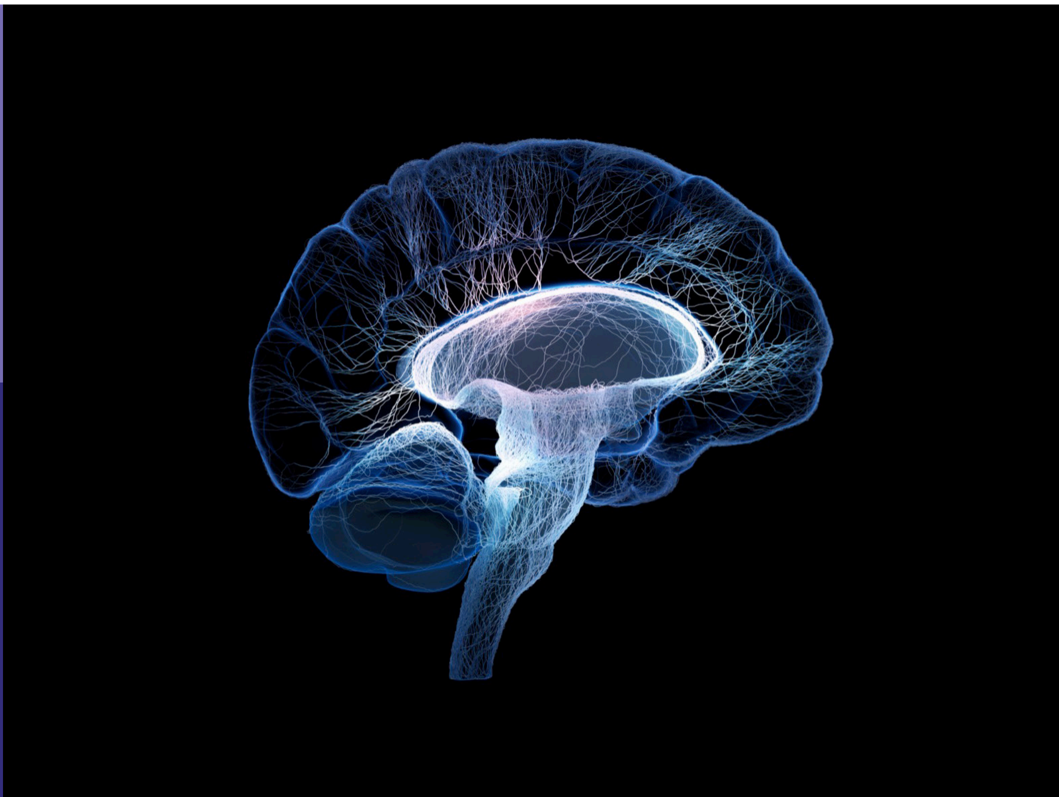
frontiers | Research Topics

Multimodal brain image fusion: Methods, evaluations, and applications

Edited by
Yu Liu, Jiayi Ma, Qiang Zhang, Wei Wei, Xun Chen and Zheng Liu

Published in
Frontiers in Neuroscience



ISSN 1664-8714
ISBN 978-2-83251-388-0
DOI 10.3389/978-2-83251-388-0

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

Multimodal brain image fusion: Methods, evaluations, and applications

Topic editors
Yu Liu — Hefei University of Technology, China
Jiayi Ma — Wuhan University, China
Qiang Zhang — Xidian University, China
Wei Wei — The First Affiliated Hospital of University of Science and Technology of China Anhui Provincial Hospital, China
Xun Chen — University of Science and Technology of China, China
Zheng Liu — University of British Columbia, Canada

Citation
Liu, Y., Ma, J., Zhang, Q., Wei, W., Chen, X., Liu, Z., eds. (2023). *Multimodal brain image fusion: Methods, evaluations, and applications*.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83251-388-0

Table of contents

frontiers | Frontiers in Neuroscience

TYPE Editorial
PUBLISHED 06 January 2023
DOI 10.3389/fnins.2022.1128938

Editorial: Multimodal brain image fusion: Methods, evaluations, and applications

Yu Liu[1]*, Jiayi Ma[2]*, Qiang Zhang[3]*, Wei Wei[4]*, Xun Chen[5]* and Zheng Liu[6]*

[1]Department of Biomedical Engineering, Hefei University of Technology, Hefei, China, [2]Electronic Information School, Wuhan University, Wuhan, China, [3]School of Mechano-Electronic Engineering, Xidian University, Xi'an, China, [4]Department of Radiology, The First Affiliated Hospital of University of Science and Technology of China, Hefei, China, [5]Department of Neurosurgery, The First Affiliated Hospital of University of Science and Technology of China, Hefei, China, [6]School of Engineering, University of British Columbia, Kelowna, BC, Canada

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Vince D. Calhoun,
Georgia State University, United States

*CORRESPONDENCE
Yu Liu
✉ yuliu@hfut.edu.cn
Jiayi Ma
✉ jyma2010@gmail.com
Qiang Zhang
✉ qzhang@xidian.edu.cn
Wei Wei
✉ weiweill@126.com
Xun Chen
✉ xunchen@ustc.edu.cn
Zheng Liu
✉ zheng.liu@ubc.ca

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 21 December 2022
ACCEPTED 28 December 2022
PUBLISHED 06 January 2023

CITATION
Liu Y, Ma J, Zhang Q, Wei W, Chen X and Liu Z (2023) Editorial: Multimodal brain image fusion: Methods, evaluations, and applications. *Front. Neurosci.* 16:1128938. doi: 10.3389/fnins.2022.1128938

KEYWORDS

multimodal medical images, image registration, image fusion, quality assessment, pattern recognition, deep learning

Editorial on the Research Topic
Multimodal brain image fusion: Methods, evaluations, and applications

Multimodal medical imaging is playing an increasingly critical role in the diagnosis and treatment of various brain diseases like glioma, Alzheimer, ischemic stroke, epilepsy, etc. Medical images with different modalities such as computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) focus on different categories of pathological information. Medical image fusion aims to combine the complementary information captured by different imaging modalities for better disease diagnosis and treatment. In recent years, medical image fusion has emerged as a very active topic with various fusion methods being proposed. In addition, the performance evaluation and downstream applications of medical image fusion are also attracting more and more attention. This Research Topic focuses on reporting advanced studies related to multimodal brain image fusion, including image fusion methods, objective evaluation approaches and specific applications in clinical problems. Twelve of the 16 articles submitted to this Research Topic were accepted for publication after a thorough peer-review process. A summary of the key research findings of these works is provided from three aspects as below.

Multimodal brain image registration, fusion and fusion quality evaluation

Image registration is the prerequisite of many medical image processing tasks such as fusion and segmentation. Wang J. et al. proposed a medical image registration method based on the bounded generalized Gaussian mixture model (BGGMM), which can thoroughly describe the joint intensity vector distribution of pixels and highlight image details. The mixture model is formulated based on a maximum likelihood framework, and is solved by an expectation-maximization algorithm. With regard to image fusion methods, Wang A. et al. presented a disentangled representation-based multimodal brain image fusion method *via* group lasso penalty using an auto-encoder-based deep learning framework, aiming to fully exploit the redundancy and complement prior relationships among multimodal source images. A complementary group lasso penalty was designed to promote the disentanglement ability and ensure the complementary feature maps of significant modality information. This study demonstrated that the disentangled representation can improve the interpretability of feature representation, leading to better fusion quality. Zhang et al. proposed a local extreme map guided multimodal brain image fusion method to improve the feature extraction ability of the guided image filter. By iteratively applying this local extreme map guided image filter, the proposed method can extract multiple scales of bright and dark features from the multimodal brain images, and integrate these salient features into one informative fused image. In addition, the proposed scheme can be incorporated with various guided filters or other similar filters in pursuit of improving their feature extraction ability. In comparison to the great attention paid to the study of image fusion methods, few works have explored dedicated quality assessment approaches for medical image fusion. To address this issue, Tang et al. proposed a novel quality assessment method for medial image fusion based on the conditional generative adversarial networks by adopting the mean opinion scores (MOS) of the radiologists as the guiding condition. They demonstrated that their proposed method outperforms several commonly-used quality assessment metrics of image fusion, with excellent agreement with subjective evaluations.

Applications of multimodal brain image fusion

Multimodal medical image fusion has been verified to be of great significance in various related high-level vision tasks such as classification and segmentation. Yi et al. proposed a multimodal classification architecture for the severity diagnosis of glaucoma. The proposed method integrates fundus images and gray scale images of the visual field as the input of the classification model. In addition, they introduced a plug-and-play classifier that adopts the Vision Transformer to extract the global dependencies of images, leading to improved accuracy of the diagnostic task. Li et al. conducted a study to investigate the stage of bi-modal fusion based on EEG and fNIRS for the classification task in hybrid brain-computer interfaces (BCIs). A Y-shaped neural network that fuses the bi-modal information in different stages was proposed. This study demonstrated that the early-stage fusion of EEG and fNIRS have significantly higher performance compared to middle-stage fusion and late-stage fusion. Liu et al. introduced both pixel-level and feature-level medical image fusion techniques for brain tumor segmentation, aiming to achieve more sufficient utilization of multimodal information. They presented a convolutional neural network (CNN)-based 3D pixel-level image fusion network to enrich the input modalities of the segmentation model and designed an attention-based feature fusion module for multimodal feature refinement. Xu et al. proposed a hybrid feature extraction network for medical image segmentation based on CNNs and Transformer. The proposed network can integrate the advantages of Transformer in capturing global contextual information and CNNs in extracting local features. Additionally, a multi-dimensional statistical feature extraction module was designed to strengthen low-dimensional texture features and enhance the segmentation performance. Tian et al. presented a method to combine light sheet microscopy (LSM) data with magnetic resonance histology (MRH) of the same specimen, with the aim of restoring the morphology of the LSM images to the in-skull geometry. They developed an image processing pipeline to restore the correct brain morphology of 3-dimensional cleared or stained mouse brain by registering the cleared brain data to MRH of the same specimen. Peng et al. introduced the minimally invasive puncture and drainage (MIPD) surgery using mixed reality holographic navigation technology (MRHNT) *via* integrating the holographic image and the real head. By wearing mixed reality holographic equipment, the precise location of intracranial hematomas, tumors, ventricles, and other structures with the perspective function can be understood, laying a theoretical foundation for implementation in neurosurgery.

Joint analysis of multimodal data

Mononen et al. conducted a study to evaluate the variability among tasks of magnetoencephalography (MEG)-functional magnetic resonance imaging (fMRI) relationship using data recorded during three distinct naming tasks from the same set of participants. The results demonstrated that the MEG-fMRI correlation pattern varies according to the performed task. In addition, the electromagnetic-hemodynamic correlation could serve as a more sensitive proxy for task-dependent neural engagement in cognitive tasks than isolated within-modality measures. Gallego-Rudolf et al. characterized the impact of the ballistocardiographic (BCG) artifact on resting-state EEG spectral properties and compared the effectiveness of seven common BCG correction methods to preserve EEG spectral features. They also assessed if these methods retained posterior alpha power reactivity to an eyes closure-opening task and compared the results from EEG-informed fMRI analysis using different BCG correction approaches.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Acknowledgments

We appreciate the contributions of all the authors, the reviewers, and the editorial board members for contributing to this Research Topic. We hope that this Research Topic can make a broad impact on the study of multimodal brain image fusion including methods, evaluations and applications.

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.

ORIGINAL RESEARCH
published: 11 March 2022
doi: 10.3389/fnins.2022.850179

Application of Fused Reality Holographic Image and Navigation Technology in the Puncture Treatment of Hypertensive Intracerebral Hemorrhage

Chen Peng[1†], Liu Yang[1†], Wang Yi[2], Liang Yidan[1], Wang Yanglingxi[1], Zhang Qingtao[1], Tang Xiaoyong[1], Yongbing Tang[1], Wang Jia[1], Yu Xing[1], Zhu Zhiqin[3] and Deng Yongbing[1]*

[1] Department of Neurosurgery, Chongqing Emergency Medical Center, Chongqing University Central Hospital, Chongqing, China, [2] QINYING Technology Co., Ltd., Chongqing, China, [3] College of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China

Objective: Minimally invasive puncture and drainage (MIPD) of hematomas was the preferred option for appropriate patients with hypertensive intracerebral hemorrhage (HICH). The goal of our research was to introduce the MIPD surgery using mixed reality holographic navigation technology (MRHNT).

Method: We provided the complete workflow for hematoma puncture using MRHNT included three-dimensional model reconstruction by preoperative CT examination, puncture trajectory design, immersive presentation of model, and real environment and hematoma puncture using dual-plane navigation by wearing special equipment. We collected clinical data on eight patients with HICH who underwent MIPD using MRHNT from March 2021 to August 2021, including the hematoma evacuation rate, operation time, deviation in drainage tube target, postoperative complications, and 2-week postoperative GCS.

Result: The workflow for hematoma puncture using MRHNT were performed in all eight cases, in which the average hematoma evacuation rate was $47.36 \pm 9.16\%$, the average operation time was 82.14 ± 15.74 min, and the average deviation of the drainage tube target was 5.76 ± 0.80 mm. There was no delayed bleeding, acute ischemic stroke, intracranial infection, or epilepsy 2 weeks after surgery. The 2-week postoperative GCS was improved compared with the preoperative GCS.

Conclusion: The research concluded it was feasible to perform the MIPD by MRHNT on patients with HICH. The risk of general anesthesia and highly professional holographic information processing restricted the promotion of the technology, it was necessary for technical innovation and the accumulation of more case experience and verification of its superiority.

Keywords: hypertensive intracerebral hemorrhage, minimally invasive puncture and drainage, mixed reality, navigation, deviation

OPEN ACCESS

Edited by:
Yu Liu,
Hefei University of Technology, China

Reviewed by:
Xin Jin,
Yunnan University, China
Weilin Xu,
Zhejiang University, China
Chenju Yi,
Sun Yat-sen University, China

Correspondence:
Deng Yongbing
1422495819@qq.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *07 January 2022*
Accepted: *08 February 2022*
Published: *11 March 2022*

Citation:
Peng C, Yang L, Yi W, Yidan L, Yanglingxi W, Qingtao Z, Xiaoyong T, Tang Y, Jia W, Xing Y, Zhiqin Z and Yongbing D (2022) Application of Fused Reality Holographic Image and Navigation Technology in the Puncture Treatment of Hypertensive Intracerebral Hemorrhage. Front. Neurosci. 16:850179. doi: 10.3389/fnins.2022.850179

1. INTRODUCTION

Stroke has become the leading cause of death in China. Hypertensive intracerebral hemorrhage (HICH) is one of the most serious complications of hypertension, with an incidence of 19–48% of strokes in China, and the high disability and mortality rates of HICH lead to a heavy social burden (Zhou et al., 2019). At present, there is no evidence for the optimal surgical treatment of HICH with surgical indications. MISTIE research demonstrated the safety profile of the minimal invasive surgery procedure revealed clot size reduction could be achieved with similar safety to standard medical treatment (Hanley et al., 2016, 2019).

The precise puncture of hematomas is the key to the success of surgeries, and the methods used include the "blind" method, which uses a freehand technique according to CT images combined with skull anatomical marks, CT-guided (Wang et al., 2009) and image-guided (Yang et al., 2014; Sun et al., 2016) puncture methods and the neuronavigation system (Chartrain et al., 2018) puncture method. However, all the above puncture methods have shortcomings, such as inaccuracy, expensive, non-portable, bulky hardware. It is important to find a more convenient, visualized, rapid, and precise puncture method.

Mixed reality has been developed based on virtual reality and augmented reality technologies. By processing holographic images, mixed reality provides virtual images and information in the real environment and provides users with immersive feelings. Users can obtain real and virtual image information at the same time by wearing special equipment (Microsoft, HoloLens) and interact with holographic images in the display environment according to their own commands. With this technology, neurosurgeons can first construct intracerebral hemorrhages and design the puncture trajectory. During surgery, the location and morphology of a hematoma can be observed from multiple angles, and precise puncture can be performed with the help of navigation.

Several studies on glioma, meningioma, intracranial aneurysm have shown that MR technology could implement a safe, effective, and minimally invasive individualized operation plan, evaluate the operation risk, and protect the tissue structure during the operation (Kockro et al., 2016; Incekara et al., 2018; Qi et al., 2021; Zhang et al., 2021). There are no reports on the application of MR technology to hematoma puncture in patients with HICH. In this research, we introduce the MIPD surgery using mixed reality holographic navigation technology (MRHNT). We provide the complete workflow, show the clinical data and results, shar our practical experience in hematoma puncture using MRHNT, and verify the accuracy and feasibility of the application of this technology.

In this research, we introduced a precise MIPD method in different parts of HICHs using mixed reality holographic navigation technology (MRHNT). We provided the complete workflow, showed the clinical data and results, shared our practical experience in hematoma puncture using MRHNT, and verified the accuracy and feasibility of the application of this technology.

2. MATERIALS AND METHODS

2.1. Clinical Datae

From March 2021 to August 2021, approved by the ethics committee of Chongqing Emergency Medical Center, HICH patients treated with MIPD by mixed reality holographic navigation technology were involved in this research. All patients signed the surgical informed consent form. Partially in accordance with the MISTIE study the inclusion criteria were following: patients with non-traumatic (spontaneous) ICH not due to a macrovascular cause such as an aneurysm or AVM were involved. All patients signed the surgical informed consent form. All patients age was 18–80 years old with GCS score ≥14 or NIHSS score ≤6, whose ICH remained the same size for at least 6 h after diagnostic CT. Our surgery involved patients with both supratentorial and supratentorial hemorrhage, with supratentorial hematoma volume of 30–50 ml, cerebellar hematoma volume of 10–15 ml, with brainstem hematoma volume of 5–10 ml. The exclusion criteria were as follows: patients with cerebral herniation due to HICH, severe cardiopulmonary disease, or coagulopathy, other patients who cannot tolerate general anesthesia, and patients with family members who refused surgery by mixed reality holographic navigation technology. We analyzed eight patients based on their preoperative and postoperative hematoma volume, hematoma evacuation rate, operation time, blood loss, deviation in drainage tube target (the distance between the tip of the drainage tube and the designed puncture trajectory target), 2-week rebleeding rate4, postoperative complications, and preoperative and 2-week post-operative GCS.

2.2. Preoperative CT Examination and Design Puncture Trajectory

All patients were required to undergo head CT examination before the operation. Patients were examined by placing three sticky analysis markers around the puncture area. After anesthesia, a bone nail was drilled through the hole in the sticky marker base, and a sterilized analysis marker was placed; these two markers were the "twin marker" and ensured no obvious deviation in the location of the markers. CT data were collected by a 64-slice CT scanner (Lightspeed VCT 6, General Electric Company, USA). Image parameters included exposure (3 mAS), thickness (5 mm), and image size (512,512). Based on hospital PACS, DICOM format data were imported into Medical Modeling and Design System software for reconstruction of the head model. Three-dimensional reconstructions were focused on the skull, hematoma, nose, and ears during head model building. Preoperative hematoma volume was measured by Medical Modeling and Design System software. According to the reconstructed head model, the puncture skull location and the hematoma target were planned to design a puncture trajectory. The designed puncture trajectory has the same diameter as the actual puncture needle. The depth of the designed puncture trajectory was also measured.

