, the SOA-based FRL replaces the narrow linewidth laser (NLL) and pulse modulator, reducing costs and simplifying the system. The system is demonstrated theoretically and validated experimentally. The adaptability of the SOA-based FRL in the heterodyne DAS system has been demonstrated in the experiments. Using the dual-pulse heterodyne detection method, the sensor system responds well to distributed acoustic detection and achieves accurate demodulation and positioning. A high signal-to-noise ratio (SNR) of 42.51 dB at 3 kHz is demonstrated as a demodulation result. The system's frequency range is 5 Hz to 5 kHz with a spatial resolution of 12 m. The proposed approach shows a broad application prospect for low-cost, large-scale, high-SNR distributed acoustic detection in maritime surveillance.

KEYWORDS

fiber optical sensor (FOS), distributed acoustic sensor (DAS), rayleigh backscattering (RBS), sound wave detection, heterodyne demodulation, semiconductor optical amplifier (SOA), fiber ring laser (FRL)

1 Introduction

Marine microseismic/acoustic monitoring is crucial for exploiting marine mineral resources and national underwater military defense [1, 2]. DAS based on phase-sensitive optical time-domain reflectometry (Φ-OTDR) has recently attracted attention in applications such as hydroacoustic detection [3, 4]. Φ-OTDR, a promising technique for DAS, has demonstrated its ability to provide real-time measurement reaching into the acoustic range [5–7]. In particular, Φ-OTDR has attracted increasing interest in hydroacoustic detection research due to its unique performance, including long sensing range, high spatial resolution, and wide dynamic range.

The DAS introduces a phase demodulation part to obtain the phase change caused by external acoustic signals compared to the traditional Φ-OTDR. Thus, several phase demodulation schemes have been developed. Generally, these schemes can be divided into homodyne [8, 9] and heterodyne [10, 11] according to different signaling methods. A heterodyne DAS using dual pulses was proposed. The system can simultaneously detect the actual waveforms of multiple vibration events with a high signal-to-noise ratio (SNR) [11]. Recently, researchers have been focusing on improving the performance of DAS to meet the requirements of practical applications. To accurately perform high-precision vibration

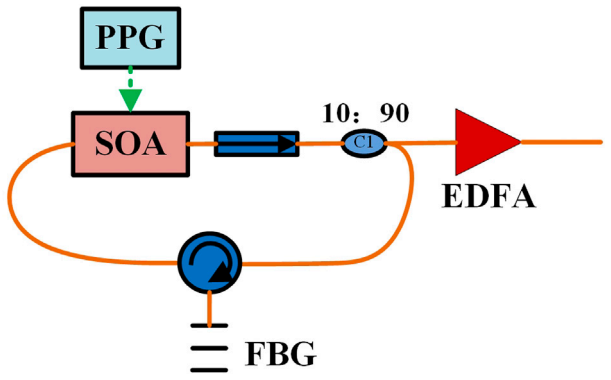


FIGURE 1
Schematic diagram of the SOA-based FRL.

identification based on high SNR acoustic signals, the system must have a high SNR level [12, 13]. To improve the SNR, signal processing techniques, such as wavelet denoising [14–18], filtering algorithms [19, 20], and pulse coding methods [21–23], are employed in the system. Aside from approaches based on signal processing, those based on optical techniques used a high-power laser [24, 25], high-performance photodetectors (PDs) [26], or optical amplifiers [27–29].

For interrogation purposes, a narrow linewidth laser (NLL) combined with a pulse modulator is typically used in DAS [30]. The pulse modulator converts the continuous light from the NLL into pulsed light. The non-ideal switching characteristics of the modulator introduce noise into the system, a significant source of the noise. Therefore, the modulator's extinction ratio (ER) is an intrinsic limiting factor that directly impacts the localization errors and reduces the SNR of Φ-OTDR [31, 32]. SOA provides an alternative idea for research due to its high ER and gains characteristics. A high visibility Φ-OTDR has been demonstrated for high-frequency vibration measurements. This was achieved by using a SOA as a modulator. This approach relied on the SOA to reduce coherent noise [5]. Furthermore, by controlling the carrier in the SOA, the SNR of the Φ-OTDR was improved in a study. A forced carrier recombination method has been proposed to improve the ER of a SOA. And experiments demonstrated 9 dB ER and 5.2 dB SNR improvement [33]. Still, the SNR is relatively low.

A dual heterodyne pulse DAS system employing a SOA-based FRL is proposed in this paper. Unlike previous DAS configurations, the SOA-based FRL operates as a pulsed mode-locked laser. The device replaces the NLL and pulse modulator, simplifying the system and reducing costs. The proposed scheme is demonstrated theoretically and experimentally. The SOA is a key component in the device, used as a gain medium in the cavity [34] and an optical pulse generator. A dual pulse heterodyne DAS system using the device is presented to verify its adaptability. The proposed system responds well to distributed acoustic signals, which achieves accurate demodulation and localization. A high SNR of 42.51 dB demodulation at 3 kHz is demonstrated in experiments. The SNR improved by more than one order of magnitude compared with the works mentioned above, such as [18], and [18, 33]. The system can recover acoustic signals with a frequency range from 5 Hz to 5 kHz with a spatial resolution of 12 m. The advantages of a simple and small structure, reliable multi-wavelength operation at room temperature, and a compact design offer potential in DAS applications.

2 Working principle and system structure

2.1 Basic principle for SOA-based FRL

Figure 1 shows the schematic of the proposed SOA-based FRL, which comprises a SOA, an FBG working in reflection mode, an optical isolator, and an optical circulator. In the device, the SOA acts as a gain medium [34] and an in-loop pulse modulator in the cavity. It is switched by a programmable pulse generator (PPG). Thus, the SOA yields a pulsed output with a repetition rate equal to the applied modulation frequency. The addition of the FBG modifies the intra-cavity loss profile. The spectral shift of the reflected light from the FBG changes as a result of exchange changes, which can be dynamically observed in the output light intensity of the filter. The FBG is also a wavelength-selective filter [30], which filters the broadband pulsed light emitted from the SOA within the bandwidth of the FBG, as shown in Figure 2. Most reflected light circulates continuously within the cavity to ensure a stable coupled output. Optical isolators are used to maintain the unidirectional nature of the cavity and to protect the SOA, which is connected to a 10:90 single mode (SM) coupler and an erbium-doped fiber amplifier (EDFA). The EDFA amplifies the output pulsed light from the SM coupler.

The SOA-based FRL operates as a switchable pulsed mode-locked laser [35], which works fundamentally differently than traditional continuous wave (CW) or mode-locked lasers. The cavity dynamics are much more complicated than either CW or mode-locked lasers because of the multiple physical effects involved. To date, a great deal of theoretical modeling and investigation of the laser dynamics in the cavities of the devices has been demonstrated [36–38]. The basic principle of the device is resonating frequency is related to the transmitted light's round-trip time, which is given by:

$$f_{cav} = \frac{v_g}{L_0 + 2L_{FBG}} \tag{1}$$

where v_g is the group velocity, and L_0 is the fixed fiber length corresponding to the length of the ring. L_{FBG} is the fiber length between the circulator and the FBG.

It is worth noting that using the SOA and FBG in the ring cavity aims to achieve a low side-mode suppression ratio (SMSR) laser [35], which shows a high stability in theory. Two main factors affect the stability of lasers: one is relaxation oscillation, and the other is mode competition [39]. Since the SOA's carrier recovery time is significantly shorter than the photon rise time in the cavity [40], the small fluctuations in optical power at the stimulated wavelength caused by external environmental disturbances decay rapidly with time. Thus, no relaxation oscillation is generated. The only factor that significantly affects the stability of the device is mode competition. As the fiber medium in fiber lasers cannot effectively compress the side modes [41], the side mode oscillation will compete with the main mode. When the mode

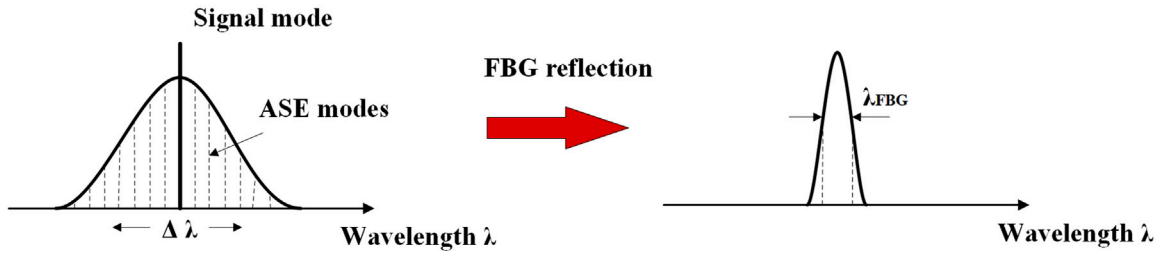


FIGURE 2
Operation principle of the SOA-based FRL.

hopping occurs, the SNR of the fiber laser output decreases due to interference between the two modes [42]. However, in the case of a fiber laser as a gain medium, the side-mode oscillation is suppressed by the gain saturation effect as long as it is in the gain saturation state [43]. Then, the laser output power becomes flat, and the SNR is improved. The side mode suppression effect is due to the nonlinear effect of the SOA. The nonlinear compression effect in the proposed SOA-based FRL is derived in detail below. The main mode and one of the side modes of the FRL are assumed to have the same small signal gain. At the input of the main mode, the main mode is $E_d = E \cdot \cos \omega_0 t$, the side mode is written as $E_s = \delta_i \cdot E \cdot \cos[(\omega_0 + \Delta\omega)t + \phi]$, and the SMSR is $\delta_i \ll 1$. Then, the SOA's optical input power is obtained as follows:

$$P_i = \frac{1}{2}(1 + \delta_i^2)E^2 + \delta_i E^2 \cos(\Delta\omega t + \phi) = \overline{P_i} + \Delta P_i(t) \tag{2}$$

The relative change in input power is defined as follows:

$$R_i = \frac{|\Delta P_i|_{max}}{\overline{P_i}} = \frac{2\delta_i}{1 + \delta_i^2} \tag{3}$$

Similarly, the relative changes of the side mode extinction ratio and power of SOA output can be derived from the small signal dynamic equation. The integral gain of SOA can be expressed as follows [44]:

$$h(t) = \int_0^l g(z,t)dz = \bar{h} + \Delta h(t) \tag{4}$$

where $\Delta h(t)$ is the time-varying term due to the input optical power, which is determined by the following equation:

$$\tau_c \cdot \frac{d\Delta h(t)}{dt} + \left(1 + \frac{\overline{P_i} \cdot e^{\bar{h}}}{P_{sat}}\right)\Delta h(t) = -\frac{\Delta P_t(t)}{P_{sat}}\left(e^{\bar{h}} - 1\right) \tag{5}$$

where τ_c is the carrier lifetime. P_{sat} is saturated power of the SOA. Define the saturation coefficient to characterize the depth of saturation as follows:

$$S = \frac{\overline{P_i} \cdot e^{\bar{h}}}{P_{sat}} \tag{6}$$

Then $\Delta h(t)$ can be obtained as follows:

$$\Delta h(t) = \frac{-SR_i(1 - e^{\bar{h}})}{(1+S)^2 + (\tau_c\Delta\omega)^2} \cdot [(1+S) \cdot \cos(\Delta\omega t + \phi) + \tau_c \cdot \sin(\Delta\omega t + \phi)] \tag{7}$$

The output power of the SOA can be written as follows:

$$P_o = P_i \cdot e^h \approx \overline{P_i} \cdot e^{\bar{h}} + \Delta P_i(t) \cdot e^{\bar{h}} + \overline{P_i} \cdot e^{\bar{h}} \cdot \Delta h(t) \tag{8}$$

Ignoring the higher-order small signals in Eq. 8, the expression for the output power can be derived as follows:

$$\begin{aligned}\Delta P_o(t) &\approx \Delta P_o(t) \cdot e^{\bar{h}} + \overline{P_i} \cdot e^{\bar{h}} \cdot \Delta h(t) \\ &= \overline{P_i} R_i e^{\bar{h}}\left[\left(1 - \frac{(1+S)S(1-e^{\bar{h}})}{(1+S)^2 + (\tau_c\Delta\omega)^2}\right)\right. \\ &\left.\cos(\Delta\omega t + \phi) - \left(1 - \frac{S(1-e^{\bar{h}})\tau_c\Delta\omega}{(1+S)^2 + (\tau_c\Delta\omega)^2}\right)\sin(\Delta\omega t + \phi)\right]\end{aligned} \tag{9}$$

The relative change in output power is

$$R_o = \frac{|\Delta P_i|_{max}}{\overline{P_i}} = \sqrt{1 - \frac{(1+S)^2 A(2-A)}{(1+S)^2 + (\tau_c\Delta\omega)^2}} \cdot R_i \tag{10}$$

where $A = \frac{S(1-e^{\bar{h}})}{(1+S)}$ for $0 < A < 1$, then it can be derived that $R_o < R_i$. When $\delta < 1$, R decreases monotonically as δ decreases, so $\delta_0 < \delta_i$. It indicates that the side mode has a smaller effective gain than the main mode. Although the side and main modes have the same small signal gain, the oscillations are suppressed by the smaller effective gain of the side mode.

Considering $\delta_i \ll 1$, then

$$\frac{R_o}{R_i} = \frac{\delta_o}{\delta_i} \cdot \frac{1 + \delta_i^2}{1 + \delta_o^2} \approx \frac{\delta_o}{\delta_i} \tag{11}$$

The ratio of the effective gain of the side mode to the main mode can be obtained from the above equation as follows:

$$\eta(\Delta\omega) = \frac{G_s}{G_d} = \left(\frac{\delta_o}{\delta_i}\right)^2 \approx \left(\frac{R_o}{R_i}\right)^2 = 1 - \frac{(1+S)^2 A(2-A)}{(1+S)^2 + (\tau_c\Delta\omega)^2} \tag{12}$$

The equation above suggests that η decreases as $\Delta\omega$ decreases. Therefore, the closer the side mode is to the main mode, the stronger the suppression effect is. The side mode farther away from the main mode has a weaker non-linear compression effect. However, the small signal gain of it is also relatively lower. Additionally, the FBG, as a frequency-selective element, makes the side mode farther away a significant loss. Thus, these frequencies far from the main mode can also be suppressed by the non-linear effect. When $\Delta\omega$ is fixed, if the average input power is high, then S and A are also relatively high, then η is low. In other words, the side modes are effectively suppressed when the SOA is

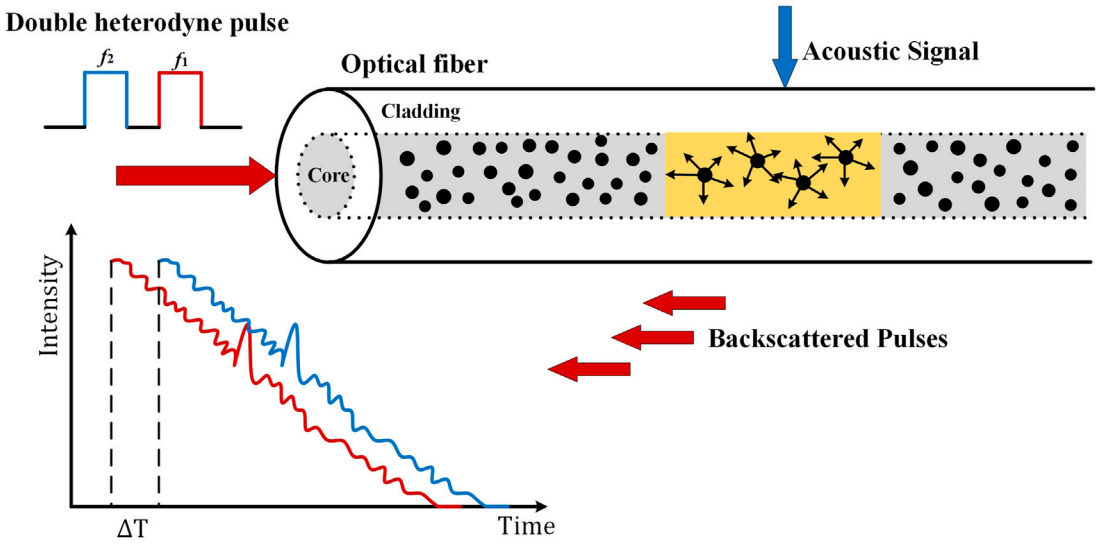


FIGURE 3
Operational diagram of the dual-pulse heterodyne Φ-OTDR employing SOA-based FRL.

in deep saturation. Thus, the device is of high stability in theory due to low SMSR output.

2.2 Principle of dual pulse heterodyne detection

The working principle of Φ-OTDRs is the interference effect of Rayleigh backscattering (RBS) light in optical fibers [30]. When subjected to external disturbances, the scattered light changes during its transmission through the fiber (intensity, phase, etc.). By detecting changes in these properties, scattered light can be identified. Observation of changes in the scattered light provides information about external disturbances.

In contrast to a conventional Φ-OTDR system, the probe pulses used in the proposed method are a pair of pulses. The principle of operation based on dual pulse heterodyne detection is shown in Figure 3. The receiver is disturbed by two RBS from the pair of probe pulses. This causes a heterodyne signal to be generated at the receiver. The acoustic signal can be obtained with a high SNR using a suitable demodulation method with OTDR. The waveform in the time domain and the signal's frequency spectrum can be effectively recovered.

Discrete the optical fiber intervals of ΔL for each probe pulse [45]at the point $Z_m = m\Delta L$. The RBS within a pulse width can be expressed as follows:

$$\boldsymbol{E}(\boldsymbol{Z_m}) = \boldsymbol{E_0}\sum_{k=m}^{m+N-1}\left(\boldsymbol{\gamma_k\sigma_k e^{j\theta_k} e^{j\varphi_k} e^{-\alpha k\Delta L}}\right) \tag{13}$$

where E_0 is the amplitude of the incident light, and α is the fiber attenuation coefficient. γ_k, σ_k and θ_k denote the polarization attenuation coefficient, Rayleigh scattering cross section and phase delay of the pulse through the point, respectively. φ_k is the phase change caused by the acoustic signal at the point. N is the total number of slices through which a single pulse passes. All slices within the pulse width have the same properties. Then, γ_k, σ_k and θ_k can be considered as constants, and the vibration point spacing is related to ΔL. The above equation can be simplified as:

$$\boldsymbol{E}(\boldsymbol{Z_m}) = \boldsymbol{S_m}\cdot\boldsymbol{e^{j\phi_m}} \tag{14}$$

where $S_m = E_0 e^{-\alpha m\Delta L}\gamma_m\sigma_m e^{j\theta_m}$, and ϕ_m denotes the total phase change caused by the vibration at that point. As long as the two pulse widths are long enough, the total expression of the detected RBS can be expressed as:

$$\boldsymbol{E}(\boldsymbol{Z_m}) = \boldsymbol{S_m}\cdot\boldsymbol{e^{j\phi_m}}\cdot\boldsymbol{e^{j\left(2\pi f_1 t+\varphi_1\right)}}\cdot\boldsymbol{S_{m-N_d}}\cdot\boldsymbol{e^{j\phi_{m-N_d}}}\cdot\boldsymbol{e^{j\left(2\pi f_2 t+\varphi_2\right)}} \tag{15}$$

where φ_1 and φ_2 are the initial phase of the dual pulses. Then the AC component of the interference signal can be expressed as:

$$\boldsymbol{I_s}(\boldsymbol{Z_m}) = \boldsymbol{S_m S_{m-N_d}}\cos\left[\boldsymbol{2\pi\Delta f t + \Phi(t) + \Delta\varphi_0}\right] \tag{16}$$

where Δf is the heterodyne frequency mentioned above. $\boldsymbol{\Phi(t)} = \boldsymbol{\phi_m - \phi_{m-N_d}}$ is phase change caused by the acoustic signal, and $\boldsymbol{\Delta\varphi_0 = \varphi_1 - \varphi_2}$.

Then, the output of the system based on the equation above can be simplified as follows:

$$\boldsymbol{I_s(t)} = \boldsymbol{A}\cos\left[\boldsymbol{2\pi\Delta f t + \Phi(t) + \Delta\varphi_0}\right] \tag{17}$$

where $\boldsymbol{A} = |\boldsymbol{S_m S_{m-N_d}}|$ is the intensity amplitude of the interferometric light, and $2\pi\Delta f t$ is the heterodyne carrier term. $\Phi(t)$ is the phase change term caused by the external vibration; $\Delta\varphi_0$ is the initial phase noise of the dual pulses.

Φ (t) can be demodulated by an in-phase/quadrature (IQ) phase demodulation algorithm [46], as shown in Figure 4. The output signal needs to be mixed with $\cos(2\pi\Delta f t)$ and $\sin(2\pi\Delta f t)$ respectively. Then, the mixed frequency terms will be low-pass filtered, giving:

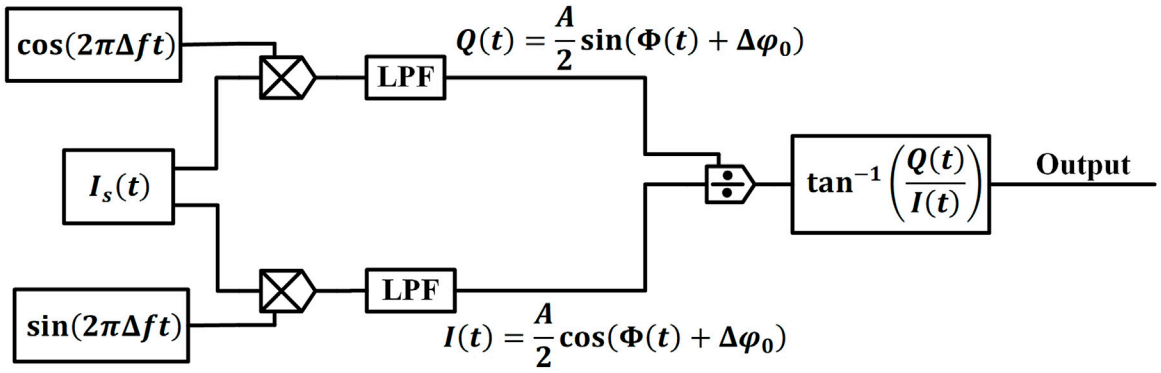


FIGURE 4
Operational diagram of the dual-pulse heterodyne detection.

$$I(t) = \frac{A}{2}\cos\left(\Phi(t) + \Delta\varphi_0\right) \tag{18}$$

$$Q(t) = \frac{A}{2}\sin\left(\Phi(t) + \Delta\varphi_0\right) \tag{19}$$

The intensity and phase information of RBS can be obtained by solving the summation and arctangent of the above two signals as follows:

$$A = \sqrt{I^2(t) + Q^2(t)} \tag{20}$$

$$\Phi(t) = \tan^{-1}\left(\frac{Q(t)}{I(t)}\right) - \Delta\varphi_0 \tag{21}$$

As the phase noise $\Delta\varphi_0$ is a slow variable, which can be filtered using a high-pass filter. Thus, the phase variation $\Phi(t)$ due to external acoustic signals can be obtained. The range of values of the arctangent method is $[-\frac{\pi}{2}, \frac{\pi}{2}]$, and can be extended by the use of the unwrapping algorithm [46].

Furthermore, as the pulses propagate forward in the fiber, the phase information from the previous point is carried back. As a result, the RBS at the location behind the interference point has the interference information. This RBS cannot complete the positioning of the interference signal. Separating the phase differences at a certain distance can eliminate the problem of phase accumulation. It is worth noting that the RBS results from coherence within a half pulse. Therefore, the length of the phase difference should be greater than half a pulse width. This should be greater than the spatial resolution in this system. By the phase difference algorithm, the phase change of the corresponding part of the fiber can be demodulated as follows:

$$\Delta\Phi(t) = \Phi_{z1}(t) - \Phi_{z2}(t) \tag{22}$$

For the demodulated amplitudes, the same is true. Thus, using the different algorithms on the demodulation results can achieve localization of the acoustic signal.

3 Experiments

3.1 SOA-based FRL output measurement

To verify the effectiveness of the SOA-based FRL, an experimental setup is constructed, as shown in Figure 5. The modulated signals are generated from a PPG. The driver circuit of the SOA amplifies them. The amplified modulated signals are applied to the SOA. The driver circuit integrates the temperature control system and the current control system of the SOA optical chip. This mainly aims to prevent damage to the optical chip during operation and improve the laser's power stability. The bandwidth of the FBG is <0.3 nm, and the reflectance index is >90%. The light from the FRL is connected to a photodetector after passing through a 20 dB attenuator (ATT). The bandwidth of the PD is 200 MHz. The signal from the PD is acquired by a high-speed data acquisition card (DAQ) with a sampling rate of 250 MHz. The output results are displayed on a personal computer (PC). The output results of applying different pulse duration and repetition frequency modulations to the SOA are shown in Figure 6. The amplitude of the SOA-based FRL is large enough for DAS. The short rise/fall SOA results also show good switching performance.