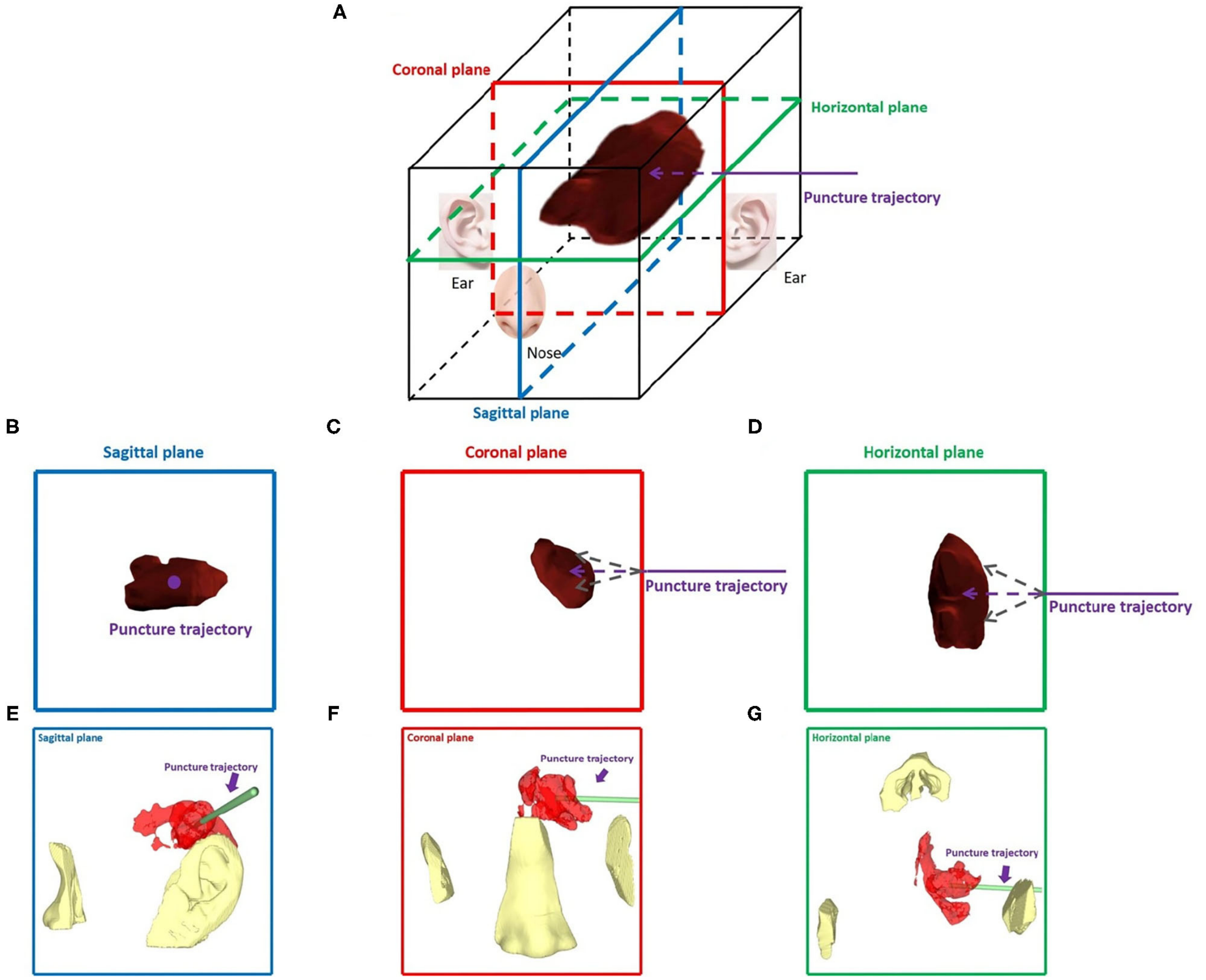


FIGURE 1 | Dual-plane navigation puncture. **(A)** The head was considered a six-sided cube with horizontal, sagittal, and coronal planes. According to the hematoma puncture trajectory designed before surgery, the puncture angle and depth were observed from two planes. **(B–G)** For example, a hematoma was punctured at the basal ganglia from the temporal region, and hematoma and puncture trajectories were observed in the sagittal, coronal, and horizontal planes, respectively. **(B–D)** Theoretical images of different planes. **(E–G)** Wearing HoloLens, images of the different planes were presented by adjusting the locations of the nose and ear.

2.3. Registration of Holographic Images

After anesthesia, three-dimensional coordinate locations data of the calibration plate, puncture needle, and three markers in head were captured by camera. We matched the preoperative reconstructed head mode with the coordinate data by MAYA software, bond location of the corresponding skull, hematoma, nose, and ears by analysis markers and imported the matched information into Microsoft HoloLens. The camera captured dynamic changes in the analysis marker location of the head and puncture needle, synchronizing holographic models with tracking software. This procedure took ∼40 min.

2.4. Dual-Plane Navigation Puncture

Innovatively, since a double-arm digital subtraction angiography device can observe vascular morphology from two angles, we considered the head to be a six-sided cube with horizontal, sagittal, and coronal planes. If the puncture trajectory was perpendicular to a plane, the other two planes could be observed to evaluate the deviation in the puncture trajectory from the horizontal and vertical directions. For example, when puncturing a hematoma at the basal ganglia from the temporal region, the puncture trajectory was perpendicular to the sagittal plane, and we observed the vertical and horizontal deviation between the puncture needle and the designed puncture trajectory from the coronal plane and horizontal plane. When wearing mixed reality holographic equipment, the three-dimensional sense of the space will be more obvious. According to the locations of the nose and ear, gestures such as rotation and movement were used to adjust the plane angle and location, and then, the image was locked. After the image was locked, the holographic image could not change due to gestures, making this method more convenient for the

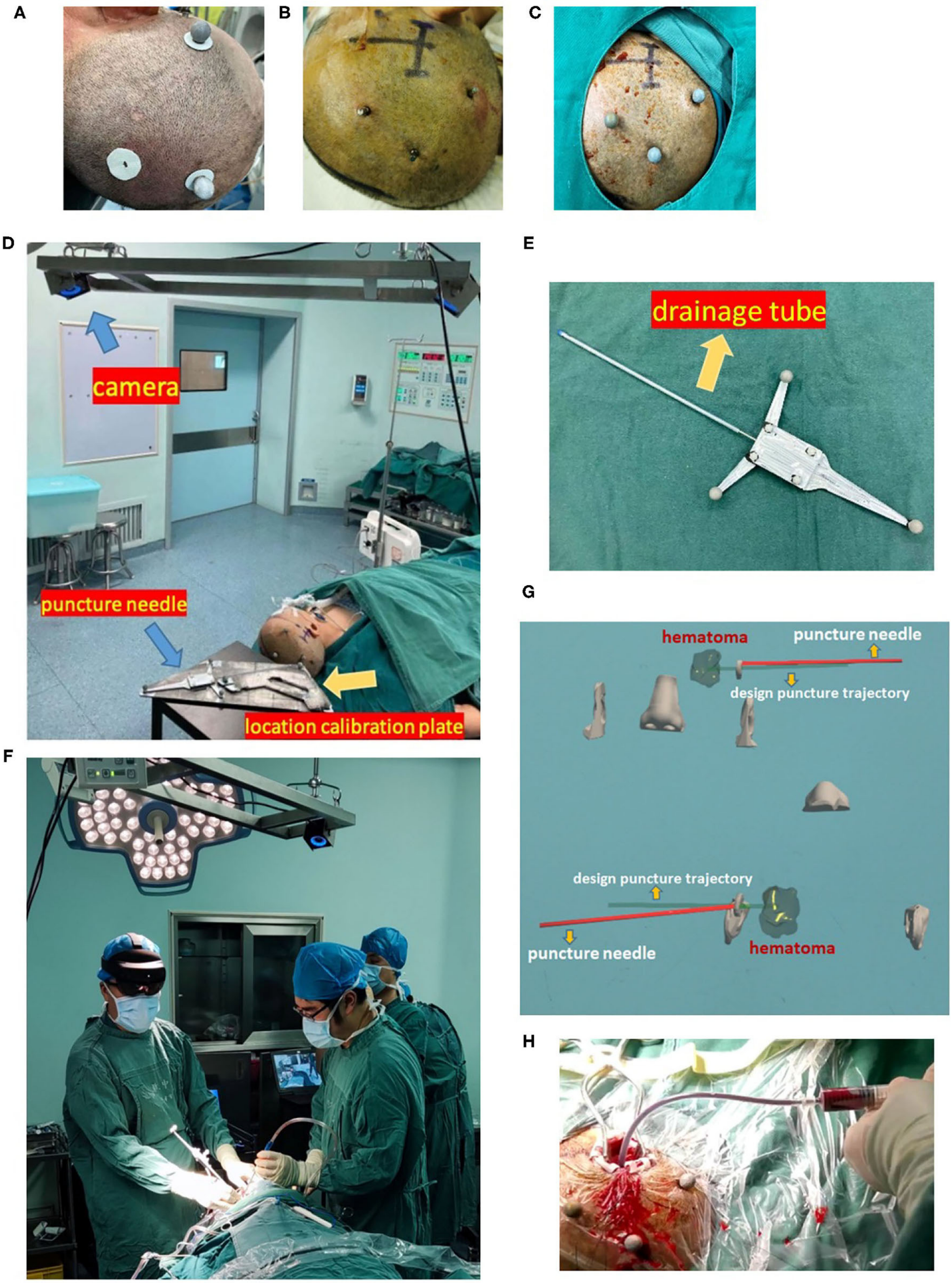


FIGURE 2 | Workflow of the minimally invasive puncture and drainage of hypertensive intracerebral hemorrhages by mixed reality holographic navigation technology. **(A)** Patients wore three sticky analysis markers around the puncture area. **(B)** A bone nail was drilled through the hole in the sticky marker base, which was replaced with sticky analysis markers, and these "twin marker" ensured no obvious deviation in the location of markers. **(C)** After disinfection, a sterilized analysis marker was installed on the bone nail. **(D)** Three-dimensional coordinate data of the location calibration plate, puncture needle, and head were captured by a camera. **(E)** Combination of a puncture needle and drainage tube. **(F)** Wearing HoloLens, the surgeon used MRHNT for hematoma puncture. **(G)** Wearing HoloLens, the surgeon actually viewed the two planes of the image. **(H)** After hematoma puncture was completed, the hematoma was aspirated from the drainage tube.

TABLE 1 | Demographic and clinical data of eight patients in the research.

Cases	Age (years), Gender	Hematoma location	Pre-operative volume (ml)	Post-operative volume (ml)	HER (%)	Deviation (mm)
1	69, M	Temporal lobe	32.17	17.23	46.44	7.08
2	47, M	Basal ganglia	33.10	12.48	62.30	5.62
3	37, M	Brainstem	5.45	3.18	41.65	4.22
4	69, M	Basal ganglia	30.34	18.81	38.00	6.04
5	43, M	Basal ganglia	31.69	19.41	38.75	5.46
6	44, M	Basal ganglia	37.22	15.32	58.84	5.87
7	44, M	Brainstem	6.32	3.12	50.63	6.13
8	67, F	Basal ganglia	35.22	20.34	42.25	5.66

puncture operation of both hands. We illustrated this method in **Figure 1**.

2.5. Surgical Procedure

The doctor wore mixed reality holographic equipment to observe the precise locations of the skull, hematoma, nose, and ears, designed puncture trajectory and actual puncture needle. According to the designed puncture trajectory, we performed skin incision and skull drilling, performed hematoma puncture according to the above dual-plane navigation puncture technology, aspirated the hematoma, retained the drainage tube, and sutured the skin. When the operator observed the puncture needle entering the hematoma target, removed the puncture needle, retained drainage tube, and connected with a 10 ml syringe to aspirate until there was no longer any fluid component of the clot. The drainage tube was tunneled subcutaneously, and connected to closed drainage system. We performed postoperative head CT examination, but did not inject rtPA or other drugs in the drainage tube as in the MISTIE study, and kept the drainage tube in low drainage for 48 h and then removed it. We provide the complete workflow of this technology in **Figure 2**.

2.6. Follow-Up Imaging and Accuracy Assessment

Head CT examination was performed immediately or 1 day after surgery. Postoperative hematoma volume was measured by a non-operator, as described above. Hematoma evacuation rate = (pre-operative hematoma volume- post-operative hematoma volume)/pre-operative hematoma volume. Accuracy assessment was defined as the deviation between the drainage tube and the planned puncture hematoma target. The deviation calculation used BLENDER2.93.3 software, which used the 3D XYZ coordinate system to visualize the deviation between the drainage tube and the target point (points 0, 0, and 0).

2.7. Statistical Analysis

Quantitative data were presented as means ± SDs. The paired t-test was used to compare the difference between the preoperative and postoperative hematoma volumes and GCS. All statistical analyses were performed using SPSS version 21 (IBM SPSS Statistics for Macintosh, IBM Corp). In all cases, a $p < 0.05$ was considered statistically significant.

3. RESULTS

From March 2021 to August 2021,8 patients with HICHs were treated with MIPD by mixed reality holographic navigation technology, including five males and three females with an average age of 52.5 ± 13.42 years (range, 37–69 years). The hematoma was located in the basal ganglia in five cases, in the brainstem in two cases and in the temporal lobe in one case. Among six patients with supratentorial hematoma, four cases of postoperative hematoma were more than 15 ml, and 1 case was more than 20 ml, with the average post-operative hematoma was 17.3 ml. The average hematoma evacuation rate in eight patients was 47.36 ± 9.16 %. There were statistically significant differences in the pre-operative and post-operative hematoma ($P = 0.002$) volumes. All operations were performed under general anesthesia, the average operation time was 82.14 ± 15.74 min, and the average intraoperative blood loss was 36.28 ± 8.14 ml. By double-plane MRHNT, the average deviation in the drainage tube target was 5.76 ± 0.80 mm. There was no delayed bleeding, acute ischemic stroke, intracranial infection, or epilepsy 2 weeks after surgery. The average preoperative GCS was 9.25 ± 2.05, while the 2-week postoperative GCS was 11.00 ± 2.39. The 2-week postoperative GCS was improved, but it was not statistically significant ($P = 0.26$) compared with the preoperative GCS. A summary of demographic and clinical characteristics is provided in **Table 1**.

4. TECHNOLOGY ADVANTAGES

Wearing sticky analysis markers for preoperative CT examination could greatly shorten the time for holographic image registration. Mixed reality holographic image technology succeeded in creating stereoscopic sensations of the skull, hematoma, designed puncture trajectory, and actual puncture needle, and this immersive holographic environment was difficult to fully express through photos or videos or even AR technology. In addition, mixed reality holographic images were transferred to the screen in real time, allowing observers without experience to share the same view with the surgeon. Innovatively, dual-plane navigation puncture technology was used with camera monitoring, which allowed the head to move. After adjusting the two observation puncture planes by gestures, the holographic image was locked to avoid image changes caused

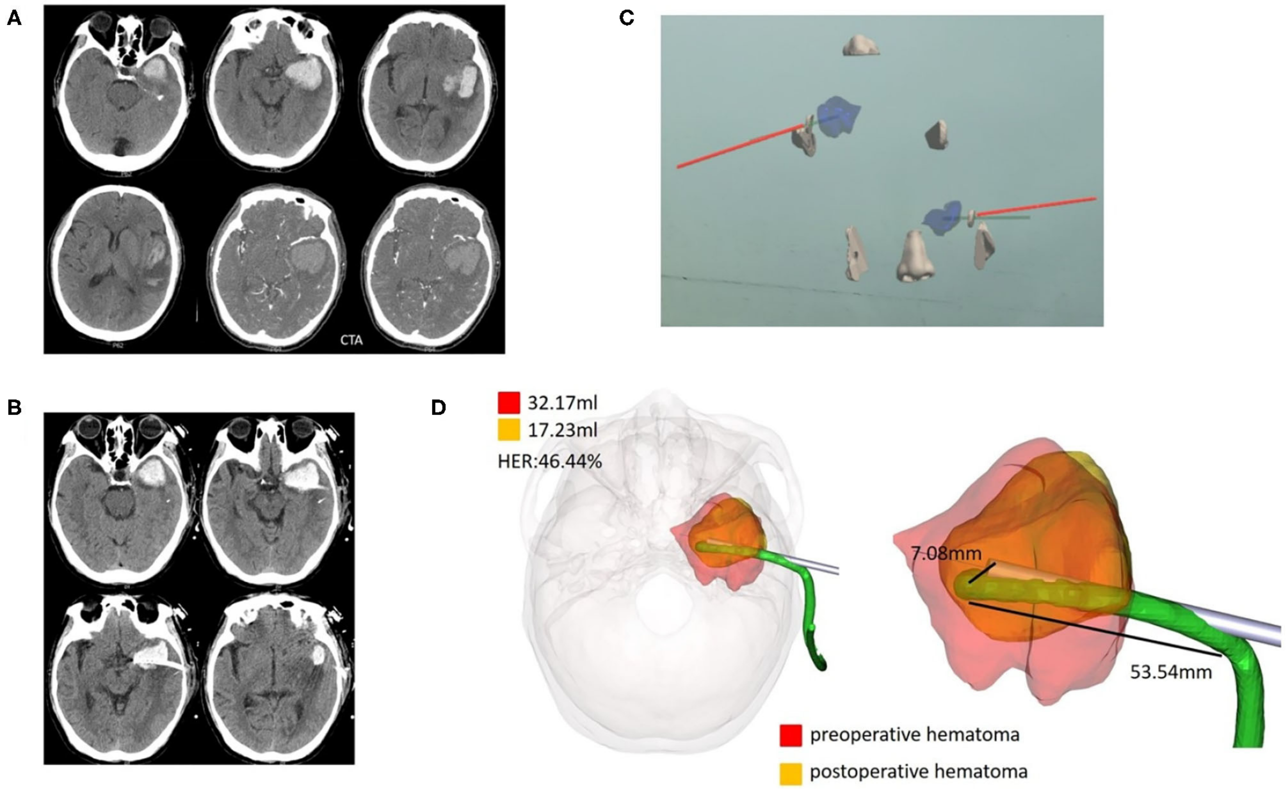


FIGURE 3 | Case 1, a 69-year-old male patient diagnosed with HICH in the left temporal lobe. **(A)** Preoperative CT showed a HICH in the left temporal lobe, excluding aneurysms, and arteriovenous malformation. **(B)** Postoperative follow-up CT. **(C)** Wearing HoloLens, the coronal, and horizontal planes were adjusted for puncture through the ear and nose locations. **(D)** For fusion of the preoperative and postoperative three-dimensional reconstruction of hematomas, the preoperative hematoma volume was 32.17 ml, the postoperative hematoma volume was 17.23 ml, and the hematoma evacuation rate was 46.44%. The length of the intracranial drainage tube was 53.54 mm, and the deviation in the drainage tube target was 7.08 mm. HER, hematoma evacuation rate.

by gestures. We used both hands to hold the head and tail of the puncture needle to enhance the stability of the puncture. We observed the puncture direction with dynamic navigation from two planes to better control the puncture deviation. We obtained a puncture deviation of 5.76 ± 0.80 mm, which was perfectly acceptable for a hematoma volume of ∼30 ml. Representative cases are presented in **Figures 3–5**.

5. DISCUSSION

In patients with HICH, MIPD of hematomas was the preferred option for appropriate patients. Surgery first required the localization of the hematoma, whereas hematomas in HICH did not require millimeter accuracy, and experienced neurosurgeons could successfully puncture the hematoma with various localization methods. Stereotactic devices and neuronavigation systems seemed overqualified for hematoma localization and were not available in many hospitals. Additionally, neuronavigation system also has the disadvantages of expensive, non-portable, and with bulky hardware. However, various puncture methods mainly rely on personal experience, and disadvantages include poor accuracy, a high failure rate, and difficulty in ensuring the homogeneity of the puncture location, which affects the surgical effect and increases the surgical risk.

Mixed reality technology integrates holographic image information into the real world by a computer. The real environment and virtual images could be spliced in the same field of view in real time for a three-dimensional display. Owing to advancements in holographic information transmission and processing. Our solution of using professional software, we take CT image and realistic environment information captured by camera to reconstruct, match, and generate holographic images. The method did not require additional cost or technical complexity, other than few professional software. This technology rendered MRHNT as much more convenient, affordable, portable, and popular.

We acquired head information using the technology to achieve holographic image and real head integration. By wearing mixed reality holographic equipment, we could understand the precise location of intracranial hematomas, tumors, ventricles, and other structures with the perspective function, which laid a theoretical foundation for implementation in neurosurgery. This technology should have a promising future in medicine, but it is still in its infancy and is in the initial stage; its application in neurosurgery has rarely been reported (Zhang et al., 2021).

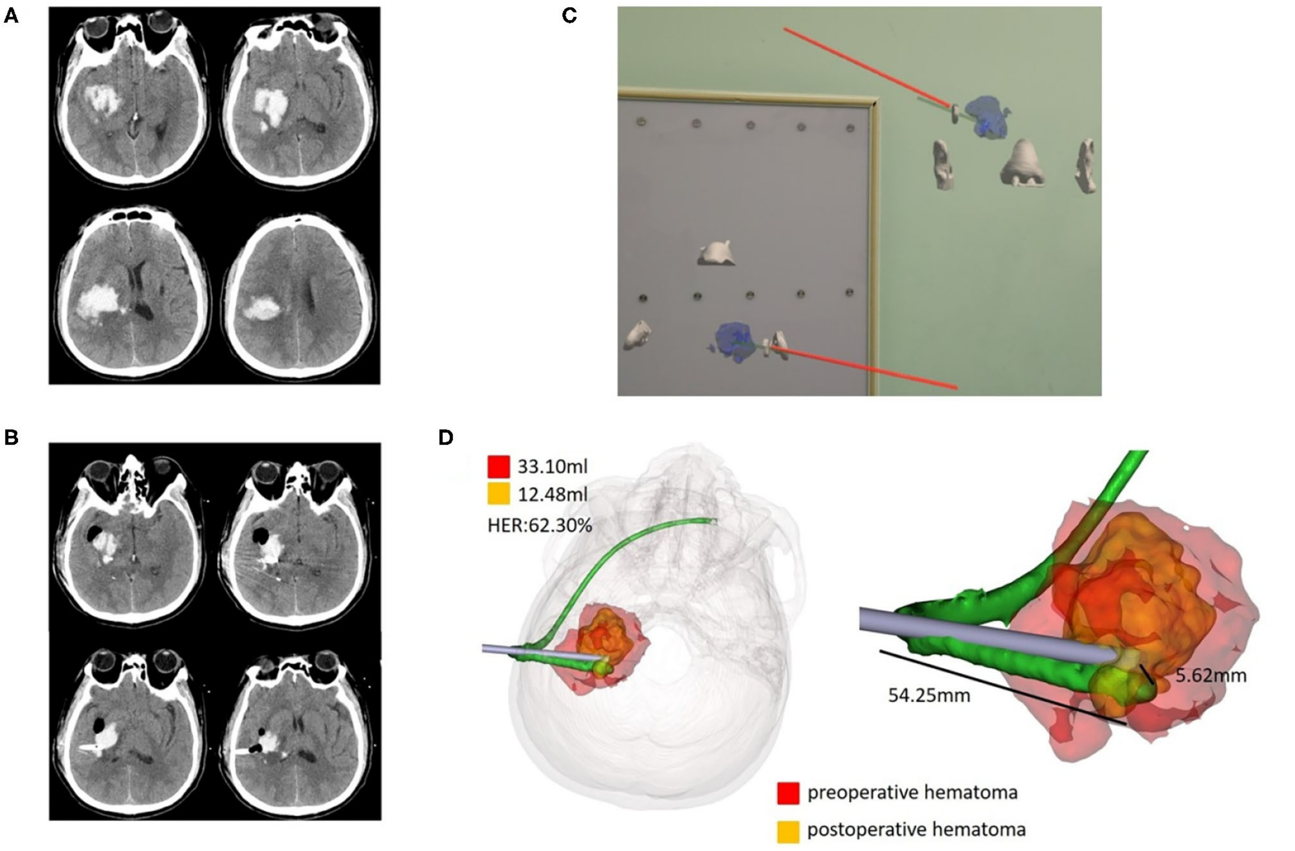


FIGURE 4 | Case 2, a 47-year-old male patient diagnosed with HICH in the right basal ganglia. **(A)** Preoperative CT showed HICH in the right basal ganglia. **(B)** Postoperative follow-up CT. **(C)** Wearing HoloLens, adjust the coronal, and horizontal planes for puncture through the ear and nose location. **(D)** Fusion of the preoperative and postoperative three-dimensional reconstructions of the hematoma. The preoperative hematoma volume was 33.10 ml, the postoperative hematoma volume was 12.48 ml, and the hematoma evacuation rate was 62.30%. The length of the intracranial drainage tube was 54.25 mm, and the deviation of the drainage tube target was 5.62 mm. HER, hematoma evacuation rate.

In the preliminary work, our team rigidly matched the holographic image processed by the computer with the head, visualized the ventricular structure, intuitively guided ventricular puncture operation, and improved the puncture accuracy compared with that of the traditional method. Moreover, mixed reality technology played a very helpful role in finding foreign bodies and locating hematomas in patients with traumatic brain injury. We reported a case of the localization of the intracranial nail and hematoma by mixed reality technology, which helped us to design the surgical incision rationally and avoid secondary injury caused by blind exploration (Li et al., 2018, 2021).

There were obvious shortcomings in the previous method, including low registration speed and rigid integration of the hologram to the head to avoid movement of the head. In particular, preliminary technology was not truly navigational; when the puncture needle was drilled into the skull, it could not be tracked.

To solve the above problems, we made several improvements as follows. First, patients wore three sticky analysis markers for head CT examination before the operation. We replaced sticky analysis markers with "twin marker" to ensure no obvious deviation in the location of markers. Then, the three-dimensional coordinate data were captured by a camera, which could shorten the holographic image registration time. Second, we abandoned the rigid matching of the holographic image and head by the eye. Alternatively, the camera captured dynamic changes in the analysis marker location of the head and puncture needle, synchronizing holographic images. This meant that even if the head location changed, the holographic image would change accordingly by analyzing the marker space distance through the camera. Third, we observed the puncture direction from dynamic navigation from the two planes to better control puncture deviation.

The results of eight patients with HICHs treated with MIPD by mixed reality holographic navigation technology revealed that the operation time and blood loss were acceptable. The hematoma evacuation rate was 47.36 ± 9.16%, the average of supratentorial postoperative hematoma volume in six patients in our research was 17.3 ml. According to the results of the MISTIE study revealed reduction in clot size to 15 ml or less was associated with functional improvement. Although GCS score improved 2 weeks after surgery, this result was not comparable

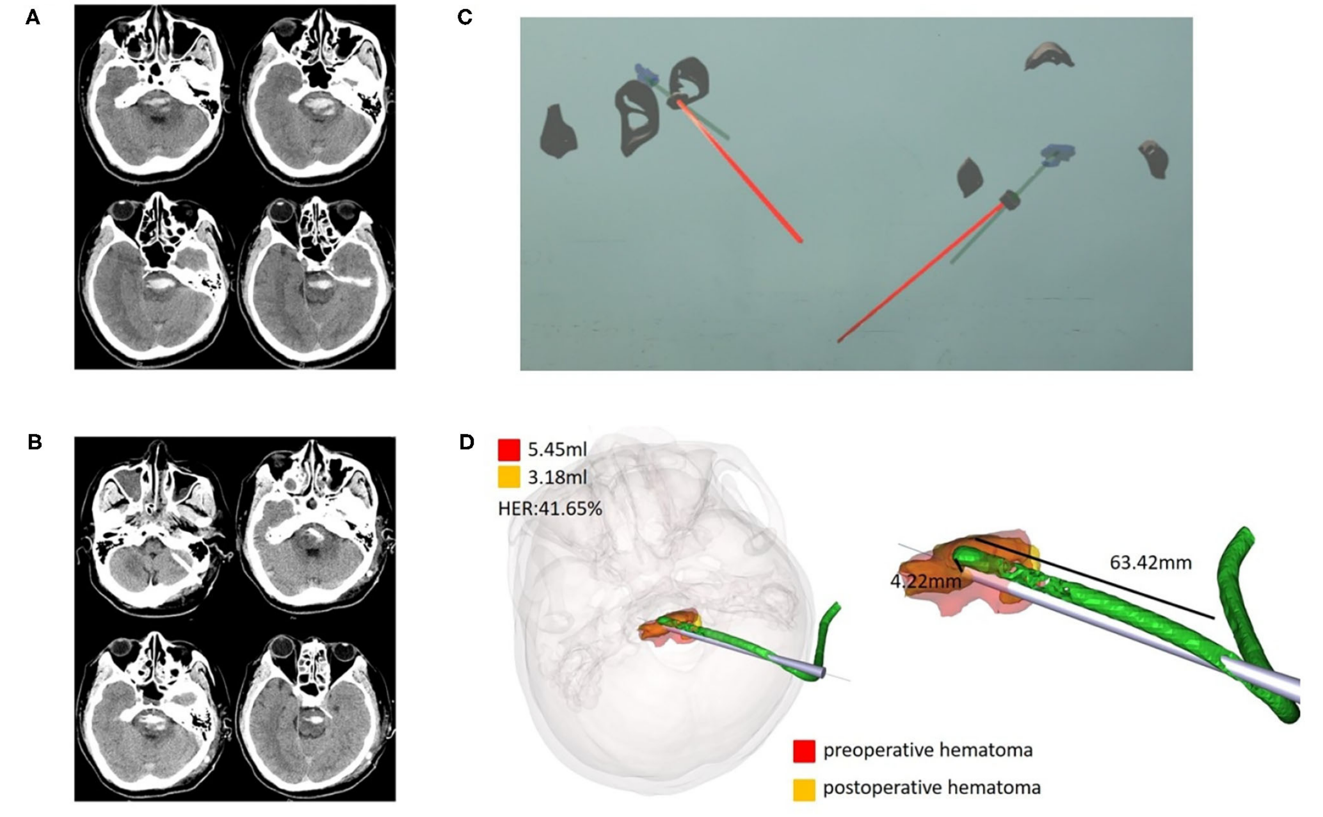


FIGURE 5 | Case 3, a 37-year-old male patient diagnosed with HICH in the brainstem. **(A)** Preoperative CT showed HICH in the brainstem. **(B)** Postoperative follow-up CT. **(C)** Wearing HoloLens, the sagittal and horizontal planes were adjusted for puncture through the ear and nose locations. **(D)** Fusion of the preoperative and postoperative three-dimensional reconstructions of the hematoma. The preoperative hematoma volume was 5.45 ml, the postoperative hematoma volume was 3.18 ml, and the hematoma evacuation rate was 41.65%. The length of the intracranial drainage tube was 63.42 mm, and the deviation in the drainage tube target was 4.22 mm. HER, hematoma evacuation rate.

with MISTIE study, considering the small number of cases, not much preoperative hematoma volume (30–40 ml), and the absence of control group.

At present, most precise and popular of the previous methods were neuronavigation systems. van Doormaal et al. (2019) reported compared to the mean fiducial registration error of conventional neuronavigation was 3.6 mm, the mean fiducial registration error of holographic neuronavigation was 4.4 mm in three patients. Other researches have shown that the navigation deviation using mixed reality holographic navigation was 4–6 mm (Incekara et al., 2018; Li et al., 2018; McJunkin et al., 2018), which was consistent with our results. We obtained a puncture deviation of 5.76 ± 0.80 mm, which was perfectly acceptable for a hematoma volume of ∼30 ml.

To improve the puncture accuracy, we have the following suggestions: 1. When drilling the skull, try to drill a larger hole, and make the puncture needle coincide exactly with the designed puncture trajectory. 2. Hold the head and tail of the puncture needle with both hands, and adjust the puncture direction horizontally and vertically at any time. 3. Mark the puncture needle with depth, and determine the puncture depth by holographic image navigation.

Our research has some limitations. At present, we have relatively few cases, so there are not enough data to verify the advancement of the technology. The new technology bears the risk of general anesthesia and takes a long time for surgery, which might make many surgeons relatively apathetic about this technology. Improving the accuracy of the puncture also requires the surgeon to spend much time in the model for consistent practice.

We believe that as science and technology drive the accelerated progress of medicine, surgical procedure will be more simplified, and new equipment and methods will be developed to improve puncture accuracy. Fusion with MRI images with white matter cellulose information to design the optimal puncture trajectory (Liu et al., 2020; Zheng et al., 2020; Zhu et al., 2021), and accumulation of more cases experience and verification of its superiority.

6. CONCLUSION

With MRHNT, neurosurgeons can first construct three-dimensional model and design the puncture trajectory. During surgery, the location and morphology of the hematoma can be observed from multiple angles, and precise puncture can be performed with the help of dual-plane navigation. The risk of general anesthesia and highly professional holographic information processing restrict the promotion of the technology, it is necessary for technical innovation and the accumulation of more case experience and verification of its superiority.

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/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of Chongqing Emergency Medical Center. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

DY and CP participated in the design and coordination of the study and drafted the manuscript. LYi and WYa participated in the clinical evaluation of the patients. WYi and ZZ interpreted data reconstruction and image fusion. CP, LYa, TX, and ZQ performed the minimally invasive surgery. WJ and YX performed the statistical analysis. All authors read and approved the final manuscript.

FUNDING

This study was financially supported by the Medical Research Project of Science and Technology Bureau Joint Health Commission, Chongqing, China (2020MSXM078).

REFERENCES

Chartrain, A. G., Kellner, C. P., Fargen, K. M., Spiotta, A. M., Chesler, D. A., Fiorella, D., et al. (2018). A review and comparison of three neuronavigation systems for minimally invasive intracerebral hemorrhage evacuation. *J. Neurointervent. Surg.* 10, 66–74. doi: 10.1136/neurintsurg-2017-013091

Hanley, D. F., Thompson, R. E., Muschelli, J., Rosenblum, M., McBee, N., Lane, K., et al. (2016). Safety and efficacy of minimally invasive surgery plus alteplase in intracerebral haemorrhage evacuation (MISTIE): a randomised, controlled, open-label, phase 2 trial. *Lancet Neurol.* 15, 1228–1237. doi: 10.1016/S1474-4422(16)30234-4

Hanley, D. F., Thompson, R. E., Rosenblum, M., Yenokyan, G., Lane, K., McBee, N., et al. (2019). Efficacy and safety of minimally invasive surgery with thrombolysis in intracerebral haemorrhage evacuation (MISTIE III): a randomised, controlled, open-label, blinded endpoint phase 3 trial. *Lancet* 393, 1021–1032. doi: 10.1016/S0140-6736(19)30195-3

Incekara, F., Smits, M., Dirven, C., and Vincent, A. (2018). Clinical feasibility of a wearable mixed-reality device in neurosurgery. *World Neurosurg.* 118, e422–e427. doi: 10.1016/j.wneu.2018.06.208

Kockro, R. A., Killeen, T., Ayyad, A., Glaser, M., Stadie, A., Reisch, R., et al. (2016). Aneurysm surgery with preoperative three-dimensional planning in a virtual reality environment: technique and outcome analysis. *World Neurosurg.* 96, 489–499. doi: 10.1016/j.wneu.2016.08.124

Li, Y., Chen, X., Wang, N., Zhang, W., Li, D., Zhang, L., et al. (2018). A wearable mixed-reality holographic computer for guiding external ventricular drain insertion at the bedside. *J. Neurosurg.* 131, 1599–1606. doi: 10.3171/2018.4.JNS18124

Li, Y., Huang, J., Huang, T., Tang, J., Zhang, W., Xu, W., et al. (2021). Wearable mixed-reality holographic navigation guiding the management of penetrating intracranial injury caused by a nail. *J. Digit. Imaging* 34, 362–366. doi: 10.1007/s10278-021-00436-3

Liu, Y., Wang, L., Cheng, J., Li, C., and Chen, X. (2020). Multi-focus image fusion: a survey of the state of the art. *Inform. Fus.* 64, 71–91. doi: 10.1016/j.inffus.2020.06.013

McJunkin, J., Jiramongkolchai, P., Chung, W., Southworth, M., Durakovic, N., Buchman, C., et al. (2018). Development of a mixed reality platform for lateral skull base anatomy. *Otol. Neurotol.* 39, e1137–e1142. doi: 10.1097/MAO.0000000000001995

Qi, Z., Li, Y., Xu, X., Zhang, J., Li, F., Gan, Z.-C., et al. (2021). Holographic mixed-reality neuronavigation with a head-mounted device: technical feasibility and clinical application. *Neurosurg. Focus* 51:E22. doi: 10.3171/2021.5.FOCUS21175

Sun, G.-C., Chen, X.-l., Hou, Y.-Z., Yu, X.-G., Ma, X.-D., Liu, G., et al. (2016). Image-guided endoscopic surgery for spontaneous supratentorial intracerebral hematoma. *J. Neurosurg.* 127, 537–542. doi: 10.3171/2016.7.JNS16932

van Doormaal, T. P. C., van Doormaal, J. A. M., and Mensink, T. (2019). Clinical accuracy of holographic navigation using point-based registration on augmented-reality glasses. *Oper. Neurosurg.* 17, 588–593. doi: 10.1093/ons/opz094

Wang, W.-Z., Jiang, B., Liu, G.-M., Li, D., Lu, C.-Z., Zhao, Y.-D., et al. (2009). Minimally invasive craniopuncture therapy vs. conservative treatment for spontaneous intracerebral hemorrhage: results from a randomized clinical trial in china. *Int. J. Stroke* 4, 11–16. doi: 10.1111/j.1747-4949.2009.00239.x

Yang, Z., Hong, B., Jia, Z., Chen, J., Ge, J., Han, J., et al. (2014). Treatment of supratentorial spontaneous intracerebral hemorrhage using image-guided minimally invasive surgery: initial experiences of a flat detector ct-based puncture planning and navigation system in the angiographic suite. *Am. J. Neuroradiol.* 35, 2170–2175. doi: 10.3174/ajnr.A4009

Zhang, C., Gao, H., Liu, Z., and Huang, H. (2021). The potential value of mixed reality in neurosurgery. *J. Craniof. Surg.* 32, 940–943. doi: 10.1097/SCS.0000000000007317

Zheng, M., Qi, G., Zhu, Z., Li, Y., Wei, H., and Liu, Y. (2020). Image dehazing by an artificial image fusion method based on adaptive structure decomposition. *IEEE Sensors J.* 20, 8062–8072. doi: 10.1109/JSEN.2020.2981719

Zhou, M., Wang, H., Zeng, X., Yin, P., Zhu, J., Chen, W., et al. (2019). Mortality, morbidity, and risk factors in China and its provinces, 1990-2017: a systematic analysis for the global burden of disease study 2017. *Lancet* 394, 1145–1158. doi: 10.1016/S0140-6736(19)30427-1

Zhu, Z., Wei, H., Hu, G., Li, Y., Qi, G., and Mazur, N. (2021). A novel fast single image dehazing algorithm based on artificial multiexposure image fusion. *IEEE Trans. Instrum. Measure*. 70, 1–23. doi: 10.1109/TIM.2020.3024335

Conflict of Interest: WYi was employed by QINYING 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.