3.2 Dual pulses measurement

As shown in Figure 7, an unbalanced Mach-Zehnder interferometer (MZI) converts the single pulse from the SOA-based FRL into periodic double pulses separated by ΔT. The single pulse is split into the MZI by a 30:70 SM coupler. The split pulses are propagated along two different paths. The short path propagates through an acoustic-optic modulator (AOM) with a frequency shift of $\Delta f = 80MHz$, and the long path is through an attenuator and a delay. The AOM is used only as a frequency shifter, driven by a DC source and an RF driver. The center wavelength of the SOA-based FRL's output is f_{cav}. Then, $f_1 = f_{cav}$ and $f_2 = f_{cav} + \Delta f$. Thus, Δf, the heterodyne frequency, also separates the dual pulses in the frequency domain. The difference in length L_d of the two arms of the MZI is 50 m, which corresponds to the pulse interval $\Delta T = \frac{nL_d}{c} \approx 245\text{ns}$. The pulse repetition frequency is 10 kHz. The PD with a bandwidth of 200 MHz detects the pulse traces from the MZI. Figure 8 shows the detected dual pulse with ten traces. The results show that when the pulse width is 240 ns, the dual pulses can also be separated in the time domain. However, when the pulse width is 250 ns, the dual pulses overlap. The experimental results show that the rise/fall time of the SOA-based FRL is less than 10 ns, which has significant advantages over the commonly used AOM [47].

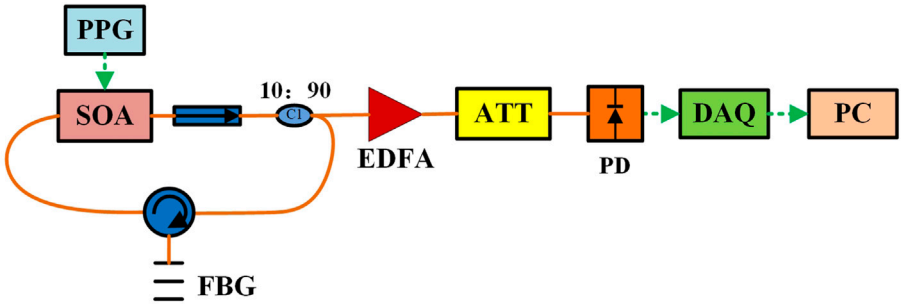


FIGURE 5
Experiment setup for SOA-based FRL output measurement.

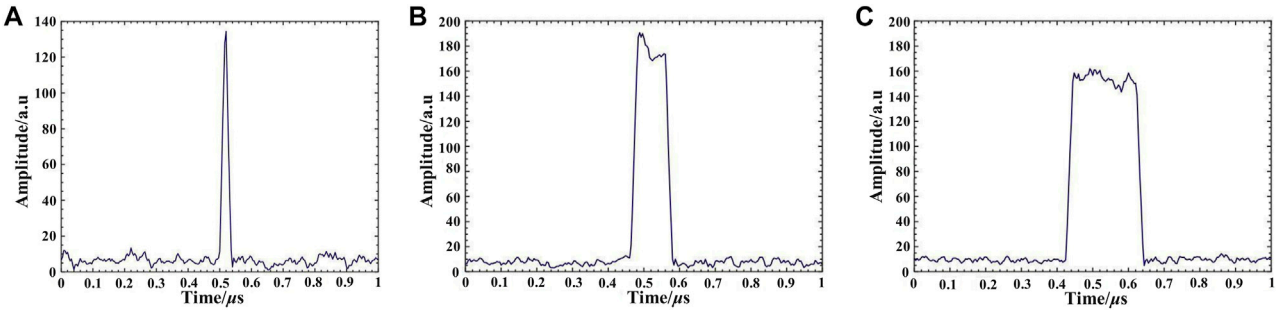


FIGURE 6
Output when modulation signals are applied when the pulse duration and repetition frequency are **(A)** 10 kHz, 20 ns, **(B)** 50 kHz, 100 ns, **(C)** 20 kHz, 200 ns.

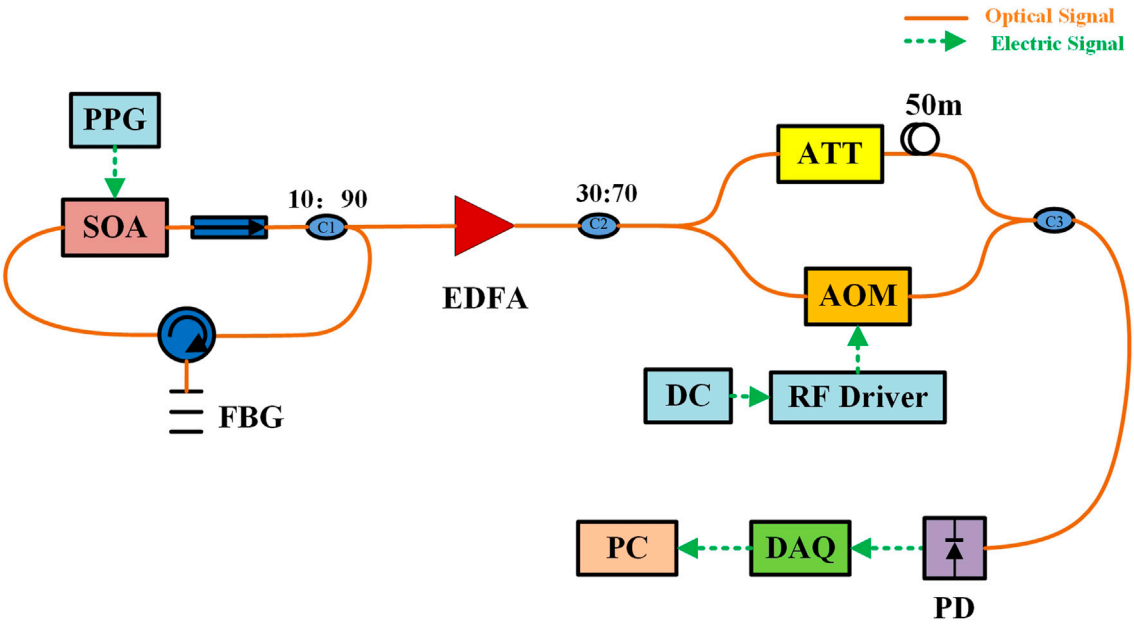


FIGURE 7
Experiment setup for dual pulse measurement.

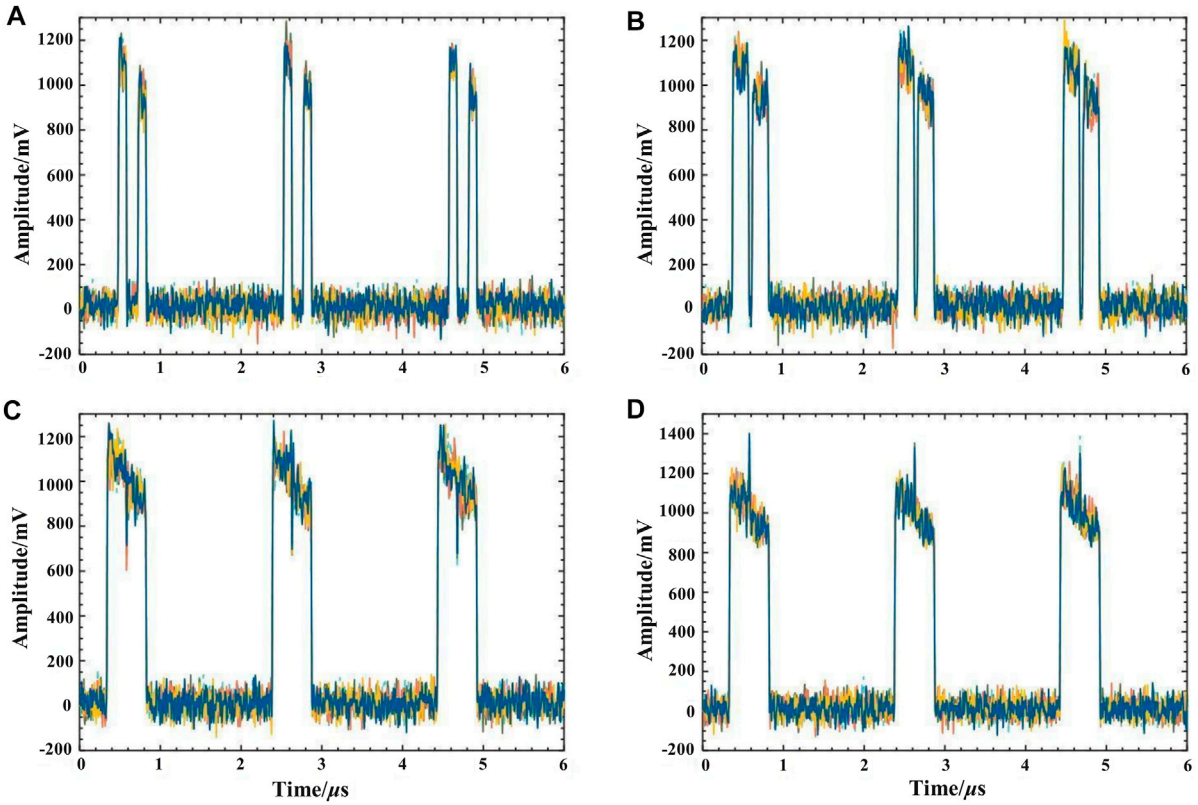


FIGURE 8
Dual pulse output when the repetition frequency is 10 kHz with different pulse durations of **(A)** 100 ns, **(B)** 150 ns,**(C)** 240 ns, and **(D)** 250 ns.

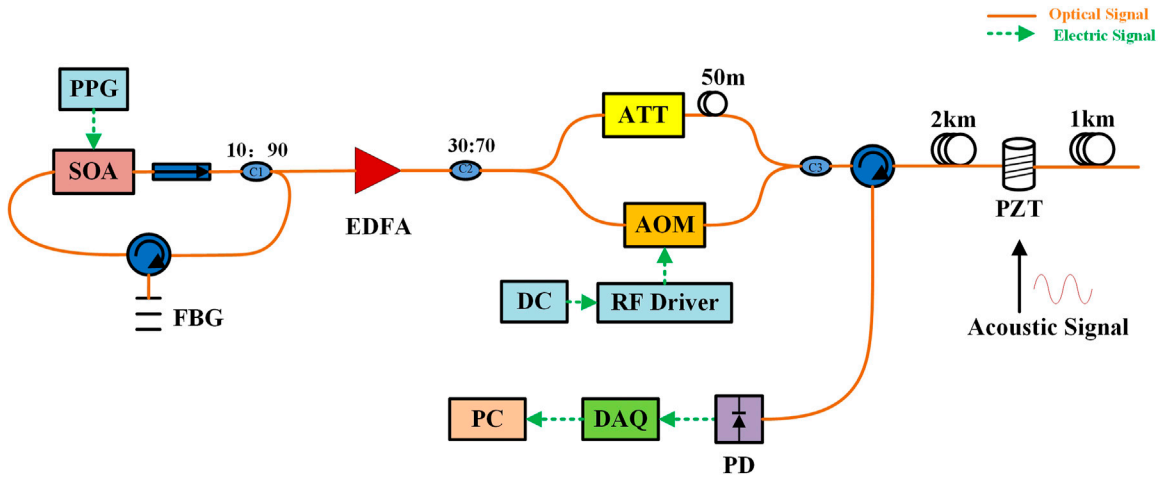


FIGURE 9
Experimental system for a dual-pulse heterodyne Φ-OTDR employing SOA-based FRL.

3.3 Acoustic signals detection

An experimental system is shown in Figure 9 to verify the ability of the proposed method to detect acoustic signals. The length of the FUT is 3 km long single-mode fiber (SMF). An arbitrary function generator (AFG) controls a piezoelectric ceramic transducer (PZT). The PZT simulates acoustic signals applied to 2 km of the FUT. In the experiments, the AOM shifted the light by 80 MHz. The PPG modulates the SOA so that the SOA-based FRL outputs an optical pulse with $\tau = 240ns$ width and $f_r = 10kHz$ repetition frequency.