METHODS
published: 02 June 2022
doi: 10.3389/fnins.2022.911957

Medical Image Registration Algorithm Based on Bounded Generalized Gaussian Mixture Model

Jingkun Wang[1†], *Kun Xiang*[2†], *Kuo Chen*[3], *Rui Liu*[2], *Ruifeng Ni*[2], *Hao Zhu*[2*] and *Yan Xiong*[1*]

[1] *Department of Orthopaedics, Daping Hospital, Army Medical University, Chongqing, China,* [2] *College of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China,* [3] *School of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China*

OPEN ACCESS

Edited by:
Yu Liu,
Hefei University of Technology, China

Reviewed by:
Xin Jin,
Yunnan University, China
Yu Zhang,
Beihang University, China
Huafeng Li,
Kunming University of Science and Technology, China

Correspondence:
Hao Zhu
zhuhao@cqupt.edu.cn
Yan Xiong
xiongyan815@163.com

† *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *03 April 2022*
Accepted: *04 May 2022*
Published: *02 June 2022*

Citation:
Wang J, Xiang K, Chen K, Liu R, Ni R, Zhu H and Xiong Y (2022) Medical Image Registration Algorithm Based on Bounded Generalized Gaussian Mixture Model. Front. Neurosci. 16:911957. doi: 10.3389/fnins.2022.911957

In this paper, a method for medical image registration based on the bounded generalized Gaussian mixture model is proposed. The bounded generalized Gaussian mixture model is used to approach the joint intensity of source medical images. The mixture model is formulated based on a maximum likelihood framework, and is solved by an expectation-maximization algorithm. The registration performance of the proposed approach on different medical images is verified through extensive computer simulations. Empirical findings confirm that the proposed approach is significantly better than other conventional ones.

Keywords: medical image registration, gray-level-based registration, multimodal, Gaussian mixture model, bounded generalized Gaussian mixture model

INTRODUCTION

Image registration is an essential part of computer vision and image processing (Visser et al., 2020), which is widely used in medical image analysis and intelligent vehicles (Zhu et al., 2013, 2017, 2021a,b, 2022). Medical image analysis is the basis for judging the patient's condition in future intelligent diagnosis and treatment or auxiliary diagnosis and treatment (Weissler et al., 2015; Yang et al., 2018). More importantly, image registration sets the stage for subsequent image segmentation and fusion (Saygili et al., 2015; Zhu et al., 2019). Current clinical practice typically involves printing images onto radiographic film and viewing them on a lightbox. The computerized approach offers potential benefits, particularly by accurately aligning the information in different images and providing tools to visualize the composite image. A key stage in this process is the alignment or registration of the images (Hill et al., 2001).

The premise of image registration is that there is a same logical part between the reference image and the floating image (Gholipour et al., 2007; Reaungamornrat et al., 2016). Image registration realizes transformation by determining the space coordinate transformation between two image pixels, which enables the corresponding region on the reference image to coincide with the floating image in space (Zhang et al., 2019). This means that the same anatomical point on the human body has the same spatial position (the same position, angle and size) on two matched images (Gefen et al., 2007).

There are two medical image registration methods: feature-based registration and gray-level-based registration (Sengupta et al., 2021). The feature-based registration method does not directly utilize the gray-level information of the image. It is based on abstracting the geometric features (such as corners, the center of the closed region, edges, contours, etc.) that remain unchanged in the image to be registered. The parameter values of the transformation model between the images to be registered are obtained by describing the features of the two images, respectively, and establishing the matching relationship (Huang, 2015). The image registration based on this feature has advantages of less computation and faster registration speed, and it is robust to changes of gray image scale. However, its registration accuracy is usually not as high as that of gray-level-based image registration (Li et al., 2020; Ran and Xu, 2020).

In the gray-level-based medical image registration method, a similarity measure function between images is established through the gray information of the entire image (Yan et al., 2020). The transformation model parameters between images are obtained by maximizing and minimizing the value of the similarity measure function (Zhang et al., 2019). The gray-level-based image registration algorithm uses all the gray information of the image in the registration process. Therefore, the precision and robustness of the obtained transformation model are higher than the feature-based image registration (Frakes et al., 2008). The commonly used gray-level-based image registration methods are sequential similarity detection algorithm (SSDA), cross-correlation, mutual information, and phase correlation (Gupta et al., 2021). Based on the traditional algorithms, Yan et al. (2010) extracted a fast and effective algorithm, SSDA. Anuta (1970) proposed an image registration technique using Fourier transform for cross-correlation image detection and calculation to improve speed performance of registration. Evangelidis and Psarakis (2008) offered a modified version of the correlation coefficient as a performance criterion for image approval. Zheng et al. (2011) proposed a cross-correlation registration algorithm based on image rotation projection to avoid rotation and interpolation steps in image registration, reducing data dimension and computational complexity. For image registration using mutual information as a similarity measure, Pluim et al. (2000) combined image gray level with spatial image information and added image gradient into the algorithm, which successfully solved the problem of finding the global optimal solution in the registration process. A direct image registration method using mutual information (MI) as an alignment metric was proposed by Dame and Marchand (2012). A set of two-dimensional motion parameters can be estimated accurately in real time by optimizing the maximum mutual information. Lu et al. (2008) proposed a new joint histogram estimation method, which utilizes Hanning's windowed since approximation function as a kernel function of partial volume interpolation. Orchard and Mann (2009) utilized the maximum likelihood clustering method of the joint strength scatter chart. The expected probability of the cluster is modeled as a Gaussian mixture model (GMM), and the expectation-maximization (EM) method is utilized for achieving solution in iterative algorithm. Sotiras et al. (2013) emphasized the technology applied to medical images and systematically presented the latest technology. The paper provided an extensive account of registration techniques in a systematic manner. Pluim et al. (2004) compared the performance of mutual information as a registration measure with that of other f-information measures. An important finding is that several measures can potentially yield significantly more accurate results than mutual information. Klein et al. (2007) compared the performance of eight non-rigid registration optimization methods of medical images. The results show that the Robbins–Monro method is the best choice in most applications. With this approach, the computation time per iteration can be lowered approximately 500 times without affecting the rate of convergence. However, the distribution range of GMM is $(-\infty, +\infty)$, and so the method could not process the target information in a fixed area.

In the field of computer vision, image pixel values are distributed over a limited area of $[0, 255]$. Therefore, the bounded generalized Gaussian mixture model (BGGMM) is used to model the image (Nguyen et al., 2014), which can more thoroughly describe the joint intensity vector distribution of the image pixels and highlight the details of the image. The BGGMM has good robustness at the same time. Therefore, based on the BGGMM, this paper models both single-modality and multimodal image registration and then solves the model under the framework of maximum likelihood estimation (Zhu and Cochoff, 2002). Experimental verification results on a large number of image data sets show that compared with the existing gray-level-based medical image registration algorithm based, the image registration accuracy of the proposed method is improved.

PROBLEM FORMULATION

Suppose that two different medical images are registered, one medical image represents the reference image, denoted by A, and the other represents the floating image, denoted by B. These two different medical images come from different sensors. Therefore, each pixel position x in the space of two medical images corresponds to a pixel value, and we use the joint intensity vector to represent the intensity value of the two images at the position. Here, I_x can be expressed as:

$$I_x = [A_x;\ B_x] \tag{1}$$

Among them, A_x and B_x, respectively, represent the pixel value of the reference image and the floating image at the pixel position x. In order to realize the registration of two images, it is necessary to assign N registration parameters to each image to describe the spatial transformation of the image. θ can represent the set of all registration parameters. Then, the joint intensity vector of the registration image after employing registration parameters can be re-expressed as I_x^θ.

The bounded generalized Gaussian mixture model (BGGMM) is used to describe the distribution of the joint intensity. The probability distribution of the joint strength vector is:

$$p(I_x^\theta|\rho) = \sum_{m=1}^{M} \tau_m BG(I_x^\theta|u_m, \sigma_m, \Lambda_m) \tag{2}$$

Where $\rho = \{u_m, \sigma_m, \Lambda_m, \tau_m\}$is the model parameters, M represents the number of bounded generalized Gaussian

(BGG) distribution components in the mixture model, u_m, σ_m and Λ_m, respectively, represent the mean, covariance, and shape parameters of the m-th BGG distribution component. $\tau_m$represents the weight of the distribution component in the mixture model and satisfies the condition $\tau_m \geq 0$ and $\sum_{m=1}^{M} \tau_m = 1$. $BG(.)$ represents a BGG distribution, i.e.,

$$BG(I_x^\theta|u_m, \sigma_m, \Lambda_m) = \frac{T(I_x^\theta|u_m, \sigma_m, \Lambda_m)H(I_x^\theta|\Omega_m)}{\int_{\partial_m} \mathrm{T}(I_x^\theta|u_m, \sigma_m, \Lambda_m)dx} \tag{3}$$

Which ∂_m represents a bounded support area, and the distribution $\mathrm{T}(I_x^\theta|u_m, \sigma_m, \Lambda_m)$ is written as

$$T(I_x^\theta|u_m, \sigma_m, \Lambda_m) = \alpha(\Lambda_m) \exp\left(-\beta(\Lambda_m)\left|\frac{I_x^\theta - u_m}{\sigma_m}\right|^{\Lambda_m}\right) \tag{4}$$

and

$$H(I_x^\theta|\Omega_m) = \begin{cases} 1, & \textit{if } I_x^\theta \textit{ belongs to } \partial_m \\ 0, & \textit{otherwise} \end{cases} \tag{5}$$

$$\alpha(\Lambda_m) = \frac{\Lambda_m\sqrt{\Gamma(3/\Lambda_m)}}{2\sigma_m\Gamma(1/\Lambda_m)\sqrt{\Gamma(1/\Lambda_m)}}, \beta(\Lambda_m) = \left[\frac{\Gamma(3/\Lambda_m)}{\Gamma(1/\Lambda_m)}\right]^{\Lambda_m/2} \tag{6}$$

Where $\Gamma(\cdot)$ is the gamma function.

Therefore, X represents the number of pixels, and the log-likelihood function of image registration is:

$$\mathcal{L}(\rho) = \sum_{x}^{X} log\ p\left(I_x^\theta\, |\rho\right) \tag{7}$$

In the framework of maximum likelihood, the hidden variable z_{xm} that is introduced to the model indicates the category of the cluster that I_x^θ belongs to, that is, it belongs to the m-th (BGG) distribution component. Therefore, the log-likelihood function of the model can be written as:

$$\mathcal{L}(\rho) = \sum_{x}^{X} log\ p\left(I_x^\theta, z_{xm}\, |\rho\right) \tag{8}$$

PARAMETERS ESTIMATION

Density Estimation

According to the above model, the EM algorithm is used to estimate various parameters involved in the model. The EM algorithm is mainly divided into two steps, step E and step M.

Step E: $Q\left(\rho, \rho^t\right) = E\left[\mathcal{L}(\rho)\left|I_x^\theta, \rho^t\right.\right]$
Step M: $\rho^{t+1} = \max\limits_{\rho} Q\left(\rho \mid \rho^t\right)$

Here t represents the t-th iteration. The final model parameters can be determined by iterating these two steps.

In step E, the probability that I_x^θ belonging to the m-th cluster is given:

$$\eta\left(z_{xm}\right) = p\left(z_{xm}|I_x^\theta, \rho\right) = \frac{\tau_m BG\left(I_x^\theta|u_m, \sigma_m, \Lambda_m\right)}{\sum_{m=1}^{M} \tau_m BG\left(I_x^\theta|u_m, \sigma_m, \Lambda_m\right)} \tag{9}$$

Where $\sum_{m=1}^{M} \eta\left(z_{xm}\right) = 1$. Using the posterior distribution $\eta\left(z_{xm}\right)$ and the current parameters $\rho^{(t)}$

$$Q\left(\rho, \rho^t\right) = E\left[\mathcal{L}(\rho)\left|I_x^\theta, \rho^t\right.\right]$$

$$= \sum_{x=1}^{X}\sum_{m=1}^{M} \eta\left(z_{xm}\right)$$

$$\left[\begin{array}{c} \log\tau_m + \log T\left(I_x^\theta|u_m, \sigma_m, \Lambda_m\right) + \\ \log H\left(I_x^\theta|\Omega_m\right) - \log \int\limits_{\partial} T\left(I_x^\theta \mid u_m, \sigma_m, \Lambda_m\right)dx \end{array}\right] \tag{10}$$

At step M, the parameters $u_m^{t+1}, \sigma_m^{t+1}, \Lambda_m^{t+1}, \tau_m^{t+1}$ at the time (t+1) are updated by the maximizing equation (10). The results are as follows:

$$u_m^{t+1} = \frac{\sum_{x=1}^{X} \eta\left(z_{xm}\right)\left(\left|I_x^\theta - u_m^t\right|^{\Lambda_m^t - 2} I_x^\theta + R_m\right)}{\sum_{x=1}^{X} \eta\left(z_{xm}\right)\left|I_x^\theta - u_m^t\right|^{\Lambda_m^t - 2}} \tag{11}$$

Where R_m represents:

$$R_m = \frac{\sum_{o=1}^{O} sign\left(u_m^t - S_{om}\right)\left|S_{om} - u_m^t\right|^{\Lambda_m - 1} H\left(S_{om}|\Omega_m\right)}{\sum_{o=1}^{O} H\left(S_{om}|\Omega_m\right)} \tag{12}$$

In formula (12), when $x \geq 0$, $sign(x)$ is equal to 1, otherwise it is equal to 0. $Som \sim T\left(I_x^\theta \mid u_m^t, \sigma_m^t, \Lambda_m^t\right)$ represents the random variable in the probability distribution $T\left(I_x^\theta \mid u_m^t, \sigma_m^t, \Lambda_m^t\right)$, o is the number of random variables S_{om}. Note that O is a large integer, and $O = 10^6$ is taken in this paper.

$$\sigma_m^{t+1} = \left[\frac{\Lambda_m^t\beta\left(\Lambda_m^t\right)\sum_{x=1}^{X} \eta\left(z_{xm}\right)\left|I_x^\theta - u_m^t\right|^{\Lambda_m^t}}{\sum_{x=1}^{X} \eta\left(z_{xm}\right)\left(1 + Gm\right)}\right]^{\frac{1}{\Lambda_m^t}} \tag{13}$$

Where Gm represents:

$$G_m = \frac{\sum_{o=1}^{O}\left[-1 + \Lambda_m^t\beta\left(\Lambda_m^t\right)\left|S_{om} - u_m^t\right|^{\Lambda_m^t}\left(\sigma_m^t\right)^{-\Lambda_m^t}\right] H\left(S_{om}|\Omega_m\right)}{\sum_{o=1}^{O} H\left(S_{om}|\Omega_m\right)} \tag{14}$$

Under the condition that other parameters are fixed, use the Newton-Raphson method to estimate Λ_m. Each iteration needs to solve the first and second derivatives of $Q\left(\rho, \rho^t\right)$ with respect to parameter Λ_m. The next iteration value of Λ_m can be expressed as:

$$\Lambda_m^{t+1} = \Lambda_m^t - \frac{\partial Q\left(\rho, \rho^t\right)}{\partial\Lambda_m}\left[\frac{\partial Q^2\left(\rho, \rho^t\right)}{\partial\Lambda_m^2} + \vartheta\right]^{-1}|_{\Lambda_m = \Lambda_m^t} \tag{15}$$

Where ϑ is the scale factor, and the derivative of $Q\left(\rho, \rho^{t}\right)$ with respect to Λ_m is given by:

$$\frac{\partial Q\left(\rho, \rho^{t}\right)}{\partial \Lambda m} = -\sum_{x=1}^{X} \eta\left(z_{xm}\right)$$
$$\left[f\left(I_x^{\theta} \mid u_m, \sigma_m, \Lambda_m\right) - \frac{\int_{\partial} T\left(I_x^{\theta} \mid u_m, \sigma_m, \Lambda_m\right) f\left(I_x^{\theta} \mid u_m, \sigma_m, \Lambda_m\right) dx}{\int_{\partial} T\left(I_x^{\theta} \mid u_m, \sigma_m, \Lambda_m\right) dx}\right] \tag{16}$$

Where:

$$f\left(I_x^{\theta} \mid u_m, \sigma_m, \Lambda_m\right) = \left[\frac{1}{\Lambda_m} + \frac{3BG\left(\frac{1}{\Lambda_m}\right) - 3BG\left(\frac{3}{\Lambda_m}\right)}{2\Lambda^{2m}}\right]$$
$$-\ BG\left(\Lambda_m\right)\left|\frac{I_x^{\theta} - u_m}{\sigma_m}\right|^{\Lambda_m} \log\left|\frac{I^{\theta x - u_m}}{\sigma_m}\right| - BG\left(\Lambda_m\right)\ \times$$
$$\left[\frac{1}{2}\log\frac{\Gamma\left(\frac{3}{\Lambda_m}\right)}{\Gamma\left(\frac{1}{\Lambda_m}\right)} + \frac{BG\left(\frac{1}{\Lambda_m}\right) - 3BG\left(\frac{3}{\Lambda_m}\right)}{2\Lambda_m}\right]\left|\frac{I_x^{\theta} - u_m}{\sigma_m}\right|^{\Lambda_m} \tag{17}$$

$$\int_{\partial} T\left(I_x^{\theta} \mid u_m, \sigma_m, \Lambda_m\right) f\left(I_x^{\theta} \mid u_m, \sigma_m, \Lambda_m\right) dx$$
$$\approx \frac{1}{O}\sum_{o=1}^{O} f\left(S_{om} \mid u_m^{t}, \sigma_m^{t}, \Lambda_m^{t}\right) H\left(S_{om} \mid \boldsymbol{\Omega}_m\right) \tag{18}$$