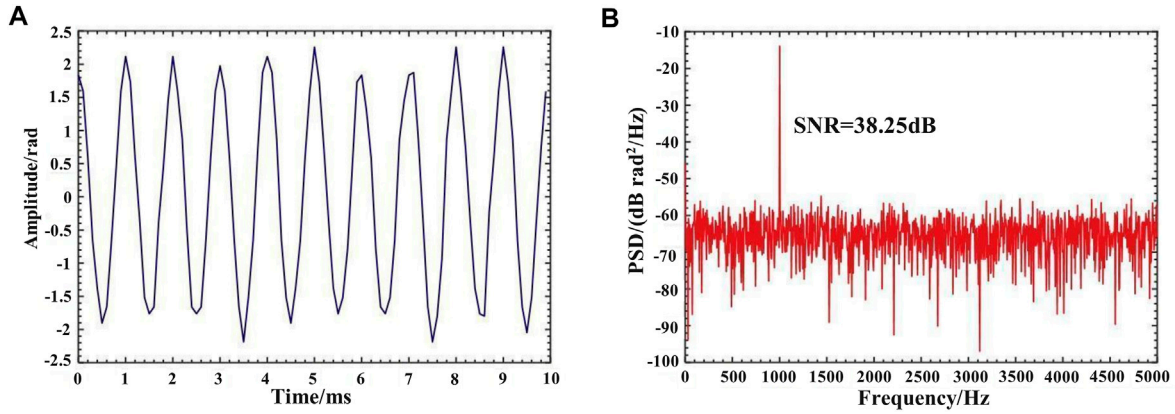


FIGURE 10
Demodulation results for 1 Vpp and 1 kHz signal **(A)** in time-domain trace **(B)** in the frequency domain.

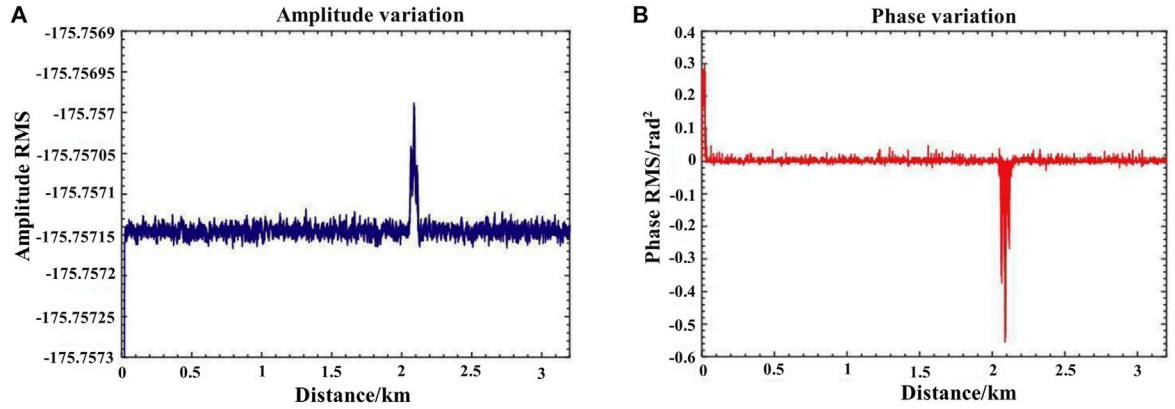


FIGURE 11
Amplitude and phase variations for 1 Vpp and 1 kHz signal.

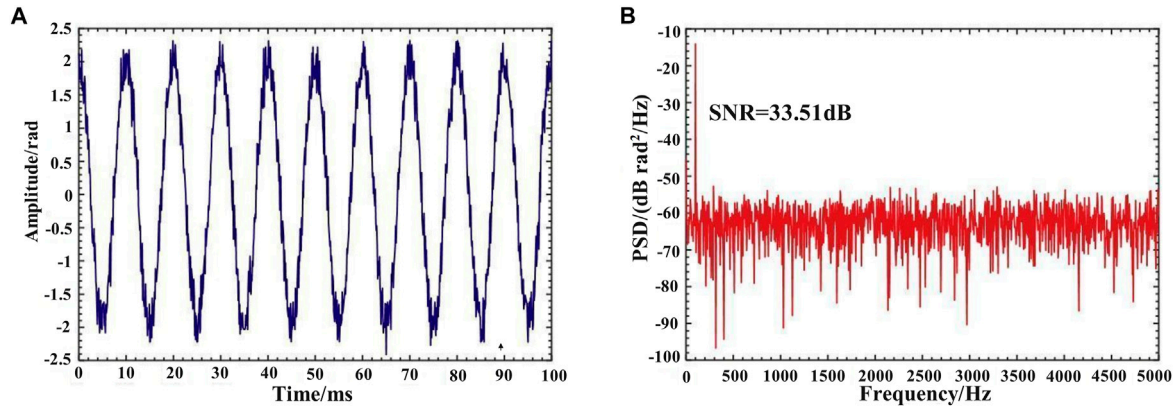


FIGURE 12
Demodulation results for 1 Vpp and 100 Hz signal **(A)** in time-domain trace **(B)** in the frequency domain.

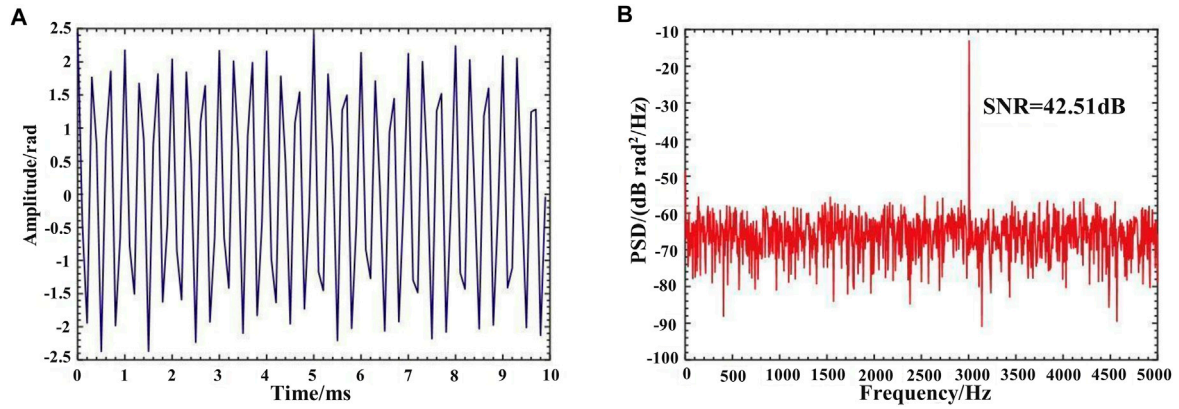


FIGURE 13
Demodulation results for 1 Vpp and 3 kHz signal **(A)** in time-domain trace **(B)** in the frequency domain.

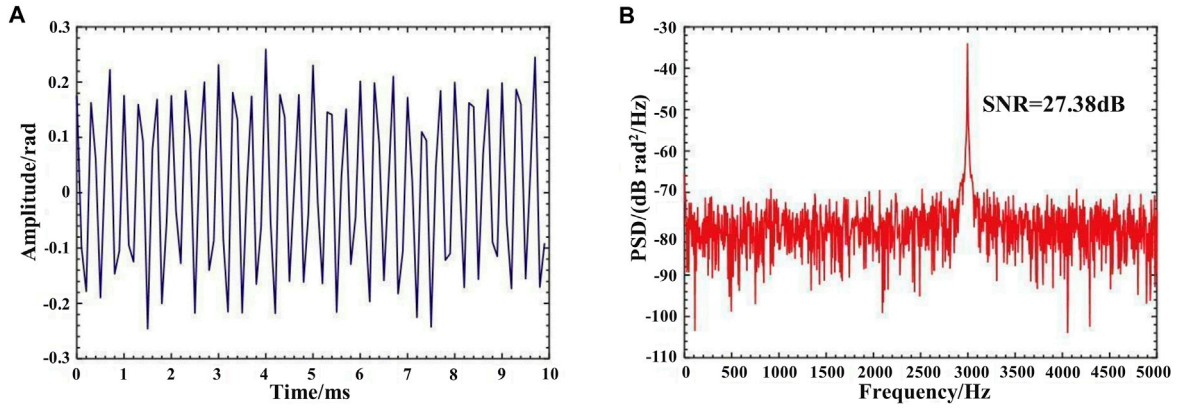


FIGURE 14
Demodulation results for 100 mVpp and 3 kHz signal **(A)** in time-domain trace **(B)** in the frequency domain.

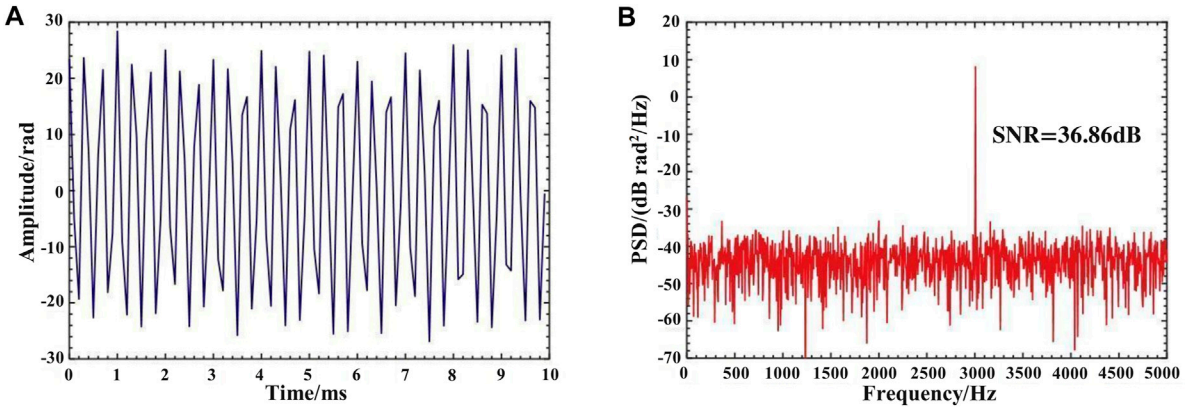


FIGURE 15
Demodulation results for 12 Vpp and 3 kHz signal **(A)** in time-domain trace **(B)** in the frequency domain.

For each measurement, $M = 2000$ traces of RBS signals are collected. In this case, the theoretical minimum detectable frequency is $\frac{f_r}{M} = 5Hz$, while the maximum is 5 kHz. And the maximum detection range is up to about $L_s = \frac{c}{2nf_rM} \approx 10.2km$. Then the given spatial resolution is $\frac{w_\tau + L_d}{2} = 11.8m$, where w_τ is the pulse width in length. The interference signals can be reconstructed by extracting specific points in each trace. The heterodyne demodulation technique algorithm is used to retrieve the signal's phase information. The demodulation process is conducted in the PC.

Figure 10 shows the demodulation results when the modulation signal frequency with a peak amplitude of 1 V applied to the PZT is 1 kHz. A power spectral density (PSD) analysis is conducted with the demodulation results. The results are in the time domain and frequency domain.

A differential algorithm can be applied to the demodulated signal, allowing the acoustic signal to be located. Figure 11 shows the amplitude and phase variance results for a 1 Vpp and 1 kHz signal. The results agree well with the location of the PZT. The results imply that the system can localize the acoustic signal accurately.

The modulation signals applied to the PZT are varied with a fixed amplitude to demonstrate the system's response to other frequencies. The sinusoidal signal applied to the PZT was fixed at a constant peak voltage of 1 V. The frequency was gradually changed for the measurement and demodulated, as shown in Figures 12, 13. As can be seen from the demodulation results, the system can accurately detect the acoustic signals at the fixed point of the fiber with a complete signal waveform. The spectrum also shows that the system can achieve a high SNR of 42.51 dB and accurately demodulate the signal frequencies.

The modulation signals applied to the PZT are varied in amplitude at a fixed frequency to demonstrate the system's dynamic response. The sinusoidal signal applied to the PZT has been set at 3 kHz; the amplitude is 100 mV and 12 V, respectively, corresponding to 0.19 rad and 22.8 rad. The results of the demodulation are shown in Figures 14, 15. The demodulation results show that the system can accurately detect acoustic signals with different amplitudes.

The experimental results illustrate that the proposed system can detect acoustic signals with different frequencies and amplitudes. The demodulation and positioning results are accurate, fitting the modulation signals applied to the PZT well. For a 3 kHz acoustic signal, the SNR can achieve 42.51 dB. Above 100 Hz, the SNR of the demodulation results remained at 27.38 dB and above.

4 Conclusion

A dual pulse heterodyne DAS system is proposed by introducing SOA-based FRL in this paper. The proposed system is investigated theoretically and experimentally. The SOA-based FRL operates as a pulsed mode-locked laser, which replaces the NLL and pulse modulator, simplifying the system and reducing costs. A narrow linewidth optical pulse output can be obtained using the device, which fully exploits the gain and good switching characteristics of SOA. The effectiveness of the device is demonstrated in a dual pulse heterodyne DAS system. The proposed system achieves accurate demodulation and locatilization. The system's frequency range is 5 Hz to 5 kHz, and a high SNR of 42.51 dB demodulation is achieved at 3 kHz. With a detection range of 10.2 km, the system's spatial resolution is 12 m. The proposed system provides an alternative idea for DAS, significantly reducing the cost and simplifying the system. It is expected to greatly benefit from cost-effective, large-scale, and high-SNR applications in DAS. As the system can be modified by adding additional FBGs to support more laser wavelengths, further work will focus on multi-wavelength SOA-based FRL and its improvements. The multi-wavelength output is expected to benefit the practical applications of DAS greatly.

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

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

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

References

1. Dandridge A. Fiber optic interferometric sensors at sea. *Opt Photon News* (2019) 30(6):34. doi:10.1364/opn.30.6.000034

2. Toky A, Singh RP, Das S. Localization schemes for underwater acoustic sensor networks - a review. *Comput Sci Rev* (2020) 37:100241. doi:10.1016/j.cosrev.2020.100241

3. Lu B, Wu B, Gu J, Yang J, Gao K, Wang Z, et al. Distributed optical fiber hydrophone based on Phi-OTDR and its field test. *Opt Express* (2021) 29(3):3147–62. doi:10.1364/oe.414598

4. Almudévar A, Sevillano P, Vicente L, Preciado-Garbayo J, Ortega A. Unsupervised anomaly detection applied to Φ-OTDR. *Sensors* (2022) 22:6515. doi:10.3390/s22176515

5. Martins HF, Martin-Lopez S, Corredera P, Filograno ML, Frazao O, Gonzalez-Herraez M. Coherent noise reduction in high visibility phase-sensitive optical time domain reflectometer for distributed sensing of ultrasonic waves. *J Lightwave Technol* (2013) 31(23):3631–7. doi:10.1109/jlt.2013.2286223

6. Wang Z, Zhang L, Wang S, Xue N, Peng F, Fan M, et al. Coherent Φ-OTDR based on I/Q demodulation and homodyne detection. *Opt Express* (2016) 24(2):853–8. doi:10.1364/oe.24.000853

7. Soriano-Amat M, Martins HF, Durán V, Costa L, Martin-Lopez S, Gonzalez-Herraez M, et al. Time-expanded phase-sensitive optical time-domain reflectometry. *Light Sci Appl* (2021) 10(1):51. doi:10.1038/s41377-021-00490-0

8. Zhong X, Gui D, Zhang B, Deng H, Zhao S, Zhang J, et al. Performance enhancement of phase-demodulation ϕ-OTDR using improved two-path DCM algorithm. *Opt Commun* (2021) 482:126616. doi:10.1016/j.optcom.2020.126616

9. Zhong X, Zhang B, Ren J, Deng H, Chen X, Ma M. A novel ϕ-OTDR system with a phase demodulation module based on sagnac balanced interferometer. *J Lightwave Technol* (2021) 39(22):7307–14. doi:10.1109/jlt.2021.3113082

10. Liu T, Wang F, Zhang X, Yuan Q, Niu J, Zhang L, et al. Interrogation of ultra-weak FBG array using double-pulse and heterodyne detection. *IEEE Photon Technol Lett* (2018) 30(8):677–80. doi:10.1109/lpt.2018.2811411

11. He X, Xie S, Liu F, Cao S, Gu L, Zheng X, et al. Multi-event waveform-retrieved distributed optical fiber acoustic sensor using dual-pulse heterodyne phase-sensitive OTDR. *Opt Lett* (2017) 42(3):442–5. doi:10.1364/ol.42.000442

12. Zinsou R, Liu X, Wang Y, Zhang J, Jin B. Recent progress in the performance enhancement of phase-sensitive OTDR vibration sensing systems. *Sensors* (2019) 19(7):1709. doi:10.3390/s19071709

13. Ibrahim ADA, Abbas K, Linda BE. SNR enhancement by DWT for improving the performance of Φ-OTDR in vibration sensing. *Photon Switching Comput* (2022) 5.

14. Qin Z, Chen L, Bao X. Wavelet denoising method for improving detection performance of distributed vibration sensor. *IEEE Photon Technol Lett* (2012) 24(7):542–4. doi:10.1109/lpt.2011.2182643

15. Shi Y, Feng H, Zeng Z. A long distance phase-sensitive optical time domain reflectometer with simple structure and high locating accuracy. *Sensors* (2015) 15:21957–70. doi:10.3390/s150921957

16. Qin Z, Chen H, Chang J. Detection performance improvement of distributed vibration sensor based on curvelet denoising method. *Sensors* (2017) 17:1380. doi:10.3390/s17061380

17. Qu S, Chang J, Cong Z, Chen H, Qin Z. Data compression and SNR enhancement with compressive sensing method in phase-sensitive OTDR. *Opt Commun* (2019) 433:97–103. doi:10.1016/j.optcom.2018.09.064

18. Li D, Lou S, Xin Q, Liang S, Sheng X. SNR enhancement of far-end disturbances on distributed sensor based on phase-sensitive optical time-domain reflectometry. *IEEE Sens J* (2021) 21(2):1957–64. doi:10.1109/jsen.2020.3019838

19. He H, Shao L, Li H, Pan W, Luo B, Zou X, et al. SNR enhancement in phase-sensitive OTDR with adaptive 2-D bilateral filtering algorithm. *IEEE Photon J* (2017) 9(3):1–10. doi:10.1109/jphot.2017.2700894

20. Adeel M, Tejedor J, Macias-Guarasa J, Lu C. Improved perturbation detection in direct detected -otdr systems using matched filtering. *IEEE Photon Technol Lett* (2019) 31(21):1689–92. doi:10.1109/lpt.2019.2940297

21. Muanenda Y, Oton CJ, Faralli S, Di Pasquale F. A cost-effective distributed acoustic sensor using a commercial off-the-shelf DFB laser and direct detection phase-OTDR. *IEEE Photon J* (2016) 8:1–10. doi:10.1109/jphot.2015.2508427

22. Wang Z-N, Zhang B, Xiong J, Fu Y, Lin S, Jiang J, et al. Distributed acoustic sensing based on pulse-coding phase-sensitive OTDR. *IEEE Internet Things J* (2019) 6:6117–24. doi:10.1109/jiot.2018.2869474

23. Soriano-Amat M, Martins HF, Duran V, Martin-Lopez S, Gonzalez-Herraez M, Fernández-Ruiz MR. Quadratic phase coding for SNR improvement in time-expanded phase-sensitive OTDR. *Opt Lett* (2021) 46(17):4406–9. doi:10.1364/ol.432350

24. Martins HF, Martín-López S, Corredera P, Ania-Castanon JD, Frazao O, Gonzalez-Herraez M. Distributed vibration sensing over 125 km with enhanced SNR using phi-OTDR over a URFL cavity. *J Lightwave Technol* (2015) 33:2628–32. doi:10.1109/jlt.2015.2396359

25. Choi KN, Taylor HF. Spectrally stable Er-fiber laser for application in phase-sensitive optical time-domain reflectometry. *Photon Technol Lett IEEE* (2003) 15(3):386–8. doi:10.1109/lpt.2003.807905

26. Qian H, Luo B, He H, Zhang X, Zou X, Pan W, et al. Phase demodulation based on DCM algorithm in Φ-OTDR with self-interference balance detection. *IEEE Photon Technol Lett* (2020) 32:473–6. doi:10.1109/lpt.2020.2979030

27. Martins HF, Martín-López S, Corredera P, Filograno ML, Frazao O, Gonzalez-Herraez M. Phase-sensitive optical time domain reflectometer assisted by first-order Raman amplification for distributed vibration sensing over >100 km. *J Lightwave Technol* (2014) 32:1510–8. doi:10.1109/jlt.2014.2308354

28. Pastor-Graells J, Nuño J, Fernández-Ruiz MR, Garcia-Ruiz A, Martins HF, Martin-Lopez S, et al. Chirped-pulse phase-sensitive reflectometer assisted by first-order Raman amplification. *J Lightwave Technol* (2017) 35:4677–83. doi:10.1109/jlt.2017.2756558

29. Du L, Shen Y, Yang S, Liang Y. Research on RP-EDF amplification characteristics based on φ-OTDR system. *Optik: Z Licht- Elektronenoptik: = J Light-and Electronoptic.* (2022) 262:169029. doi:10.1016/j.ijleo.2022.169029

30. Muanenda Y Recent advances in distributed acoustic sensing based on phase-sensitive optical time domain reflectometry. *J Sens* (2018) 2018(1):1–16. doi:10.1155/2018/3897873

31. Liu X, Jin B, Bai Q, Wang Y, Wang D, Wang Y. Distributed fiber-optic sensors for vibration detection. *Sensors* (2016) 16(8):1164. doi:10.3390/s16081164

32. Aktas M, Maral H, Akgun T. A model-based analysis of extinction ratio effects on phase-OTDR distributed acoustic sensing system performance. *SPIE Conf Phys Simulation Optoelectronic Devices* (2018).

33. Chen Y, Mao B-M, Zhou B, Guo C, Lin Z. Improving the SNR of the phase-OTDR by controlling the carrier in the SOA. *J Mod Opt* (2020) 67:1241–6. doi:10.1080/09500340.2020.1827071

34. Zulkifli MZ, Hassan NA, Awang NA, Ghani Z, Harun S, Ahmad H. Multi-wavelength fiber laser in the S-band region using a Sagnac loop mirror as a comb generator in an SOA gain medium. *Laser Phys Lett* (2010) 7(9):673–6. doi:10.1002/lapl.201010046

35. Madrigal J, Fraile-Peláez FJ, Zheng D, Barrera D, Sales S. Characterization of a FBG sensor interrogation system based on a mode-locked laser scheme. *Opt Express* (2017) 25(20):24650–7. doi:10.1364/oe.25.024650

36. Slepneva S, Kelleher B, O'Shaughnessy B, Hegarty S, Vladimirov A, Huyet G. Dynamics of Fourier domain mode-locked lasers. *Opt Express* (2013) 21(16):19240–51. doi:10.1364/oe.21.019240

37. Todor S, Biedermann B, Huber R, Jirauschek C. Balance of physical effects causing stationary operation of Fourier domain mode-locked lasers. *J Opt Soc Am B* (2012) 29(4):656–64. doi:10.1364/josab.29.000656

38. Jirauschek C, Biedermann B, Huber R. A theoretical description of Fourier domain mode locked lasers. *Opt Express* (2009) 17(26):24013–9. doi:10.1364/oe.17.024013

39. Ahmed M, Yamada M. Influence of instantaneous mode competition on the dynamics of semiconductor lasers. *IEEE J Quan Electron* (2002) 38(6):682–93. doi:10.1109/jqe.2002.1005419

40. Reale A, Di Carlo A, Lugli P. Gain dynamics in traveling-wave semiconductor optical amplifiers. *IEEE J Sel Top Quan Electron* (2001) 7(2):293–9. doi:10.1109/2944.954142

41. Feng X, Tam H-Y, Liu H, Wai P. Multiwavelength erbium-doped fiber laser employing a nonlinear optical loop mirror. *Opt Commun* (2006) 268(2):278–81. doi:10.1016/j.optcom.2006.07.010

42. Yamada M Theory of mode competition noise in semiconductor injection lasers. *IEEE J Quan Electron* (1986) 22(7):1052–9. doi:10.1109/jqe.1986.1073087

43. Hu Z, Davanco M, Blumenthal DJ. Extinction ratio improvement by strong external light injection and SPM in an SOA for OTDM pulse source using a DBR laser diode. *IEEE Photon Technol Lett* (2003) 15(10):1419–21. doi:10.1109/lpt.2003.818258

44. Connelly MJ. Wideband semiconductor optical amplifier steady-state numerical model. *IEEE J Quan Electron* (2001) 37(3):439–47. doi:10.1109/3.910455

45. Park J, Lee W, Taylor HF. Fiber optic intrusion sensor with the configuration of an optical time-domain reflectometer using coherent interference of Rayleigh backscattering. *Opt Fiber Optic Sensor Syst* (1998).