The second derivative of $Q\left(\rho, \rho^{t}\right)$ with respect to Λ_m is:

$$\frac{\partial Q^2\left(\rho, \rho^{t}\right)}{\partial \Lambda_m^2} = -\sum_{x=1}^{X} \eta\left(\phi_{xm}\right)$$
$$\left[\begin{array}{c} g\left(I_x^{\theta} \mid u_m, \sigma_m, \Lambda_m\right) + \frac{\left(\int_{\partial} T\left(I_x^{\theta} \mid u_m,\sigma_m,\Lambda_m\right) f\left(I_x^{\theta} \mid u_m,\sigma_m,\Lambda_m\right) dx\right)^2}{\left(\int_{\partial} T\left(I_x^{\theta} \mid u_m,\sigma_m,\Lambda_m\right) dx\right)^2} - \\ \frac{\int_{\partial} T\left(I_x^{\theta} \mid u_m,\sigma_m,\Lambda_m\right)\left[f^2\left(I_x^{\theta} \mid u_m,\sigma_m,\Lambda_m\right) + g\left(I_x^{\theta} \mid u_m,\sigma_m,\Lambda_m\right)\right] dx}{\int_{\partial} T\left(I_x^{\theta} \mid u_m,\sigma_m,\Lambda_m\right) dx} \end{array}\right] \tag{19}$$

Where,

$$g\left(I_x^{\theta} \mid u_m, \sigma_m, \Lambda_m\right) = \frac{\partial f\left(I_x^{\theta} \mid u_m,\sigma_m,\Lambda_m\right)}{\partial \Lambda_m}$$
$$= \left[\frac{-1}{\Lambda_m^2} - \frac{3BG\left(\frac{1}{\Lambda_m}\right)}{2\Lambda_m^4} - \frac{3BG\left(\frac{1}{\Lambda_m}\right)}{\Lambda_m^3} + \frac{9BG\left(\frac{3}{\Lambda_m}\right)}{2\Lambda_m^4} + \frac{3BG\left(\frac{3}{\Lambda_m}\right)}{\Lambda_m^3}\right]$$
$$-\beta\left(\Lambda_m\right)\left|\frac{I_x^{\theta} - u_m}{\sigma_m}\right|^{\Lambda_m}\left(\log\left|\frac{I_x^{\theta} - u_m}{\sigma_m}\right|\right)^2 - \tag{20}$$
$$\beta\left(\Lambda_m\right) \times \left[\begin{array}{c} \frac{1}{2} log\frac{\Gamma\left(\frac{3}{\Lambda_m}\right)}{\Gamma\left(\frac{1}{\Lambda_m}\right)} + \frac{BG\left(\frac{1}{\Lambda_m}\right) - 3BG\left(\frac{3}{\Lambda_m}\right)}{2\Lambda_m} + \\ \frac{-BG'\left(\frac{1}{\Lambda_m}\right) + 9BG'\left(\frac{3}{\Lambda_m}\right)}{2\Lambda_m^3} \end{array}\right]^2$$
$$\left|\frac{I_x^{\theta} - u_m}{\sigma_m}\right|^{\Lambda_m} - \beta\left(\Lambda_m\right) \times \left[\frac{1}{2}\log\frac{\Gamma\left(\frac{3}{\Lambda_m}\right)}{\Gamma\left(\frac{1}{\Lambda_m}\right)} + \frac{BG\left(\frac{1}{\Lambda_m}\right) - 3BG\left(\frac{3}{\Lambda_m}\right)}{2\Lambda_m}\right]$$
$$\left|\frac{I_x^{\theta} - u_m}{\sigma_m}\right|^{\Lambda_m} \log\left|\frac{I_x^{\theta} - u_m}{\sigma_m}\right|$$

$$\int_{\partial} T\left(I_x^{\theta} \mid u_m, \sigma_m, \Lambda_m\right)$$
$$\left[f^2\left(I_x^{\theta} \mid u_m, \sigma_m, \Lambda_m\right) + g\left(I_x^{\theta} \mid u_m, \sigma_m, \Lambda_m\right)\right] dx$$
$$\approx \frac{1}{O}\left[\sum_{o=1}^{O} f^2\left(S_{om} \mid u_m^{t}, \sigma_m^{t}, \Lambda_m^{t}\right) + f\left(s_{om} \mid u_m^{t}, \sigma_m^{t}, \Lambda_m^{t}\right)\right]$$
$$H\left(s_{om} \mid \boldsymbol{\Omega}_m\right) \tag{21}$$

Finally, update the estimate of the prior probability τ_m^{t+1} that can be expressed as:

$$\tau_m^{t+1} = \frac{1}{X}\sum_{x=1}^{X} \eta\left(z_{xm}\right) \tag{22}$$

Motion Parameters Estimation

Optimize the corresponding parameter θ by deriving the result of $Q\left(\rho, \rho^{t}\right)$ to θ as 0:

$$\frac{\partial Q\left(\rho, \rho^{t}\right)}{\partial \theta} = 0 \tag{23}$$

In order to find the appropriate model movement parameter θ to satisfy the equation (23), introduce a small movement increment $\tilde{\theta}$ and replace θ with as the estimated parameter. The following is obtained by using approximate linear space transformation:

$$I_x^{\theta+\tilde{\theta}} = I_x^{\theta} + \frac{\partial I_x^{\theta^{T}}}{\partial \theta}\tilde{\theta} \tag{24}$$

Incorporate formula (23) into formula (24) and the following can be obtained:

$$\left\{\sum_{x=1}^{X}\left[\sum_{m=1}^{M} \eta\left(z_{xm}\right) \Lambda_m \beta(\Lambda_m) \frac{\partial I_x^{\theta}}{\partial \theta}\left(\sigma_m^{t+1}\right)^{-1} \frac{\partial I_x^{\theta\,T}}{\partial \theta}\right]\right\}\tilde{\theta}$$
$$= -\sum_{x=1}^{X}\left[\sum_{m=1}^{M} \eta\left(z_{xm}\right) \Lambda_m \beta(\Lambda_m) \frac{\partial I_x^{\theta}}{\partial \theta}\left(\sigma_m^{t+1}\right)^{-1}\left(I_x^{\theta} - u_m^{t+1}\right)\right] \tag{25}$$

The optimization of the registration parameters can be achieved by solving the movement increment $\tilde{\theta}$ in equation (25).

Implementation

In summary, the proposed image registration algorithm based on the BGGMM is shown in **Algorithm 1** and **Figure 1**. This paper regards M BGG distribution components in the joint intensity scatter plot of the registered image as M clusters, uses the k-mean method to find the cluster centers and compares parameter initialization of the BGGMM model. This paper initializes $\Lambda_m = 2$. Secondly, this paper also utilizes multi-resolution image registration, and the resolutions are set [0.1 0.2 1], respectively. The image is first registered at low resolution and then high resolution, and the registration result at each resolution can be used as the next resolution registration. Therefore, the calculation time can be reduced, and the algorithm convergence can be accelerated in the iterative process of the proposed algorithm.

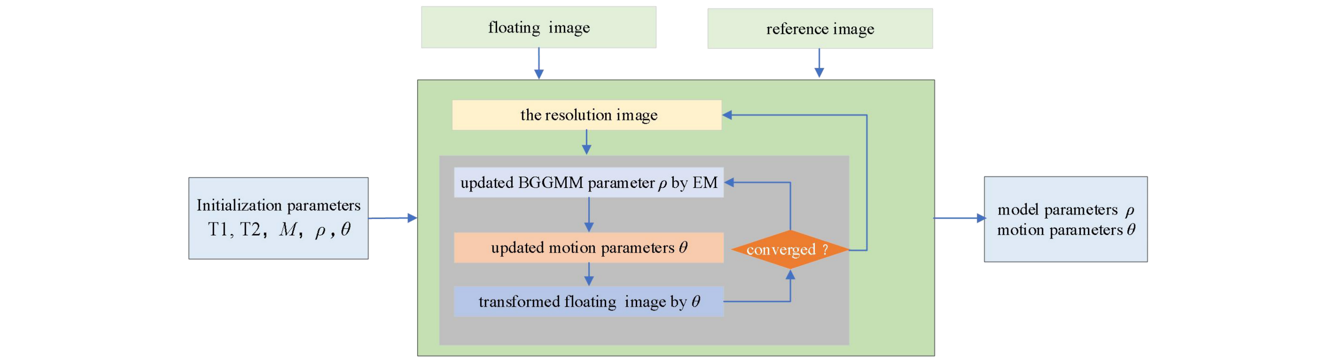

floating image
reference image
the resolution image
Initialization parameters
T1, T2, M, ρ, θ
updated BGGMM parameter ρ by EM
updated motion parameters θ
converged ?
transformed floating image by θ
model parameters ρ
motion parameters θ

FIGURE 1 | Flowchart of medical image registration.

The EM algorithm is first used to estimate the BGGMM model parameter ρ on the joint intensity scatter plot. After the optimal BGGMM model parameter ρ is estimated for T1 times, the motion adjustment is performed. This paper introduces a small movement increment and iterates T2 times to update the motion parameters, ensuring the optimal parameters are obtained. Finally, iterate repeatedly until convergence to achieve image registration.

Algorithm 1: Description of algorithm for medical image registration based on BGGMM.

Input: reference image A, floating image B, the number of clusters *M* of the BGGMM, the number of iterations T1, T2
Output: BGGMM model parametersρ, motion parameters θ

```
Initialization: k-mean initializes BGGMM model parameters $\rho$
for each scale do
    Get the $I'_x$ under the resolution image
    $I'_x$ applies motion parameters to get $I^{\theta'}_x$
    while not converged do
        for T1 iterations do
          Update BGGMM model parameter $\rho$ (step E and M)
        end for
        for T2 iterations do
          Move increment $\tilde{\theta}$
          Update exercise parameters
          Apply updated motion parameters to $I'_x$
        end
    end while
end for
```

EXPERIMENT

The computer environment of experiments in this paper is Intel(R) Core (TM) i5-7300HQ CPU @ 2.50 GHz with 8 GB RAM, while the operating system is 64-bit Windows 10.0. All simulations are implemented using MATLAB R2020b.

The mutual information method (MI) (Lu et al., 2008), the enhanced correlation coefficient (ECC) (Evangelidis and Psarakis, 2008) and the ensemble registration approach (ER) (Orchard and Mann, 2009) are compared to evaluate the performance of the proposed method. The average pixel displacement (PAD) (Li et al., 2016) is used as a registration error to objectively measure the performance of different approaches. In the successful registration case, the value of the PAD is zero. The larger the PAD, the more significant deviation and the lower registration accuracy. If PAD is greater than 3, the registration is considered to have failed.

MURA (Rajpurkar et al., 2017) and Altas (Yu and Zheng, 2016) public image data sets are used to verify the performance of these methods. Details about two image datasets and experiments are reported, as shown in **Table 1**, where the bold values indicate the best results. The *t*-test is used to test the significance of the difference between the PAD results of the BGGMM method and the other three

TABLE 1 | The pad results of image registration on public data sets.

Method/dataset	**Public dataset**	
	MURA images	**Atlas images**
MI	1.9162	0.7168
ECC	8.1494	10.7606
ER	6.5182	9.1342
Proposed method	**0.2271**	**0.6801**

The bold values indicated the best results.

TABLE 2 | The *t*-test results of the pad results of BGGMM versus other image registration methods on public data sets.

Database	**Method**		***p*-value**
MURA	BGGMM	MI	0.132
		ECC	0.000
		ER	0.000
Atlas	BGGMM	MI	0.034
		ECC	0.001
		ER	0.000

registration methods in image registration on public data sets. $P < 0.05$ means the difference is statistically significant, and the comparison results are summarized in **Table 2**. Both in the MURA and Atlas data sets, the PAD results of the BGGMM method were minor, and the differences were statistically significant compared to the PAD results of the ECC and ER methods ($P < 0.05$). In the MURA data set, the difference between the PAD results of the BGGMM

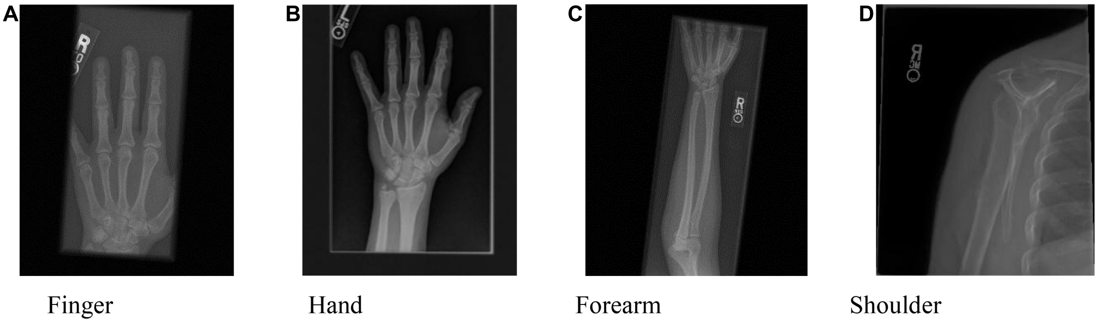


FIGURE 2 | One slice of the MURA dataset. **(A)** Finger, **(B)** Hand, **(C)** Forearm, and **(D)** Shoulder.

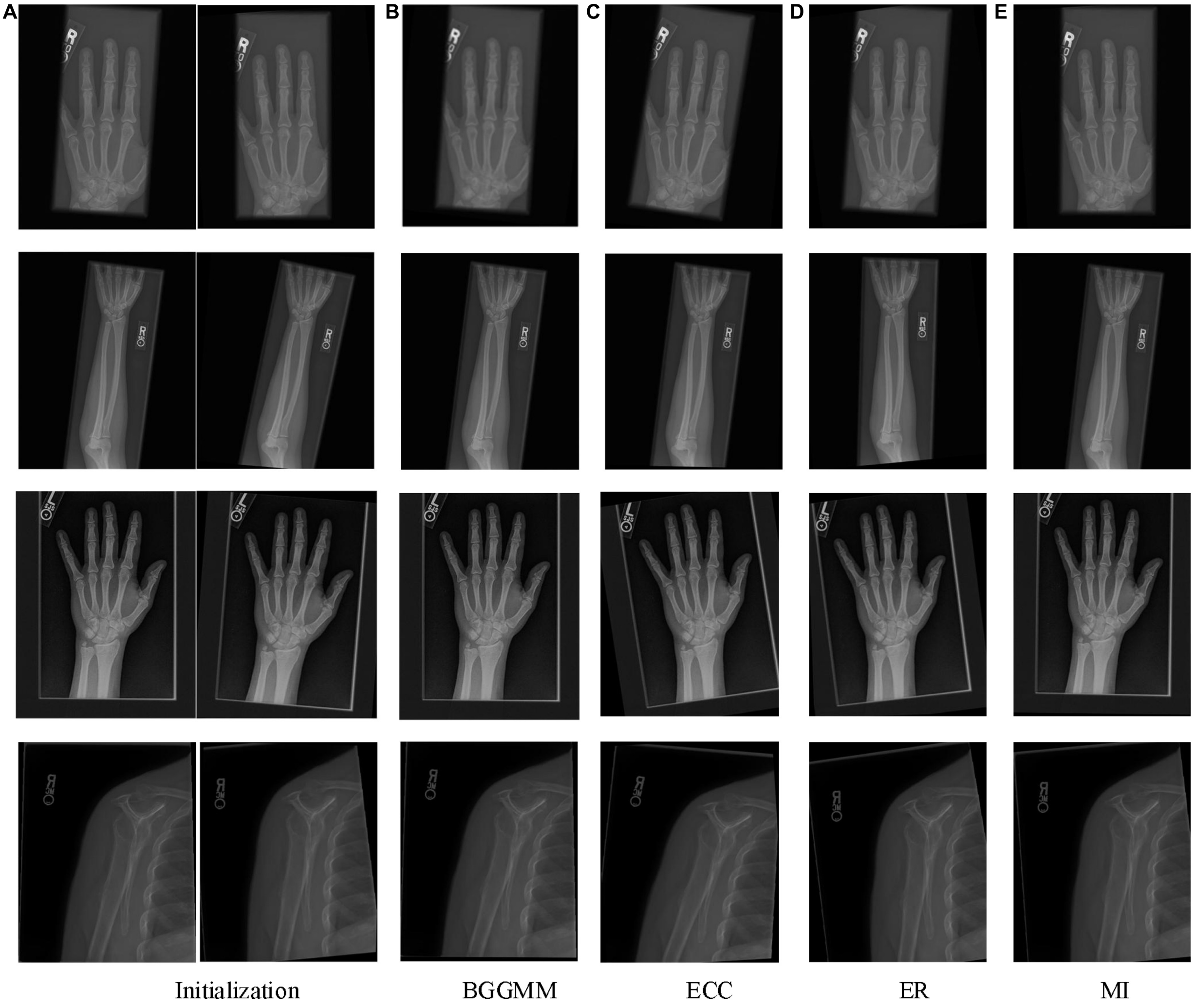
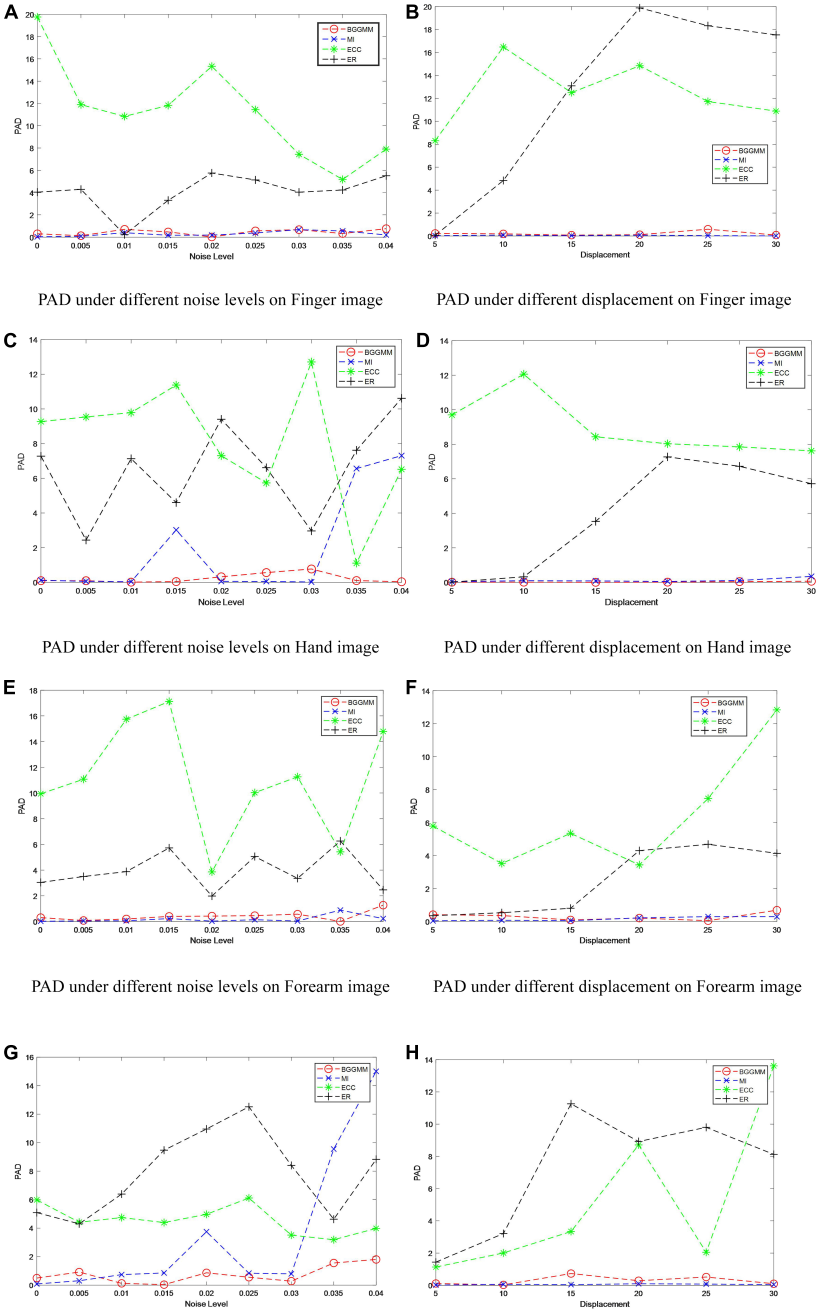


FIGURE 3 | The registration results of four methods in four skeleton images of MURA dataset. **(A)** Initialization. **(B)** BGGMM. **(C)** ECC. **(D)** (ER). **(E)** MI.

method and the MI method was not statistically different ($P > 0.05$). However, in the Atlas dataset, the PAD results of the BGGMM method were smaller than those of the MI method, and the difference was statistically significant ($P < 0.05$).