46. Gao X, Hu W, Dou Z, Li K, Gong X. A method on vibration positioning of Φ-OTDR system based on compressed sensing. *IEEE Sens J* (2022) 22(16):16422–9. doi:10.1109/jsen.2022.3191863

47. Oh JM, Koo SG, Lee D, Park SJ. Enhancement of the performance of a reflective SOA-based hybrid WDM/TDM PON system with a remotely pumped erbium-doped fiber amplifier. *J Lightwave Technol* (2008) 26(1):144–9. doi:10.1109/jlt.2007.913073

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 28 July 2023
DOI 10.3389/fphy.2023.1240555

OPEN ACCESS

EDITED BY
Xukun Yin,
Xidian University, China

REVIEWED BY
Jing Ba,
Jiangsu University of Science and Technology, China
Ke Lu,
Nanjing University of Information Science and Technology, China
Heng Du,
Nanjing Institute of Technology (NJIT), China

*CORRESPONDENCE
Xiaojuan Mao,
1017284834@qq.com
Pingping Gu,
gupingping@ntu.edu.cn
Hengrong Ju,
juhengrong@ntu.edu.cn

RECEIVED 15 June 2023
ACCEPTED 10 July 2023
PUBLISHED 28 July 2023

CITATION
Fan X, Mao X, Cai T, Sun Y, Gu P and Ju H (2023), Sensor data reduction with novel local neighborhood information granularity and rough set approach. *Front. Phys.* 11:1240555. doi: 10.3389/fphy.2023.1240555

Sensor data reduction with novel local neighborhood information granularity and rough set approach

Xiaoxue Fan[1], Xiaojuan Mao[2]*, Tianshi Cai[1], Yin Sun[3], Pingping Gu[1,4]* and Hengrong Ju[1]*

[1]School of Information Science and Technology, Nantong University, Nantong, China, [2]Department of Respiratory Medicine, The Sixth People's Hospital of Nantong, Affiliated Nantong Hospital of Shanghai University, Nantong, China, [3]Jiangsu Vocational College of Business, Nantong, China, [4]School of Transportation and Civil Engineering, Nantong University, Nantong, China

Data description and data reduction are important issues in sensors data acquisition and rough sets based models can be applied in sensors data acquisition. Data description by rough set theory relies on information granularity, approximation methods and attribute reduction. The distribution of actual data is complex and changeable. The current model lacks the ability to distinguish different data areas leading to decision-making errors. Based on the above, this paper proposes a neighborhood decision rough set based on justifiable granularity. Firstly, the rough affiliation of the data points in different cases is given separately according to the samples in the neighborhood. Secondly, the original labels are rectified using pseudo-labels obtained from the label noise data that has been found. The new judgment criteria are proposed based on justifiable granularity, and the optimal neighborhood radius is optimized by the particle swarm algorithm. Finally, attribute reduction is performed on the basis of risky decision cost. Complex data can be effectively handled by the method, as evidenced by the experimental results.

KEYWORDS

justifiable granularity, sensor data, local neighborhood decision rough set model, attribute reduction, granular computing

1 Introduction

In sensor data processing systems, researchers are often confronted with large amounts of multimodal and complex sensing data. To deal with these sensing data, data description and data reduction are pivotal process. For data acquisition, rough sets based models are considered as effective approaches in recent years [1, 2]. Rough set theory [3] was proposed in 1982 by Pawlak as a mathematical tool for analyzing and handling imprecise, inconsistent, and incomplete information. Traditional rough set theory lacks fault tolerance and does not take errors in the classification process into account at all. Pawlak et al. proposed the probabilistic rough set model to improve rough set theory using probabilistic threshold [4]. A probabilistic rough set model has been introduced to Bayesian decision theory by [5]. Further, Yao proposed a three-way decision theory on the basis of decision rough set theory [6].

Currently, many scholars have been improving the research on decision rough sets from different aspects. [7] proposed the theoretical framework of local rough set. [8] proposed

local neighborhood rough set, which integrated the neighborhood rough set and local rough set. [9] combined Lebesgue and entropy measure, and proposed a novel attribute reduction approach. [10] introduced the pseudo-label into rough set, and proposed a pseudo-label neighborhood relationship, which can distinguish samples by distance measure and pseudo-labels.

As mentioned above, scholars proposed equivalent modifications to the neighborhood decision rough set approach from multiple perspectives. However, for complex sensor data processing, neighborhood decision rough set methods still face some challenges. For example, in practical applications, complex data distribution is often uneven. In addition, the presence of abnormal data can also greatly weaken the performance of rough models and cannot correctly classify abnormal data points. For the issues mentioned above, this paper proposes a local strategy to improve the calculation process of rough membership. Additionally, the neighborhood of sample is optimized by the particle swarm optimization method (PSO algorithm) to offer the optimal neighborhood granularity for the model and carry out attribute reduction.

The remainder of this paper is structured as follows. Section 2 introduces the relevant basic theories. Section 3 presents a decision rough calculation method based on justifiable granularity. Six datasets are chosen in Section 4 to evaluate the suggested methodology. Section 5 summarizes the full text.

2 Preliminary notion

2.1 Neighborhood relation and rough set

The construction of equivalence relations for numerical type data first requires the discretization of the original data, and this method will inevitably cause the loss of information. On the basis of neighborhood relations, a neighborhood rough set model was proposed by Hu et al. [11–13].

Assume that information system is expressed as $S = (U, AT = C \cup D, f, V)$.Among them, $U = \{x_1, x_2, \ldots, x_n\}$ represents a collection of non-empty limited objects, AT stands for the set of attributes, containing conditional attribute set C and decision attribute set D.

Definition 1. Suppose the information system is $S = (U, AT = C \cup D, f, V)$, $\forall x \in U$, $B \subseteq C$, the δ-neighborhood of x in B is defined as:

$$\delta_B(x) = \left\{ y \in U \middle| dis_B(x, y) \leq \delta, \delta > 0 \right\} \tag{1}$$

where dis $(\bullet)$ represents the distance between any objects, using Euclidean distance commonly.

Definition 2. Suppose the information system is $S = (U, AT = C \cup D, f, V)$, $\forall x \in U$, $X \subseteq U$, $B \subseteq C$, the rough affiliation $\mu_B(x)$ of x to X in B is defined as:

$$\mu_B(x) = P(X|\delta_B(x)) = \frac{|X \cap \delta_B(x)|}{|\delta_B(x)|} \tag{2}$$

where $P(X|\delta_B(x))$ represents the conditional probability of classification, and $|\bullet|$ represents the number of elements in the combination.

Definition 3. Suppose the information system is $S = (U, AT = C \cup D, f, V)$, $X \subseteq U, B \subseteq C$, the lower and upper approximations of the decision D in B are defined as:

$$\overline{\delta_B}(X) = \{x \in U | \delta_B(x) \cap X \neq \varnothing\} \tag{3}$$

$$\underline{\delta_B}(X) = \{x \in U | \delta_B(x) \subseteq X\} \tag{4}$$

The following definitions apply to the positive, negative, and boundary regions of X in B:

$$POS_B(X) = \underline{\delta_B}(X) = \{x \in U \mid P(X|\delta_B(X)) = 1\} \tag{5}$$

$$NEG_B(X) = U - \underline{\delta_B}(x) = \{x \in U \mid P(X|\delta_B(x)) = 0\} \tag{6}$$

$$BND_B(X) = \overline{\delta_B}(x) - \underline{\delta_B}(x) = \{x \in U | 0 < P(X|\delta_B(x) < 1\} \tag{7}$$

From the above definition, it can be found that the conditions on which the neighborhood rough set is based in taking both acceptance and rejection decisions are too severe and lack a certain degree of fault tolerance. Only elements that are completely correctly classified are grouped into the positive domain. Alternatively, only elements that are completely misclassified are classified in the negative domain. The result of such a definition makes the boundary domain too large.

2.2 Rough set with neighborhood decision

The rough set model for decision-making put forth by Yao et al. [5] lacks the ability to directly process numerical data. In order to address this weakness, a rough set model of decision theory based on neighborhood was proposed by Li et al. [14] through the integration of the neighborhood rough set and the decision rough set.

The decision rough set has two important elements: $\Omega = \{X, \sim X\}$ and $Action = \{a_P, a_B, a_N\}$. When different decision-making actions are taken, different losses will occur. λ_{PP}, λ_{BP}, λ_{NP} respectively represent the cost of a_P, a_B and a_N when X owns the object, λ_{PN}, λ_{BN}, λ_{NN} respectively represent the cost of a_P, a_B and a_N when X is not the owner of the object. Through cost risk analysis, the solution formula of (α, β) is given [5] as follows:

$$\alpha = \frac{\lambda_{PN} - \lambda_{BN}}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{BP} - \lambda_{PP})} \tag{8}$$

$$\beta = \frac{\lambda_{BN} - \lambda_{NN}}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})} \tag{9}$$

In addition, Yao proposed three decision theories based on decision rough set model [5], including P rule, N rule and B rule.

Definition 4. Suppose the information system $S = (U, AT = C \cup D, f, V)$, $X \subseteq U$, $B \subseteq C$, then the P, B, and N rules of X on δ-neighborhood under attribute set B are defined as:

P rule: if $x \in U$, $P(X|\delta_B(X)) \geq \alpha$, then $x \in POS_B(X)$;

B rule: if $x \in U$, $\beta < P(X|\delta_B(X)) < \alpha$, then $x \in BND_B(X)$;

N rule: if $x \in U$, $P(X|\delta_B(X)) \leq \beta$, then $x \in NEG_B(X)$.

3 Neighborhood decision rough set model based on justifiable granularity

To solve the problems discussed above, this article first introduces the local neighborhood rough set model to eliminate the interference of some noise data on the approximate set.

3.1 Local rough neighborhood decision model

Definition 5. Suppose the information system $S = (U, AT = C \cup D, f, V)$, $X \subseteq U$, $B \subseteq C$, then the X of the attribute set B is related to the upper and lower approximation sets of the δ-neighborhood based local rough set, which are defined as:

$$\overline{\delta_B^L}(X) = \left\{x \in X \middle| P(X|\delta_B(x)) > \beta\right\} \tag{10}$$

$$\underline{\delta_B^L}(X) = \{x \in X | P(X|\delta_B(x)) \geq \alpha\} \tag{11}$$

The following definitions apply to the positive, negative, and boundary regions of X in B:

$$POS_B(X) = \underline{\delta_B^L}(X) = \{x \in X \mid P(X|\delta_B(X)) \geq \alpha\}; \tag{12}$$

$$NEG_B(X) = U - \underline{\delta_B^L}(x) = \{x \in X \mid P(X|\delta_B(x)) \leq \beta\}; \tag{13}$$

$$BND_B(X) = \overline{\delta_B^L}(x) - \underline{\delta_B^L}(x) = \left\{x \in X \middle| \beta < P(X|\delta_B(x) < \alpha\right\}. \tag{14}$$

The most significant difference between the local neighborhood rough set model and the neighborhood rough set model is the different search scope when finding the upper and lower approximation sets. In the neighborhood rough set model, finding the approximation set for each decision category requires traversing all the data points in the data set. However, in the local neighborhood rough set model, the focus is on the data points of the same category, and only the data points of the same decision category need to be traversed. This greatly reduces the computational effort and increases the computational speed [14]. This model not only improves computational efficiency, but also eliminates the interference of noisy points.

In addition, the traditional method of calculating rough affiliation does not take into account the complexity of the data. In this paper, the calculation process of affiliation degree is improved for the affiliation degree, and the process is as follows:

Suppose $S = (U, AT = C \cup D, f, V)$ is a decision system, $U/D = \{X_1, X_2, ..., X_d\}$ is the decision attribute of all objects U in the decision attribute set D, $\forall x \in U$, the neighborhood of x is expressed as $\delta(x)$, the decision value of the information system is $L = \{1, 2, ..., d\}$. Now suppose that the decision value of the sample x to be investigated is q.

(1) $|\delta(x)| \leq N$ (N represents a small positive integer), this paper sets rough membership degree to $P(X|\delta(x)) = e^{-5}$.

(2) $L_x = q$, $\forall x_i \in \delta(x)$, $|L_{x_i}| = 1$ and $L_{x_i} = q$, this paper sets rough membership degree to $\min[1, p_0 + s \times (|\delta(x)| - N)]$, where p_0 represents the initial probability value and N represents the minimum number of neighborhoods, s represents the search step.

(3) $L_x = q$, $\forall x_i \in \delta(x) - x_i$, $|L_{x_i}| = 1$ and $|L_{x_i}| \neq q$, rough membership degree is set to $P(X|\delta(x)) = 0$.

Depending on which of the data points in the neighborhood information granularity are specifically situated, above rules is used to define the rough membership function for each category of data points.

Based on the above discussion, this paper designs the following Algorithm 1 to calculate the upper and lower approximation sets and identify the anomalous data. Different from the classical method that only considers the upper and lower approximation sets, Algorithm 1 not only identifies label noise data points and outlier data points based on the neighborhood information, making the upper and lower approximation sets more accurate. It also appends category information to the label noise data, which is referred to as pseudo-tagging in this paper.

```
Input: S = (U, AT = C ∪ D), neighborhood radius δ, cost matrix λ.
Output: lower approximate δ(X_q), upper approximate δ̄(X_q), outlier points set O, labeled noise points set Noise, and predicted pseudo-labels set Noise′.
Segmentation of the entire dataset by tag categories U/D = {X_1, X_2, ..., X_d}.
Using the cost matrix, the threshold value α and β are calculated according to Eqs 8, 9.
For ∀x_i ∈ X_q
  Compute the δ-neighborhood δ(x_i) of x_i on the conditional attribute set C and obtain the label category L_δ(x_i) = {1, 2, ..., d}.
end
If |δ(x_i)| ≤ N
    P(X_q|δ(x_i)) = e^-5.
    O = O ∪ {x_i}.
End
If |L_δ(x_i)| = 1 & L_x_i = q
    P(X_q|δ(x_i)) = 0.
    Noise = Noise ∪ {x_i}.
    Noise′ = Noise′ ∪ {x_i}
End
If 1 < |L_δ(x_i)| < d
    P(X_q|δ(x_i)) = |δ(x_i) ∩ X_q| / |δ(x_i)|.
End
If P(X_q|δ(x_i)) ≥ α
    δ(X_q) = δ(X_q) ∪ {x_i}.
    If P(X_q|δ(x_i)) > β
     δ̄(X_q) = δ̄(X_q) ∪ {x_i}.
  End
End
Return δ(X_q), δ̄(X_q), O, Noise, Noise′.
```

Algorithm 1 The upper and lower approximation sets of local neighborhood rough set.

Algorithm 1 detects outliers and labeled noisy points, as well as enables the detection of data points for high-density areas. In fact,

TABLE 1 Dataset description.

No.	Datasets	Sample	Attribute	Class
1	Banknote Authentication	1372	5	2
2	Cardiotocography	2126	23	10
3	Glass Identification	214	10	7
4	Ionosphere	351	34	2
5	Sonar	208	60	2
6	WDBC	569	31	2

some samples are not always considered as outlier data or noise, and their decisions sometimes depend on the choice of neighborhood radius.

3.2 Selection of neighborhood information granularity based on justifiable granularity

According to the above-mentioned rough set model, a smaller neighborhood radius contains very little information, while a larger radius may cause the next approximate set to be an empty set. This paper introduces the justifiable granularity criterion [15, 16]. There are generally two functions in the construction of information granules, namely, covering function and particularity function.

The coverage function describes how much data is in the constructed information granule. This paper designs the coverage index function as shown below:

$$\mathrm{cov}(\delta) = \max[0, F_1 + F_2] \tag{15}$$

where $F_1 = \frac{1}{|X_q|}\left(\sum_{x\in X_q}\left(|\delta_q(x)| - \max_{\substack{j=1,\ldots,d\\ j\neq q}}|\delta_j(x)|\right)\right)$ and $F_2 = \frac{1}{|\delta(x)|}$ $(|POS(X_q)|\text{-}|BND(X_q)|)$.

The coverage index function mentioned above is considered from two perspectives, namely, neighborhood information granularity and approximate set. In terms of specificity criteria, the smaller the neighborhood radius, the better. Therefore, the specificity function can be designed as: $\mathrm{sp}(\delta) = 1 - \delta$.

Obviously, the two are contradictory. Therefore, the function for optimized performance can be written as the multiplication of specificity and coverage, which is: $Q = \mathrm{cov}(\delta) \times \mathrm{sp}(\delta)$.

In this way, the optimal neighborhood about X_q can be obtained. To further elaborate, the cumulative behavior can be represented in terms of the decision partition set $U/D = \{X_1, X_2, ..., X_d\}$ as follows: $Q = Q_1 + Q_2 + ... + Q_d$, where Q_1, Q_2,. . ., Q_d correspond to the optimized value of each decision class.

To achieve the optimal Q value and the corresponding optimal neighborhood radius. In this paper, PSO algorithm is used for optimization [17, 18], which is an evolutionary algorithm based on population intelligence, proposed by Drs. Kennedy and Eberhart in 1995. In this paper, we use the PSO algorithm to intelligently optimize the selection of neighborhoods and select the appropriate granularity as a way to improve the accuracy of decision making.

Moreover, to update the dataset, one can utilize the noise identification strategy along with the set of predicted pseudo-decision labels. The main steps are described in Algorithm 2.

```
1: Obtain the optimal neighborhood radius δ using PSO optimization algorithm;
2: Execute Algorithm 1 to obtain the approximation set, the set of outlier points, the set of labeled noise points, and the pseudo-tags of labeled noise points;
3: Updating decision labels for noisy data based on pseudo-labels;
4: Update the approximation set using the modified decision system.
```

Algorithm 2 Update of rough approximation set in label noise injection environment.

3.3 Attribute reduction based on neighborhood decision rough set model

In this paper the risky decision cost will be used to reduce the attributes. It comes from the Bayesian decision process, which is comparable to the classical rough set. Risky decision costs for P, N and B rule can be separately expressed as:

$$COST_{POS} = \sum_{X_j\in U/D}\sum_{x\in POS(X_j)}\sum_{k=1}^{m}\left(\lambda_{PP}^{k}\bullet P\left(X_j\middle|[x]_{C_k}\right) + \lambda_{PN}^{k}\bullet P\left(\sim X_j\middle|[x]_{C_k}\right)\right) \tag{16}$$

$$COST_{NEG} = \sum_{X_j\in U/D}\sum_{x\in NEG(X_j)}\sum_{k=1}^{m}\left(\lambda_{NP}^{k}\bullet P\left(X_j\middle|[x]_{C_k}\right) + \lambda_{NN}^{k}\bullet P\left(\sim X_j\middle|[x]_{C_k}\right)\right) \tag{17}$$

$$COST_{BND} = \sum_{X_j\in U/D}\sum_{x\in BND(X_j)}\sum_{k=1}^{m}\left(\lambda_{BP}^{k}\bullet P\left(X_j\middle|[x]_{C_k}\right) + \lambda_{BN}^{k}\bullet P\left(\sim X_j\middle|[x]_{C_k}\right)\right) \tag{18}$$

As discussed above, the cost of making a risky decision for all decision rules can be obtained as:

$$COST_B = COST_{POS} + COST_{NEG} + COST_{BND} \tag{19}$$

Obviously, the higher $COST_B$, the greater the significance of the attribute becomes evident.

Definition 6. Suppose the information system $S = (U, AT = C \cup D, f, V)$, $B \subseteq C$, $a \notin B$, the significance of an attribute is defined as:

$$sig(a, B, D) = COST_{B\cup a}(D) - COST_B(D) \tag{20}$$

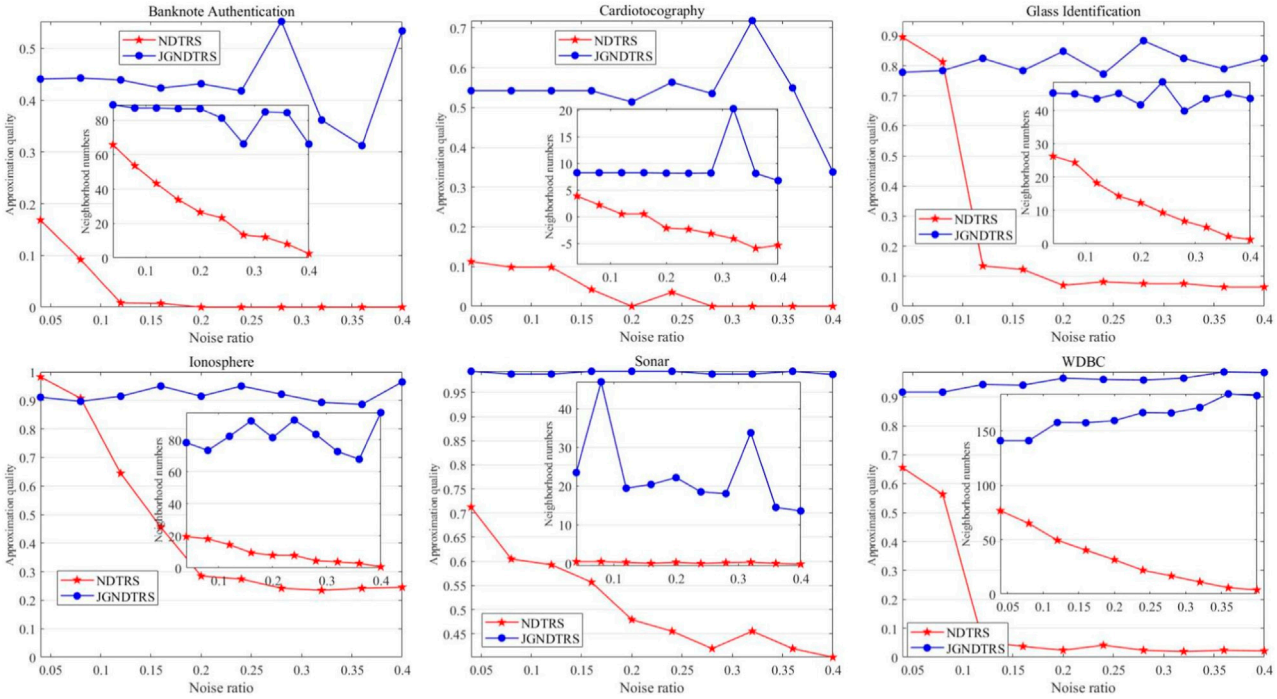


FIGURE 1
Comparison of AQ and NN with different noise ratios.

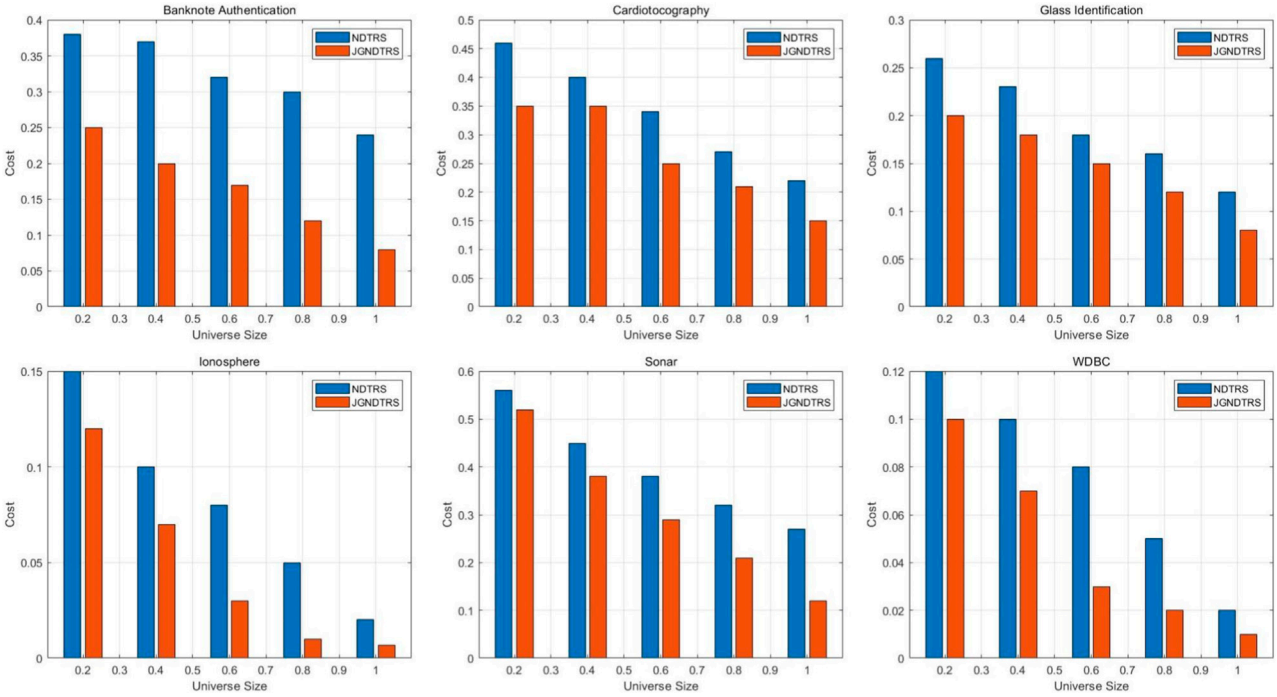


FIGURE 2
The cost of attribute reduction comparison under different universe sizes.