Musculoskeletal Radiographs Dataset

The proposed approach is tested on an ensemble of MURA images. The test set is from the Large Dataset for Abnormality Detection in Musculoskeletal Radiographs (MURA) project's training data set. One slice of this dataset is depicted in **Figure 2**.



PAD under different noise levels on Finger image

PAD under different displacement on Finger image

PAD under different noise levels on Hand image

PAD under different displacement on Hand image

PAD under different noise levels on Forearm image

PAD under different displacement on Forearm image

PAD under different noise levels on Shoulder image

PAD under different displacement on Shoulder image

FIGURE 4 | PAD of different methods under different noise levels and different displacements in MURA dataset. **(A)** PAD under different noise levels on Finger image. **(B)** PAD under different displacement on Finger image. **(C)** PAD under different noise levels on Hand image. **(D)** PAD under different displacement on Hand image. **(E)** PAD under different noise levels on Forearm image. **(F)** PAD under different displacement on Forearm image. **(G)** PAD under different noise levels on Shoulder image. **(H)** PAD under different displacement on Shoulder image.

The initial image to be registered is generated by random translation and rotation transformation, and the pixel and angle transformation parameters ranges are [−20, 20] and [−10, 10], respectively. This paper sets $M = 6$, that is, the number of BGG distribution components in the initial model is 6. The MURA dataset included 12,173 patients, 14,863 studies, and 40,561 multi-view radiographs. Each study belonged to one of the seven standard upper limb radiology study types: fingers, elbows, forearms, hands, humerus, shoulders, and wrists. Each study was manually marked as normal or abnormal by the radiologist.

The PAD values of the MURA dataset are summarized in the first column of **Table 1**. The average registration error of the proposed BGGMM method is significantly lower than other methods. The BGGMM method is more advantageous in edge retention and information content of source images. The registration results of the four methods are shown in **Figure 3**,

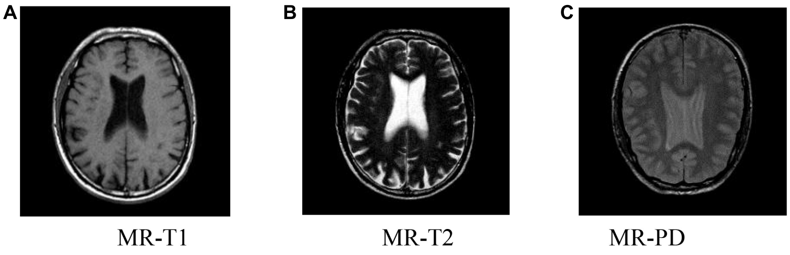


FIGURE 5 | Brain slice images from the Atlas dataset. **(A)** MR-T1, **(B)** MR-T2, **(C)** MR-PD.

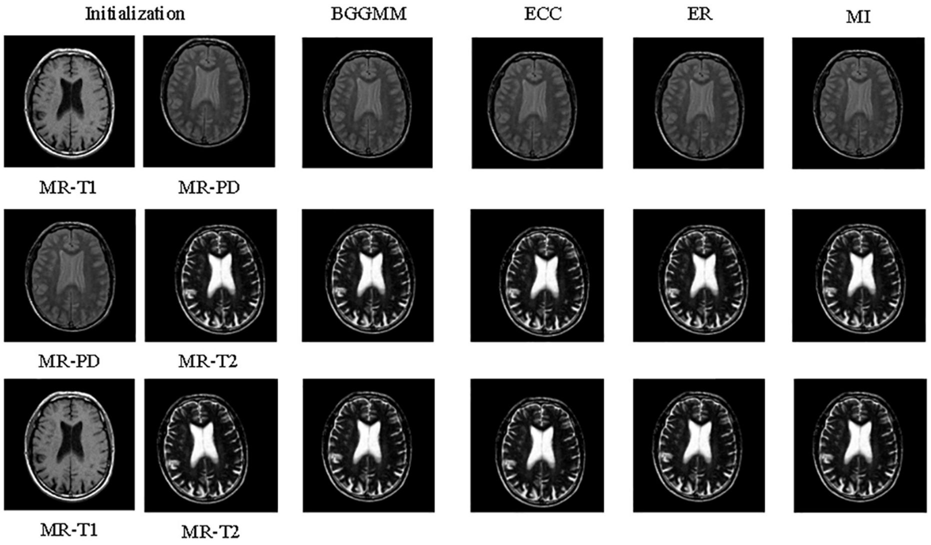


FIGURE 6 | The registration results of four methods in the brain images of Altas dataset.

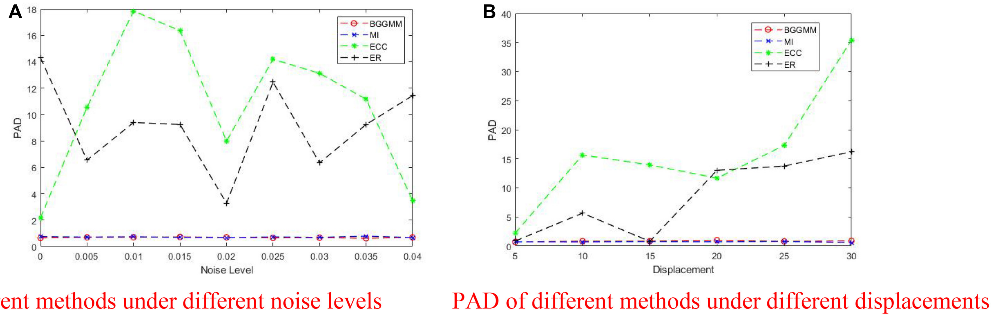


FIGURE 7 | PAD of BGGMM, ECC, ER, and MI methods under different noise levels and different displacements. **(A)** PAD of different methods under different noise levels. **(B)** PAD of different methods under different displacements.

which register the source image and transform the image with rotation and translation. In these four methods, registration is performed to the source image, and rotation, translation and transformation is performed to the image. **Figure 3A** shows the source image and the image to be registered.

With different noise levels, Gaussian noise is used as the independent variable in finger images of the experiment, and the noise level increases incrementally to test the performance of BGGMM. The mean value of Gaussian noise is 0, and the variance ranges from 0 to 0.04. As shown in **Figure 4A**, the excellent registration performance of several comparison algorithms can be observed. Among them, the registration error of the ER algorithm is the largest. The registration error of the BGGMM algorithm is lower than other methods under different noise levels.

The registration performance of the algorithm on Finger images is also tested under different displacement situations, as shown in **Figure 4B**. The displacement is added by moving the image *t* pixels horizontally and vertically, where the change range of *t* is 0–30, that is, the variation of the horizontal axis in **Figure 4B**. It is not difficult to see that the registration performance of this algorithm is better than other algorithms under different displacements. Among them, the ECC algorithm has poor anti-displacement interference, which is regarded as a registration failure. The ER algorithm has a good registration effect under the condition of small displacement. The BGGMM algorithm has the best performance when the change in displacement is large. Similarly, **Figures 4C–H** show the PAD value of different methods on Hand images, Forearm images, and Shoulder images under different noise levels and different displacements. The proposed method has the lowest registration error and the best registration performance.

Altas Dataset

Altas dataset is a multimodal dataset that includes more than 13,000 MRI and CT images of patients with brain diseases. Among them, MRI images have images with T1, T2, and PD weights. At the same time, it also includes the lesion images of patients with different lesion times. The image in which the MRI has T1, T2, and PD weights is selected, as shown in **Figure 5**. The initial image to be registered is generated by random translation and rotation transformation, and the pixel and angle transformation parameters ranges are [−20, 20] and [−10, 10], respectively. This paper sets $M = 6$, that is, the number of BGG distribution components in the initial model is 6.

The PAD values of Altas dataset are summarized in the second column of **Table 1**. The average registration error of the proposed BGGMM method is significantly lower than other methods. The BGGMM method has an advantage in preserving the edge information of the source image. The registration results of the four methods are shown in **Figure 6**. In these four methods, two different modality images are used to register separately.

The registration performance of BGGMM, ECC, and ER methods is tested under different Gaussian noises. According to the registration results in **Figure 7A**, the comparison of registration effects under different Gaussian noises can be obtained. The mean value of Gaussian noise is 0, and the variance ranges from 0 to 0.04. Among them, the registration error of the ECC algorithm is the largest. The PAD value of other algorithms mentioned above in this experiment is greater than 3, which is regarded as registration failures. The BGGMM algorithm has the lowest PAD value and has good registration performance.

As shown in **Figure 7B**, the displacement is added by moving the image *t* pixels horizontally and vertically, where the change range of *t* is 0–30. When the displacement changes considerably, the error generated by the ER algorithm becomes larger and exceeds the effective range. As the change in displacement increases, the PAD value of our BGGMM algorithm is still unaffected, always maintaining a low level and performing better among the four algorithms.

CONCLUSION

A medical registration method based on a BGGMM is proposed in this paper. Firstly, a BGGMM is applied to model the joint intensity vector distribution of the medical image. The proposed approach then formulates the model as an ML framework and estimates the parameters of models applying an EM algorithm. The experimental results indicate that the proposed BGGMM significantly improves registration performances on medical images compared with benchmark methods. The effect of this method is more pronounced when dealing with source images with more interference information and larger offsets. In the future, the research on medical image fusion will be carried out based on BGGMM image registration, which will provide convenience for medical image analysis.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

YX and HZ conceived and designed the study. JW and KX conducted most of the experiments and data analysis and wrote the manuscript. KC, RL, and RN participated in collecting materials and assisting in drafting manuscripts. All authors reviewed and approved the manuscript.

FUNDING

This study was supported by the Talents' Innovation Ability Training Program of Army Medical Center of PLA of China (Clinical Medicine Technological Innovation Ability Training Project, Grant No. 2019CXLCC017), the Key Talents Support Project of Army Medical University (Grant No. B-3261), the General Program of the National Natural Science Foundation of China (Grant Nos. 62073052 and 61905033), and the General Program of the Chongqing Natural Science Foundation of China (Grant No. cstc2021jcyj-msxmX0373).

REFERENCES

Anuta, P. E. (1970). Spatial registration of multispectral and multitemporal digital imagery using fast Fourier transform techniques. *IEEE Trans. Geosci. Electron.* 8, 353–368. doi: 10.1109/tge.1970.271435

Dame, A., and Marchand, E. (2012). Second-order optimization of mutual information for real-time image registration. *IEEE Trans. Image Process.* 21, 4190–4203. doi: 10.1109/TIP.2012.2199124

Evangelidis, G. D., and Psarakis, E. Z. (2008). Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Trans. Pattern Anal. Mach. Intell*. 30, 1858–1865. doi: 10.1109/TPAMI.2008.113

Frakes, D. H., Dasi, L. P., Pekkan, K., Kitajima, H. D., Sundareswaran, K., Yoganathan, A. P., et al. (2008). A new method for registration-based medical image interpolation. *IEEE Trans. Med. Imaging* 27, 370–377. doi: 10.1109/TMI.2007.907324

Gefen, S., Kiryati, N., and Nissanov, J. (2007). Atlas-based indexing of brain sections via 2-D to 3-D image registration. *IEEE Trans. Biomed. Eng.* 55, 147–156. doi: 10.1109/TBME.2007.899361

Gholipour, A., Kehtarnavaz, N., Briggs, R., Devous, M., and Gopinath, K. (2007). Brain functional localization: a survey of image registration techniques. *IEEE Trans. Med. Imaging* 26, 427–451. doi: 10.1109/TMI.2007.892508

Gupta, S., Gupta, P., and Verma, V. S. (2021). Study on anatomical and functional medical image registration methods. *Neurocomputing* 452, 534–548. doi: 10.1016/j.neucom.2020.08.085

Hill, D. L., Batchelor, P. G., Holden, M., and Hawkes, D. J. (2001). Medical image registration. *Phys. Med. Biol.* 46, R1–R45. doi: 10.1088/0031-9155/46/3/201

Huang, K. T. (2015). Feature Based Deformable Registration of Three-Dimensional Medical Images for Automated Quantitative Analysis and Adaptive Image Guidance. *Int. J. Radiat. Oncol. Biol. Phys.* 93, E558–E559.

Klein, S., Staring, M., and Pluim, J. P. W. (2007). Evaluation of Optimization Methods for Nonrigid Medical Image Registration Using Mutual Information and B-Splines. *IEEE Trans. Image Process*. 16, 2879–2890. doi: 10.1109/tip.2007.909412

Li, Q., Li, S., Wu, Y., Guo, W., Qi, S., Huang, G., et al. (2020). Orientation-independent Feature Matching (OIFM) for Multimodal Retinal Image Registration. *Biomed. Signal Process. Control* 60:101957. doi: 10.1016/j.bspc.2020.101957

Li, Y., He, Z., Zhu, H., and Wu, Y. (2016). Jointly registering and fusing images from multiple sensors. *Inform Fus.* 27, 85–94. doi: 10.1016/j.inffus.2015.05.007

Lu, X., Zhang, S., Su, H., and Chen, Y. (2008). Mutual information-based multimodal image registration using a novel joint histogram estimation. *Comput. Med. Imaging Graph.* 32, 202–209. doi: 10.1016/j.compmedimag.2007.12.001

Nguyen, T. M., Wu, Q. J. and Zhang, H. (2014). Bounded generalized Gaussian mixture model. *Pattern Recognit.* 47, 3132–3142.

Orchard, J., and Mann, R. (2009). Registering a multisensor ensemble of images. *IEEE Trans. Image Process*. 19, 1236–1247. doi: 10.1109/tip.2009.2039371

Pluim, J. P. W., Maintz, J. B. A., and Viergever, M. A. (2000). Image registration by maximization of combined mutual information and gradient information. *IEEE Trans. Med. Imaging* 19, 809–814. doi: 10.1109/42.876307

Pluim, J. P. W., Maintz, J. B. A., and Viergever, M. A. (2004). f-information measures in medical image registration. *IEEE Trans. Med. Imaging* 23, 1508–1516. doi: 10.1109/TMI.2004.836872

Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., and Mehta, H. (2017). A. MURA Dataset: Towards Radiologist-Level Abnormality Detection in Musculoskeletal Radiographs. 1, 2–215. doi: 10.48550/arXiv.1712.06957

Ran, Y., and Xu, X. (2020). Point cloud registration method based on SIFT and geometry feature. *Optik* 203:163902. doi: 10.1016/j.ijleo.2019.163902

Reaungamornrat, S., De Silva, T., Uneri, A., Vogt, S., Kleinszig, G., Khanna, A. J., et al. (2016). MIND demons: symmetric diffeomorphic deformable registration of MR and CT for image-guided spine surgery. *IEEE Trans. Med. Imaging* 35, 2413–2424. doi: 10.1109/TMI.2016.2576360

Saygili, G., Staring, M., and Hendriks, E. A. (2015). Confidence estimation for medical image registration based on stereo confidences. *IEEE Trans. Med. Imaging* 35, 539–549. doi: 10.1109/TMI.2015.2481609

Sengupta, D., Gupta, P., and Biswas, A. (2021). A survey on mutual information based medical image registration algorithms. *Neurocomputing* 486, 174–188. doi: 10.1109/TMI.2003.815867

Sotiras, A., Davatzikos, C., and Paragios, N. (2013). Deformable Medical Image Registration: a Survey. *IEEE Trans. Med. Imaging* 47, 3132–3142. doi: 10.1109/TMI.2013.2265603

Visser, M., Petr, J., Müller, D. M., Eijgelaar, R. S., Hendriks, E. J., Witte, M., et al. (2020). Accurate MR image registration to anatomical reference space for diffuse glioma. *Front. Neurosci.* 14:585. doi: 10.3389/fnins.2020.00585

Weissler, B., Gebhardt, P., Dueppenbecker, P. M., Wehner, J., Schug, D., Lerche, C. W., et al. (2015). A digital preclinical PET/MRI insert and initial results. *IEEE Trans. Med. Imaging* 34, 2258–2270. doi: 10.1109/TMI.2015.2427993

Yan, X. C., Wei, S. M., Wang, Y. E., and Xue, Y. (2010). AGV's image registration algorithm based on SSDA. *Sci. Technol. Eng.* 10, 696–699.

Yan, X., Zhang, Y., Zhang, D., and Hou, N. (2020). Multimodal image registration using histogram of oriented gradient distance and data-driven grey wolf optimizer. *Neurocomputing* 392, 108–120. doi: 10.1016/j.neucom.2020.01.107

Yang, W., Zhong, L., Chen, Y., Lin, L., Lu, Z., Liu, S., et al. (2018). Predicting CT image from MRI data through feature matching with learned nonlinear local descriptors. *IEEE Trans. Med. Imaging* 37, 977–987. doi: 10.1109/TMI.2018.2790962

Yu, W., and Zheng, G. (2016). "Atlas-Based Reconstruction of 3D Volumes of a Lower Extremity from 2D Calibrated X-ray Images," in *International Conference on Medical Imaging and Augmented Reality*, (Cham: Springer International Publishing), 366–374. doi: 10.1007/978-3-319-43775-0_33

Zhang, J., Ma, W., Wu, Y., and Jiao, L. (2019). Multimodal remote sensing image registration based on image transfer and local features. *IEEE Geosci. Remote Sens. Lett.* 16, 1210–1214. doi: 10.1109/lgrs.2019.2896341

Zheng, L., Wang, Y., and Hao, C. (2011). Cross-correlation registration algorithm based on the image rotation and projection," in *2011 4th International Congress on Image and Signal Processing*, (Shanghai, China: IEEE), 1095–1098.