A scheme based on neighborhood decision rough sets is designed for forward search to achieve the optimal reduction. Its specific steps are shown in Algorithm 3.

```
RED = ∅.
For a_i ∈ C − RED
  Calculate sig(a_i, B, D) = COS T_{RED∪a_i}(D) − COS T_RED(D).
End
Select a_k which satisfies sig(a_k, B, D) = max_i(sig(a_i, RED, D)).
If sig(a_k, B, D) > 0
  RED = RED ∪ {a_k}.
Else
  Break.
End
Return RED
```

Algorithm 3. Attribute reduction based on neighborhood decision rough set model.

3.4 Evaluation index

To assess the effectiveness of the suggested approach, this article discusses the following two evaluation indicators: the lower approximation and information granularity.

Approximation quality (*AQ*): Given decision information system $S = (U, AT = C \cup D, f, V)$, $A \subseteq C$, the approximate quality of A relative to D [19] is defined as:

$$\gamma = \frac{\left| \cup \ \delta_A\left(X_q\right) \right|}{|U|}, (q = 1, 2, ..., d) \tag{21}$$

The γ value is expressed as the ratio of the number of objects correctly classified by the conditional attribute set A to the number of all objects in the decision information system. The performance of the proposed granularity description is evaluated in terms of the lower approximation.

Neighborhood number(NN): $x \in U$, suppose $x \in X_q$. $\delta_q(x)$ is the set of data points with decision label q in the neighborhood of x. Therefore, the categories of similar decision label data and different data in the neighborhood can be described as:

$$NN = \sum_{x \in X_q} \left(\left|\delta_q(x)\right| - \sum_{j=1, j \neq q}^{d} \left|\delta_j(x)\right| \right) \tag{22}$$

The larger value of NN indicates that the information granularity provides greater information value to the decision maker and more reasonable granularity.

4 Experiment analysis

In this section, six UCI datasets are utilized to illustrate the feasibility and validity of the suggested methodology. Table 1 describes the relevant information of the datasets.

Parameter setting of PSO algorithm, initialize the particle swarm size to 300, a maximum of 100 iterations is permitted, the individual experience learning factor $c_1 = 1.49445$, the social experience learning factor $c_2 = 1.49445$, the top flight speed of the particle is 0.5 and the allowable error is set to 0.1. For the purpose of assessing the effectiveness of the inertia weight w, consider the use of a linear differential decreasing inertia weight [20], which is expressed as:

$$\frac{dw}{dk} = -\frac{2\left(w_{start} - w_{end}\right)}{T_{\max}^2} \times k \tag{23}$$

$$w(k) = w_{start} - \frac{\left(w_{start} - w_{end}\right)}{T_{\max}^2} \times k^2 \tag{24}$$

where w_{start} represents the initial inertia weight, w_{end} represents the inertia weight when the iteration reaches the maximum number, k represents the current iteration number, and $T_{\max}$ is the maximum iteration number. Set $w_{start} = 0.9$ and $w_{end} = 0.4$.

Figure 1 show the performance of γ and NN respectively. The neighborhood decision rough set model based on reasonable granularity proposed in this paper is abbreviated as JGNDTRS, and NDTRS stands for traditional neighborhood decision rough set. Various noise ratios are represented on the x-axis of each subfigure, which corresponds to a dataset. It can be seen intuitively from the figure that as the noise ratio increases, the approximate quality and NN of NDTRS both show a downward trend. Regarding various noise ratios, the JGNDTRS can obtain the best and relatively stable values of γ and NN in all datasets. Furthermore, JGNDTRS has remarkable performance in identifying anomalous data such as high-density and sparse-density region data points as well as label noise points.

Figure 2 shows the comparison of the cost of JGNDTRS and NDTRS when performing attribute reduction. A dataset is represented by each subplot, and various Universe sizes are shown on the x-axis. Through closer observation, we can conclude that the decision cost of both JGNDTRS and NDTRS shows a decreasing trend as the size of Universe increases. In each dataset, the decision cost of JGNDTRS is always lower than that of NDTRS, regardless of the value of the Universe size. This indicates that JGNDTRS has a superior performance with less cost used in performing attribute reduction.

5 Conclusion

The proposed neighborhood decision rough set model compensates the lack of fault tolerance of classical rough sets. However, there are some challenges in the existing models when dealing with complex data. In this paper, we propose a neighborhood decision rough set model based on justifiable granularity. Firstly, the calculation of rough affiliation is improved according to the number of data points in the neighborhood and the corresponding decision label categories. Secondly, to rectify the original labels, provide pseudo-labels for the noisy data points that are found. A justifiable granularity criterion is introduced and the optimal neighborhood radius is obtained by PSO algorithm. Finally, the risky decision cost is used for attribute reduction. The results of the experiments

demonstrate that the neighborhood decision rough set model based on justifiable granularity has significant performance in identifying abnormal data points and can enhance classification performance. In the future work, the attribute reduction of the neighborhood decision rough set based on justifiable granularity will be further investigated.

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 authors.

Author contributions

The idea was proposed by PG and HJ; XF and XM simulated the algorithm, wrote the paper and polish the English, TC and YS analysed the data designed the experiments. All authors contributed to the article and approved the submitted version.

Funding

This work was supported the National Natural Science Foundation of China under Grant 62006128, Jiangsu Innovation and Entrepreneurship Program.

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

1. Liu J, Lin Y, Du J, Zhang H, Chen Z, Zhang J. Asfs: A novel streaming feature selection for multi-label data based on neighborhood rough set. *Appl Intell* (2023) 53: 1707–24. doi:10.1007/s10489-022-03366-x

2. Wang W, Guo M, Han T, Ning S. A novel feature selection method considering feature interaction in neighborhood rough set. *Intell Data Anal* (2023) 27:345–59. doi:10.3233/IDA-216447

3. Pawlak Z. Rough sets. *Int J Parallel Program* (1982) 11:341–56. doi:10.1007/BF01001956

4. Pawlak Z, Wong S, Ziarko W. Rough sets: Probabilistic versus deterministic approach. *Int J Man Mach Stud* (1988) 29:81–95. doi:10.1016/S0020-7373(88)80032-4

5. Yao Y, Wong S. A decision theoretic framework for approximating concepts. *Int J Man Mach Stud* (1992) 37:793–809. doi:10.1016/0020-7373(92)90069-W

6. Yao Y. Three-way decisions with probabilistic rough sets. *Inf Sci* (2010) 180:341–53. doi:10.1016/j.ins.2009.09.021

7. Qian Y, Liang X, Wang Q, Liang J, Liu B, Skowron A, et al. Local rough set: A solution to rough data analysis in big data. *Int J Approx Reason* (2018) 97:38–63. doi:10.1016/j.ijar.2018.01.008

8. Wang Q, Qian Y, Liang X, Guo Q, Liang J. Local neighborhood rough set. *Knowl Based Syst* (2018) 153:53–64. doi:10.1016/j.knosys.2018.04.023

9. Sun L, Wang L, Ding W, Qian Y, Xu J. Neighborhood multi-granulation rough sets-based attribute reduction using lebesgue and entropy measures in incomplete neighborhood decision systems. *Knowl Based Syst* (2020) 192:105373. doi:10.1016/j.knosys.2019.105373

10. Yang X, Liang S, Yu H, Gao S, Qian Y. Pseudo-label neighborhood rough set: Measures and attribute reductions. *Int J Approx Reason* (2019) 105:112–29. doi:10.1016/j.ijar.2018.11.010

11. Hu Q, Liu J, Yu D. Mixed feature selection based on granulation and approximation. *Knowl Based Syst* (2008) 21:294–304. doi:10.1016/j.knosys.2007.07.001

12. Hu Q, Yu D, Liu J, Wu C. Neighborhood rough set based heterogeneous feature subset selection. *Inf Sci* (2008) 178:3577–94. doi:10.1016/j.ins.2008.05.024

13. Lin Y, Hu Q, Liu J, Chen J, Duan J. Multi-label feature selection based on neighborhood mutual information. *Appl Soft Comput* (2016) 38:244–56. doi:10.1016/j.asoc.2015.10.009

14. Li W, Huang Z, Jia X, Cai X. Neighborhood based decision-theoretic rough set models. *Int J Approx Reason* (2016) 69:1–17. doi:10.1016/j.ijar.2015.11.005

15. Pedrycz W, Homenda W. Building the fundamentals of granular computing: A principle of justifiable granularity. *Appl Soft Comput* (2013) 13:4209–18. doi:10.1016/j.asoc.2013.06.017

16. Wang D, Liu H, Pedrycz W, Song W, Li H. Design Gaussian information granule based on the principle of justifiable granularity: A multi-dimensional perspective. *Expert Syst Appl* (2022) 197:116763. doi:10.1016/j.eswa.2022.116763

17. Cui Y, Meng X, Qiao J. A multi-objective particle swarm optimization algorithm based on two-archive mechanism. *Appl Soft Comput* (2022) 119:108532. doi:10.1016/j.asoc.2022.108532

18. Deng H, Liu L, Fang J, Yan L. The application of SOFNN based on PSO-ILM algorithm in nonlinear system modeling. *Appl Intell* (2023) 53:8927–40. doi:10.1007/s10489-022-03879-5

19. Hu X, Cercone N. Learning in relational databases: A rough set approach. *Comput Intell* (1995) 11:323–38. doi:10.1111/j.1467-8640.1995.tb00035.x

20. Salgotra R, Singh U, Singh S, Mittal N. A hybridized multi-algorithm strategy for engineering optimization problems. *Knowl Based Syst* (2021) 217:106790. doi:10.1016/j.knosys.2021.106790

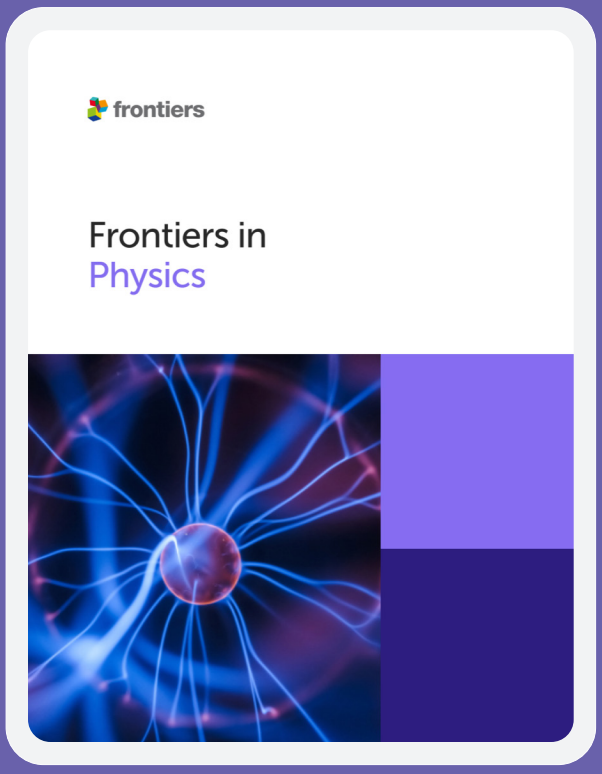
Frontiers in Physics

Investigates complex questions in physics to understand the nature of the physical world

Addresses the biggest questions in physics, from macro to micro, and from theoretical to experimental and applied physics.

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