Zhu, H., Leung, H., and He, Z. (2013). State estimation in unknown non-Gaussian measurement noise using variational Bayesian technique. *IEEE Trans. Aerosp. Electron. Syst.* 49, 2601–2614. doi: 10.1109/TAES.2013.6621839

Zhu, H., Mi, J., Li, Y., Yuen, K. V., and Leung, H. (2021a). VB-Kalman Based Localization for Connected Vehicles with Delayed and Lost Measurements: theory and Experiments. *IEEE ASME Trans. Mech*. 49, 2601–2614. doi: 10.1109/TMECH.2021.3095096

Zhu, H., Zhang, G., Li, Y., and Leung, H. (2021b). A novel robust Kalman filter with unknown non-stationary heavy-tailed noise. *Automatica* 127:109511. doi: 10.1016/j.automatica.2021.109511

Zhu, H., Yuen, K. V., Mihaylova, L., and Leung, H. (2017). Overview of environment perception for intelligent vehicles. *IEEE Trans. Intell. Transp. Syst.* 18, 2584–2601. doi: 10.1109/TITS.2017.2658662

Zhu, H., Zhang, G., Li, Y., and Leung, H. (2022). An adaptive kalman filter with inaccurate noise covariances in the presence of outliers. *IEEE Trans. Automat. Control* 67, 374–381. doi: 10.1109/TAC.2021.3056343

Zhu, H., Zou, K., Li, Y., Cen, M., and Mihaylova, L. (2019). Robust non-rigid feature matching for image registration using geometry preserving. *Sensors* 19:2729. doi: 10.3390/s19122729

Zhu, Y. M., and Cochoff, S. M. (2002). Likelihood maximization approach to image registration. *IEEE Trans. Image Process*. 11, 1417–1426. doi: 10.1109/TIP.2002.806240

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.

ORIGINAL RESEARCH
published: 29 June 2022
doi: 10.3389/fnins.2022.939472

A Multimodal Classification Architecture for the Severity Diagnosis of Glaucoma Based on Deep Learning

Sanli Yi[1], *Gang Zhang*[1], *Chaoxu Qian*[2], *YunQing Lu*[2], *Hua Zhong*[2]* *and Jianfeng He*[1]*

[1] *School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China,* [2] *First Affiliated Hospital of Kunming Medical University, Kunming, China*

Edited by:
Yu Liu,
Hefei University of Technology, China

Reviewed by:
Juan Cheng,
Hefei University of Technology, China
Zhiqin Zhu,
Chongqing University of Posts and Telecommunications, China

***Correspondence:**
Hua Zhong
zhoculist@163.com
Jianfeng He
jfenghe@kmust.edu.cn

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *09 May 2022*
Accepted: *01 June 2022*
Published: *29 June 2022*

Citation:
Yi S, Zhang G, Qian C, Lu Y, Zhong H and He J (2022) A Multimodal Classification Architecture for the Severity Diagnosis of Glaucoma Based on Deep Learning. Front. Neurosci. 16:939472. doi: 10.3389/fnins.2022.939472

Glaucoma is an optic neuropathy that leads to characteristic visual field defects. However, there is no cure for glaucoma, so the diagnosis of its severity is essential for its prevention. In this paper, we propose a multimodal classification architecture based on deep learning for the severity diagnosis of glaucoma. In this architecture, a gray scale image of the visual field is first reconstructed with a higher resolution in the preprocessing stage, and more subtle feature information is provided for glaucoma diagnosis. We then use multimodal fusion technology to integrate fundus images and gray scale images of the visual field as the input of this architecture. Finally, the inherent limitation of convolutional neural networks (CNNs) is addressed by replacing the original classifier with the proposed classifier. Our architecture is trained and tested on the datasets provided by the First Affiliated Hospital of Kunming Medical University, and the results show that the proposed architecture achieves superior performance for glaucoma diagnosis.

Keywords: glaucoma, computer-aided diagnosis, multimodal fusion, classification, multi-layer perceptron

INTRODUCTIONS

Glaucoma is a major eye health problem that leads to irreversible visual impairment (Mirzania et al., 2020). Because glaucoma initially tends to affect marginal vision and may still be asymptomatic until the middle stage, most patients are not treated in time, and further damage can occur (Yang et al., 2020). Thus, the detection and especially the severity classification of glaucoma is beneficial for ophthalmologists to analyze the condition of patients and develop follow-up treatment plans.

Fundus images, optical coherence tomography (OCT), and visual field are used as public data in the clinic. OCT can accurately evaluate the thickness of the retinal nerve fiber layer (RNFL) by tomography technology (Bowd et al., 2022). Fundus images reflect the vascular status of the eyes by contrast agent injection, and Chan et al. (2014) demonstrated that mono fundus images can provide an equal diagnostic accuracy for glaucomatous optic neuropathy evaluation when compared to stereoscopic images. The gray scale image of the visual field manifests the defect of the patient's visual field by brightness transformation (Wroblewski et al., 2009). Compared with OCT, fundus images and visual fields are easier to obtain and can be directly used to diagnose glaucoma

(Wroblewski et al., 2009; Chan et al., 2014). The diagnosis of pathological images is crucial but time-consuming and laborious; thus, reliable computer-assisted diagnosis (CAD) of glaucoma has continued to expand in the recent years (Zheng et al., 2019). The diagnostic approaches by the above technologies for glaucoma can be divided into two categories. One is the single-path method, of which the input is single type data. For example, Wroblewski et al. (2009) used support vector machines (SVMs) to provide a valid clinical diagnosis of glaucoma based solely on visual field data. Escamez et al. (2021) developed a classifier for predicting glaucoma eyes based on peripapillary retinal nerve fiber layer (RNFL) thicknesses measured with OCT. The other is a multimodal fusion image, which is a combination of two or more types of data. For instance, Bizios et al. (2011) and Chen et al. (2019) employed multimodal fusion approaches to diagnose glaucoma by integrating OCT and visual field data and OCT and fundus images.

Nevertheless, there are at least three problems to be resolved. First, the inferior resolution of the common gray scale of the visual field affects the feature extraction of convolutional neural networks (CNNs) in the task of glaucoma diagnosis. Second, the majority of studies focused on employing a single type of data to simply diagnose health and glaucoma, whereas the diagnosis of glaucomatous severity is more significant for ophthalmologists (Rajendrababu et al., 2021). Third, some studies using CNNs to capture features still had difficulty meeting the requirements of accuracy in practical diagnostic tasks. The main reason is that each convolution kernel of CNNs focuses only on the feature information of itself and its boundary while lacking the ability to model some long-range dependencies in glaucoma images (Yao et al., 2021).

To address these challenges, we propose a multimodal classification architecture based on deep learning for the severity classification of glaucoma. In this architecture, first, the gray scale image of the visual field is reconstructed with a higher resolution in the preprocessing stage, which is conducive to the feature extraction of the proposed architecture. Second, the fundus image and reconstructed visual field gray scale image are integrated to obtain multimodel information for the classification task and then transferred into CNN models for feature extraction. Third, we construct an efficient classifier to address the limitation of CNNs. This adopts the multilayer perceptron (MLP) of vision transformer (Dosovitskiy et al., 2020) (ViT) to further extract global sequence information and can be directly connected after CNNs to replace its original classifier. The main contributions of this paper are as follows:

- A multimodal classification architecture based on deep learning is constructed for the task of severity classification of glaucoma. The gray scale image of the visual field is reconstructed with a higher resolution in the preprocessing stage, in which a more subtle gray scale division unit is modeled to provide more detailed feature information in the glaucoma diagnosis task.
- The proposed architecture fuses the fundus image and visual field gray scale image as the input to provide more information for the feature extraction of the network. This architecture realizes a 4-classification of glaucoma to present its severity, which is more convenient for ophthalmologists.
- To offset the limitation of CNNs, we propose a plug-and-play classifier which adopts the multilayer perceptron (MLP) of ViT to extract the global dependencies of images. Meanwhile, the proposed classifier can easily replace the original classifier of CNNs and significantly improve the accuracy of the diagnostic task.

BACKGROUND AND RELATED WORKS

In this section, the latest progress of deep learning and its application in the field of glaucoma diagnosis are reviewed.

Development of Deep Learning

In the recent years, deep learning algorithms, especially CNNs, have made significant progress. The introduction of ImageNet (Krizhevsky et al., 2017) provided an initial explanation for the conception of deep learning. Subsequently, Simonyan and Zisserman (2014) and Iandola et al. (2017) proposed visual geometry group (VGG) and SqueezeNet, respectively; they increased the depth of the network while keeping the perception field unchanged and improving the performance of the networks. Meanwhile, He et al. (2016) and Huang et al. (2016) introduced functional modules such as residual and dense modules to enhance the performance of CNNs. Due to these improvements, CNNs are widely applied in the field of CAD. However, CNNs lack the ability to model the global dependencies of images because of their inherent limitations. Recently, transformer (Vaswani et al., 2017), which is capable of modeling long-range sequence features, attracted tremendous attention in the computer vision field. Dosovitskiy et al. (2020) introduced a transformer into the image task and successfully used embedded 2-dimensional (2D) image patches as an input sequence to achieve comparable representation with CNNs. Therefore, to obtain better performance in the task of glaucoma diagnosis, it will be of greater significance to combine transformer to offset the limitations of the CNN model.

Deep Learning for Glaucoma Diagnosis

Many deep learning algorithms have been employed in the fields of glaucomatous classification (Gour and Khanna, 2020; Wang et al., 2020; Singh et al., 2021). Raja et al. (2020) used a CNN to segment the retinal layer based on OCT data and calculate the cup-to-disk ratio (CDR). This achieved 94.6% accuracy in the glaucoma prediction task. Li et al. (2019) employed visual field data collected from hospitals to identify glaucoma, and the accuracy reached 87.6%. Kim et al. (2018) and Guo et al. (2020) diagnosed and localized fundus images by VGG16 and UNet++ networks to classify glaucoma and achieved an accuracy of 91.2% and an area under the curve (AUC) of 90.1%, respectively. Bajwa et al. (2020) and Ibrahim et al. (2022) both proposed a two-stage framework: the former detected and located optic disks on fundus images and then classified them as healthy or glaucoma; the latter preprocessed glaucoma disease data by normalization and the mean absolute deviation method in the

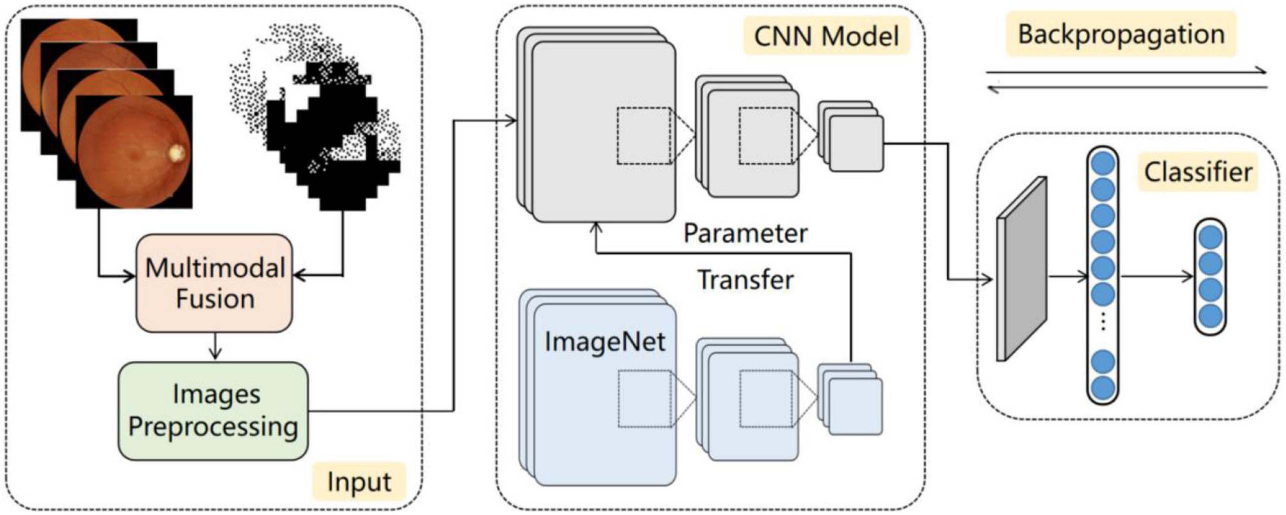


FIGURE 1 | Diagram of proposed architecture.

TABLE 1 | Distribution of dataset.

		Normal (class 0)	Early (class 1)	Intermediate (class 2)	Terminal (class 3)
Quantity	Original	87	171	79	165
	augmented	174	171	158	165

first stage and trained a deep learning model through the artificial algae optimization algorithm later. They achieved an AUC of 87.4% and an F1 score of 98.15%.

Different from the above works, Bizios et al. (2011) used a multimodal fusion approach to diagnose glaucoma by fusing OCT and standard automated visual field data and improved the AUC by 3.3% compared with single data. Chen et al. (2019) employed residual UNet to segment enhanced OCT and fundus images and then integrated the extracted features, achieving an accuracy rate of 96.88%. Kang et al. (2020) fused cup-to-disk and retinal nerve fiber layer features for the diagnosis of glaucoma. In the work of Liu et al. (2014), the limitation of the performance of a single modality was overcome by integrating patient personal data, major ocular image features, and important genome SNP features. This approach obtained the best AUC compared with a single modality.

MATERIALS AND METHODS

The workflow of the proposed multimodal classification architecture is shown in **Figure 1** and has three parts: input, CNN model, and classifier. First, the fundus image and reconstructed gray scale image of the visual field are fused into a multimodal fusion image, which are preprocessed and then sent into the CNN model. Second, as the feature extraction backbone of our architecture, the CNN model uses four ordinary CNNs to extract the feature information of the input image. These CNNs are pretrained by transfer learning technology to adapt to the task of small-scale datasets. Finally, the global dependencies of the feature maps are extracted by the proposed classifier to offset the limitations of the CNNs.

Input

Datasets

The dataset of this paper is provided by the First Affiliated Hospital of Kunming Medical University. It contains 502 fundus images and 502 visual field reports from 274 individuals, and both eyes of each individual were used in the study. Fundus images and visual field reports were acquired by a Topcon fundus camera TRC-50DX and Intelligent Video Surveillance (ISV) automatic computerized perimetry, and each image was labeled by two professional physicians. The datasets were rated from class 0 to 3 based on the severity of glaucoma, representing normal ($n = 87$), early ($n = 171$), intermediate ($n = 79$), and terminal glaucoma ($n = 165$), respectively. Related information of the dataset is listed in **Table 1**. Meanwhile, to overcome the challenges of training on imbalanced data by CNNs, we augmented normal eyes from 87 to 174 and intermediate glaucoma from 79 to 158 through data augmentation technology and balanced the ratio of all categories of data to ∼1:1:1:1. Finally, 1,336 images of the two types of data were applied to our deep learning architecture. The data sample is depicted in **Figure 2**.

Preprocessing

The preprocessing consists of two parts: data augmentation and normalization, and improving the resolution of the visual field gray scale image by reconstructing gray scale units.

Augmentation and Normalization

As shown in **Table 1**, the distribution of each category in the dataset is severely imbalanced, which may skew the diagnosis of CNNs toward more data-intensive types. To address this problem, we use data augmentation technology such as rotation, flipping, brightness, and contrast adjustment to form a dataset with the sample number of each category being almost equal. Meanwhile, to make the data more suitable for the pretraining of CNNs based on ImageNet, of which the default input resolution is 224 × 224, the images are resized to 224 × 224 pixels by bilinear interpolation.

Reconstruction of Visual Field Gray Scale Images

As depicted in **Figure 3A**, the gray scale image of the visual field is constructed based on the numerical value map, and each gray scale value in the image is represented by a gray scale unit. In the ordinary gray scale image, due to its low resolution (each gray scale unit represents a value with a span of 5 dB) (**Figure 3B**), much information is lost in the training process of CNNs, thus affecting the ability of CNNs to extract subtle features. In this paper, to solve this problem, a more subtle gray scale unit and corresponding gray scale image are established in which the gray scale unit is divided into 1 dB to retain the subtle features of the gray scale image (**Figure 3C**).

Multimodal Fusion

In this paper, the proposed multimodal classification architecture fuses fundus images and visual field gray scale images through an image concatenation approach and then transfers it into the CNN model to capture sufficient feature information. This is different from other studies. For instance, Chen et al. (2019) input images into CNNs for extracting features and then fused the extracted features to diagnose glaucoma. Such a fusion method changes the extracted features during the fusion, so the fused feature information is not reliable. Our proposed architecture fuses multimodal images before training, avoiding the mutual interference of features while improving the performance of glaucoma diagnosis.

CNN Model

Here, four CNNs (VGG 19, SqueezeNet, ResNet 50, and DenseNet 121) are adopted to extract the primary features of the fusion image in the proposed architecture. The details are shown in **Figure 4**.

VGG

Visual geometry group has a very systematic architecture. With the deepening of the network, the size of the input image is gradually compressed, but the number of convolution kernels is constantly increasing to explain the reduction in image size. Briefly, abundant 3 × 3 convolutional kernels are accumulated to replace the macrokernels to enhance the depth and width of the network. Thus, the higher the number of activation functions, the richer the extracted features and the stronger the recognition ability of the classification task.

SqueezeNet

SqueezeNet replaces the 3 × 3 convolutional kernel with abundant 1 × 1 kernels to reduce the computational cost and accelerate the training process of CNNs, with approximate results of AlexNet on the ImageNet dataset. The network is widely employed for large-scale datasets due to its light weight and high efficiency.

ResNet

Different from VGG, ResNet solves the degradation problem of deep networks by connecting the residuals of feature mapping from one layer to the subsequent through residual connections on its basis. Researchers can train deeper networks to improve task representation by solving ill-posed problems.

DenseNet

DenseNet, based on ResNet's theory, connects one layer to all subsequent layers by skipping connections, achieving dense skip connections. With further architectural transformations, the internal representation of DenseNet becomes significantly different from ResNets.

One key aspect is the use of network name suffixes in **Figure 4**. Roughly speaking, the number of layers in the network

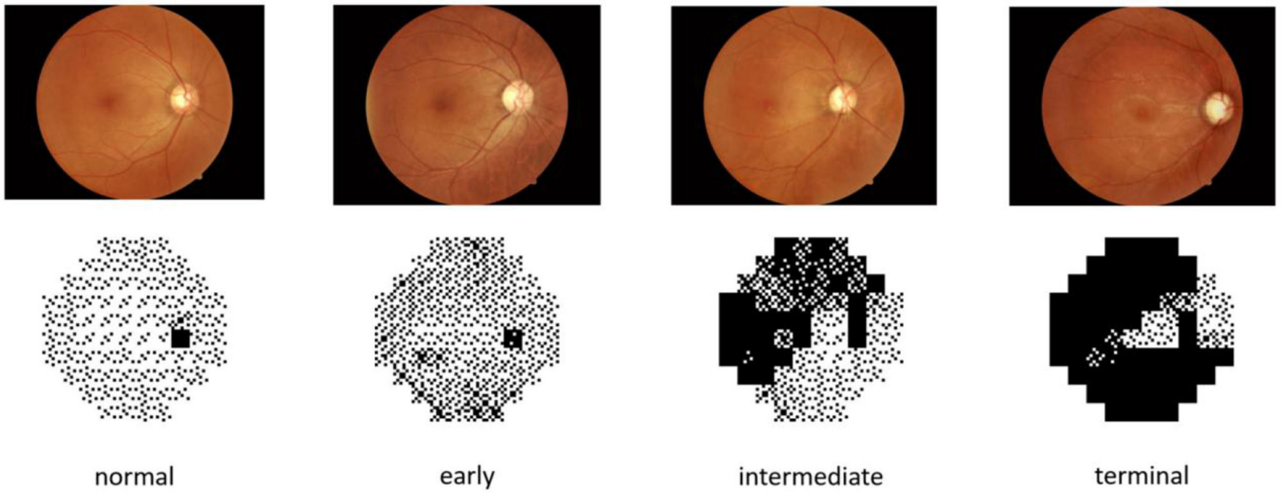


FIGURE 2 | Samples of different severities.

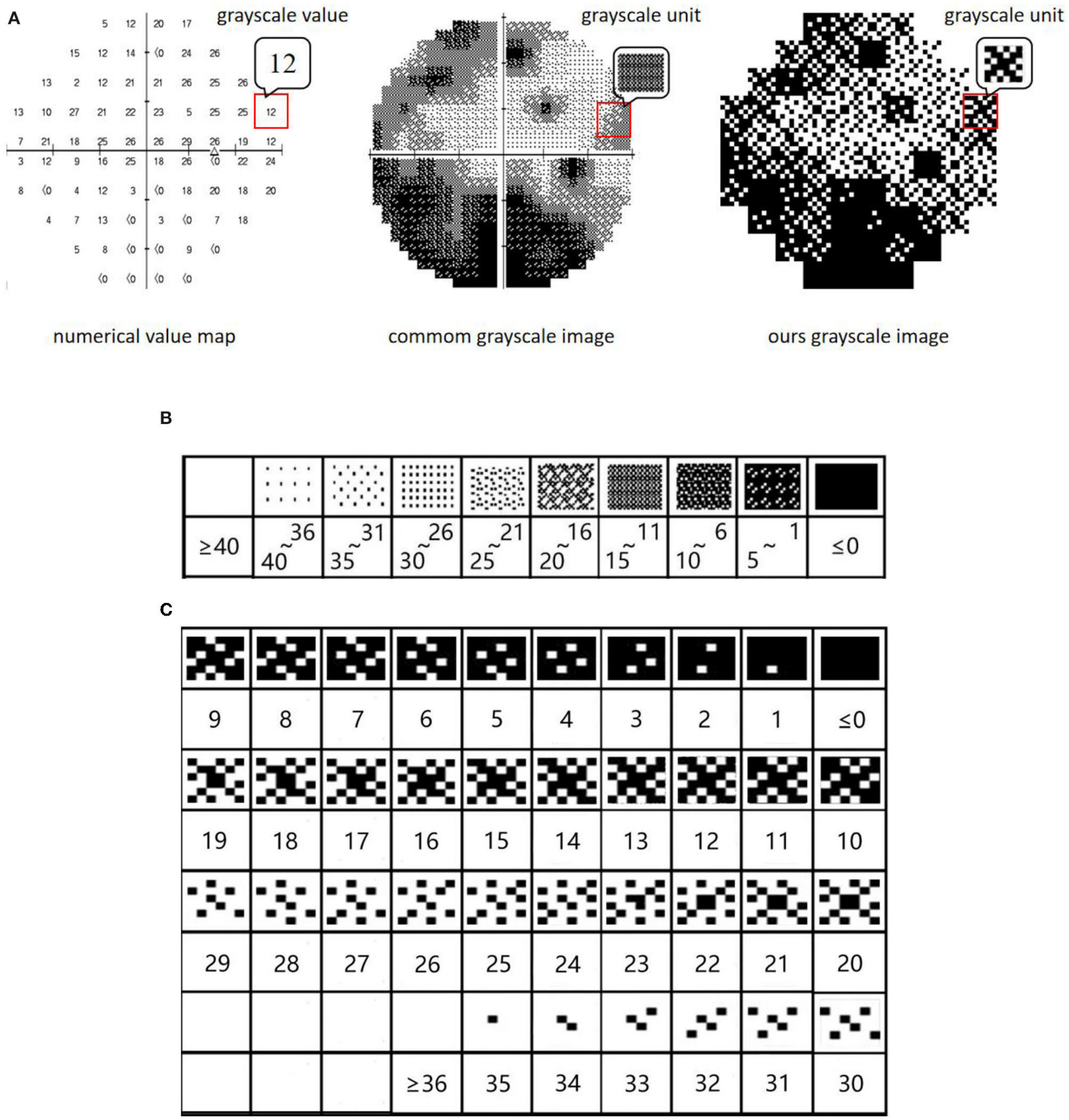


FIGURE 3 | (A) Gray scale images of visual field. **(B)** Ordinary gray scale units. **(C)** New gray scale units.

is represented as "19," "50," and "121." As you can see, the layers of the selected networks range from relatively shallow to extremely deep. This is intentional, as it leads to more architectural diversity.

Classifier

As the classifiers of CNNs are usually composed of a fully connected layer or maxpooling functions (**Figure 5A**), they lack the ability to model the long-range dependencies of glaucoma images. Therefore, we propose an effective classifier replacing the originals to offset their limitation in this paper, which is constructed by the MLP of ViT. As mentioned above, ViT can extract the global dependencies, and inspired by (Melas-Kyriazi, 2021), such an ability can be realized by its multilayer perceptron (MLP) alone, so it is employed in our classifier. **Figure 5B** shows an overview of this module.

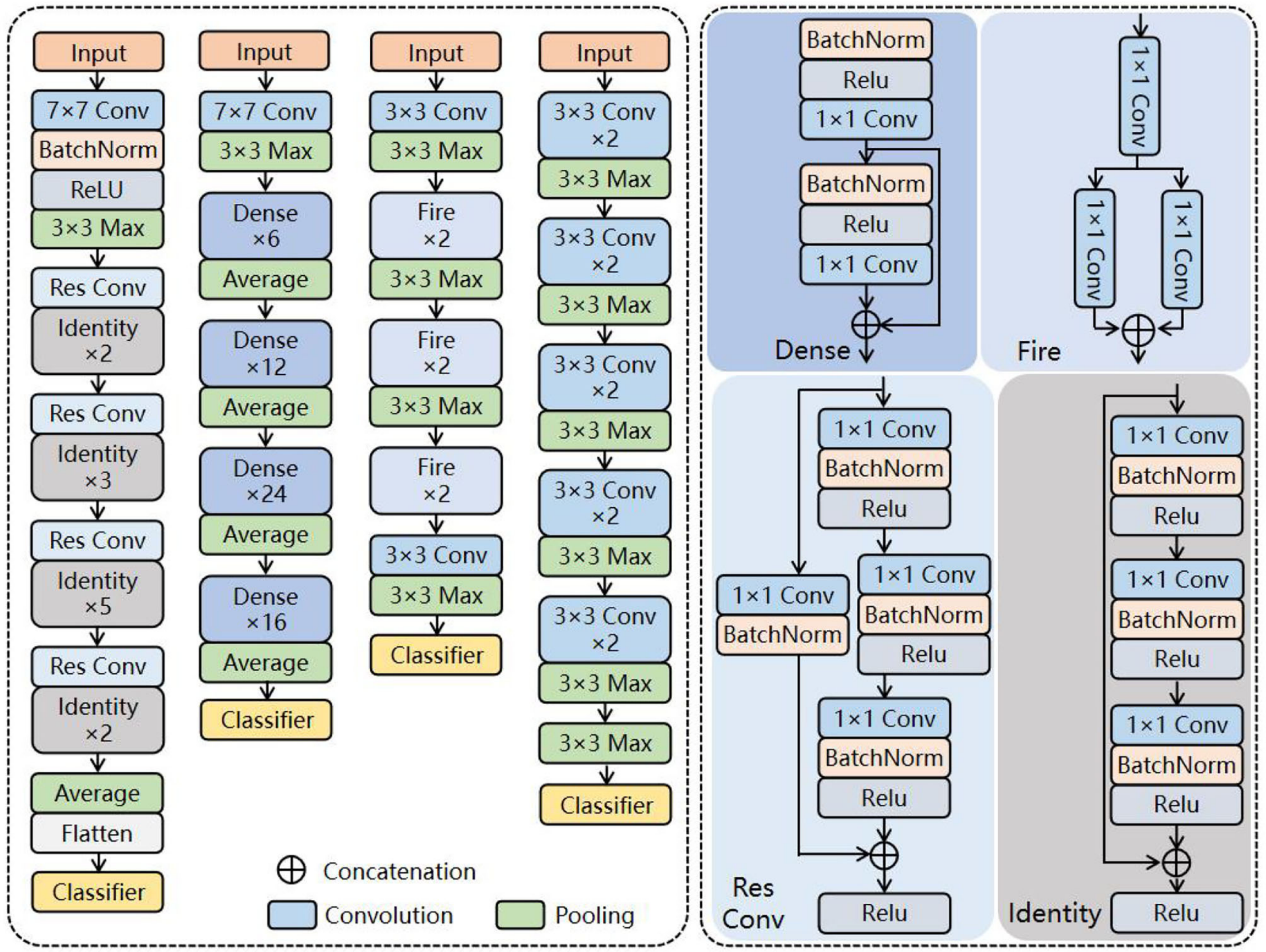


FIGURE 4 | Backbone of proposed architecture.

First, the input feature map $X_{in} \in \mathbb{R}^{H\times W\times C}$ is sent into a 1 × 1 convolutional layer to extract local features and change the dimension to match the next layer. The output of this layer is $X_1 \in \mathbb{R}^{H\times W\times C'}$, where (H, W) is the resolution of the initial image, C is the number of initial dimensions, and C ′ is the number of convoluted dimensions.

Second, a patch embedding process including image reshaping and image patch compression is performed. The feature map X_1 is reshaped into an N sequence of flattened 2D patches X_p^i (Equation 1):

$$X_p^i = P \times P \times C\, , i \in \{1, 2, \cdots, N\} \qquad (1)$$

where (P, P) is the resolution of each image patch, and N = H×W/P2 is the generating number of image patches. Then, X_p^i is compressed into a D-dimensional embedding space by a trainable linear projection for the MLP layer (Equation 2).

$$X_2 = \left[X_p^1E; X_p^2E; \cdots ; X_p^NE\right] + E_{pos}\ (2) \qquad (2)$$

where $E \in \mathbb{R}^{(P^2\times C')\times D}$ is the embedding projection of the patch, $E_{pos} \in \mathbb{R}^{N\times D}$ is the positional embedding, and X2 is the encoded image sequence.

Third, the processed data sequence X2 is transferred into the MLP layer (Equations 3, 4).

$$X2' = \mathrm{Dropout}(\mathrm{Gelu}(\mathrm{FC}(X2))) \qquad (3)$$

$$X3 = \mathrm{Dropout}(\mathrm{FC}(X2')) \qquad (4)$$

where Gelu and Dropout are activation functions used to prevent network overfitting and improve training accuracy. FC is a fully connected layer which transforms the convolution output of the two-dimensional feature map into a one-dimensional vector.

Finally, the output of the MLP layer is subsequently rearranged to the initial size of the input image $X_{out} \in \mathbb{R}^{H\times W\times C}$(Eq. 5), and the glaucoma category is predicted by a classifier.

$$\mathrm{Xout} = \mathrm{rearrange}(X3, (hw)(p1p2\,c) \rightarrow c(hp1)(wp2)) \qquad (5)$$

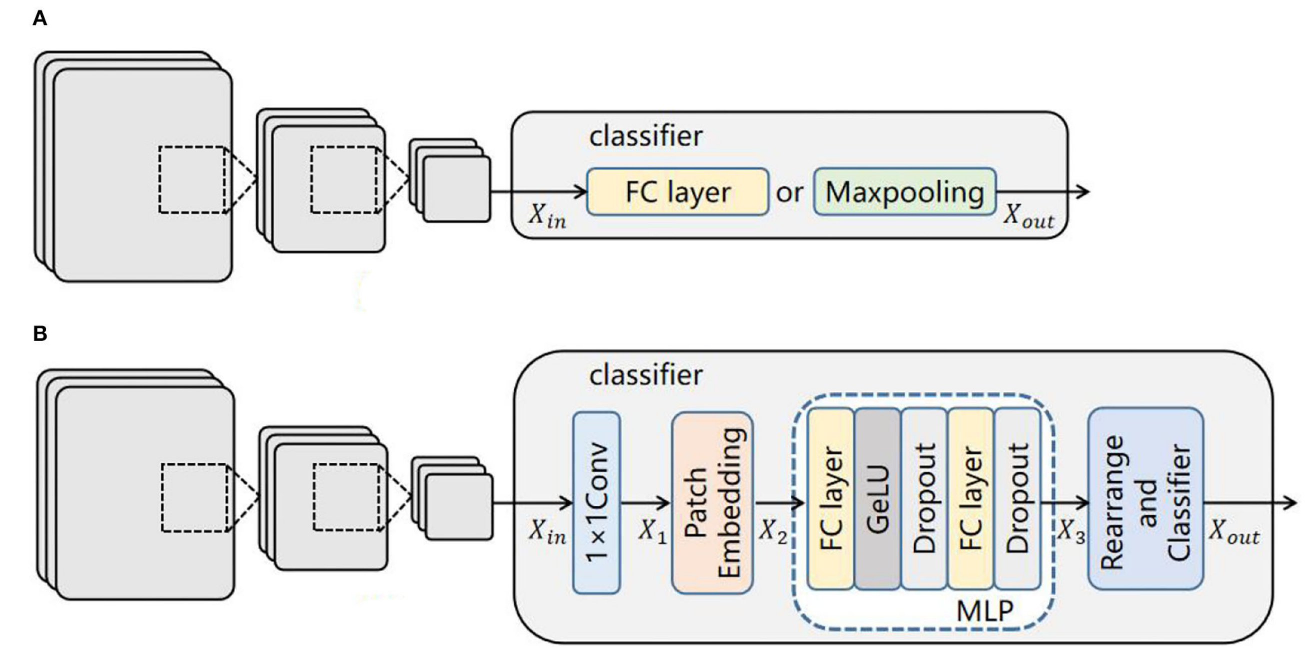


FIGURE 5 | Comparison of classifier structures: **(A)** classifier structure of CNNs; **(B)** our classifier structure.

TABLE 2 | Comparison of performances before and after reconstructed gray scale image.

	Ordinary gray image					Reconstructed gray scale image				
	Accuracy	**F1 score**	**Kappa**	**Jaccard**	**Recall**	**Accuracy**	**F1 score**	**Kappa**	**Jaccard**	**Recall**
SqueezeNet 1_1	0.772	0.753	0.690	0.623	0.772	0.793	0.779	0.724	0.652	0.793
Vgg 19	0.757	0.749	0.674	0.613	0.757	0.882	0.880	0.842	0.788	0.882
ResNet 50	0.797	0.795	0.729	0.665	0.797	0.918	0.918	0.890	0.849	0.918
DenseNet 121	0.790	0.787	0.720	0.659	0.790	0.888	0.889	0.849	0.803	0.888
Average*	0.779	0.771	0.703	0.640	0.779	0.870	0.866	0.826	0.773	0.870

Average = average value of above four CNNs.

TABLE 3 | Results of fundus images.

	Accuracy	**F1 score**	**Kappa**	**Jaccard**	**Recall**
SqueezeNet 1_1	0.696	0.662	0.595	0.528	0.696
Vgg 19	0.704	0.692	0.604	0.559	0.704
ResNet 50	0.687	0.682	0.581	0.534	0.687
DenseNet 121	0.716	0.707	0.622	0.559	0.716
Average	0.701	0.686	0.600	0.545	0.701

Evaluation Criteria

To evaluate the effectiveness of the proposed methods, we employ the accuracy, Jaccard score, Kappa score, recall, and F1 score. Accuracy indicates the proportion of the correct sample number in the total sample number. Recall represents the number of samples predicted to be positive out of the total number of true positive samples. The F1 score is the ratio of accuracy to recall. The Jaccard score evaluates the similarity and diversity of samples. The Kappa score assesses the consistency between the predicted classification results and actual results, and we employ it to evaluate the efficiency of multiclassification architectures.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Jaccard score} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

$$\text{F1 Score} = \frac{2 \bullet \text{precision} \bullet \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{P}_\text{e} = \frac{(\text{TP}+\text{FN})(\text{TP} + \text{FP})+(\text{TN} + \text{FN})\ (\text{TN} + \text{FP})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})^2}$$

$$\text{Kappa score} = \frac{\text{Accuracy} - \text{P}_\text{e}}{1 - P_e}$$

TABLE 4 | Results of multimodal fusion.

CNN model	Class no.	Acc	AUC	Spec	Sen	F1	Kappa	Avg.Acc	Avg.F1	Avg.AUC
SqueezeNet1_1	Class 0	0.948	0.965	1.0	0.930	0.909	0.873	0.896	0.895	0.931
	Class 1	0.926	0.866	0.743	0.990	0.839	0.792			
	Class 2	0.948	0.955	0.969	0.942	0.896	0.816			
	Class 3	0.970	0.939	0.879	1.0	0.935	0.916			
VGG 19	Class 0	0.956	0.970	1.0	0.940	0.921	0.890	0.911	0.910	0.956
	Class 1	0.948	0.900	0.800	1.0	0.889	0.856			
	Class 2	0.956	0.971	0.942	1.0	0.914	0.885			
	Class 3	0.963	0.924	0.848	1.0	0.918	0.894			
ResNet 50	Class 0	0.971	0.980	0.900	1.0	0.947	0.927	0.918	0.919	0.953
	Class 1	0.934	0.887	0.923	0.936	0.842	0.801			
	Class 2	0.934	0.963	0.848	0.978	0.897	0.848			
	Class 3	0.971	0.929	1.0	0.964	0.923	0.857			
DenseNet 121	Class 0	0.971	0.980	0.900	1.0	0.947	0.928	0.918	0.920	0.939
	Class 1	0.929	0.871	0.920	0.930	0.821	0.777			
	Class 2	0.907	0.963	0.848	0.936	0.857	0.788			
	Class 3	0.950	0.946	0.862	0.973	0.877	0.846			

where TP is true positive, indicating the number of images correctly classified by the classification algorithm; FN is false negative, indicating the number of images incorrectly classified into other categories by the classification algorithm; TN is true negative, indicating that the classification algorithm correctly classifies non-category images into other categories; and FP is false-positive, indicating that the classification algorithm incorrectly classifies non-category images into such categories.

EXPERIMENT AND DISCUSSION

In this section, the experimental setup of our study is introduced. Then, four experiments are conducted to present the effectiveness of our architecture. Finally, the results are shown and discussed in detail.

Experimental Setup

The experiments are conducted on a server equipped with an NVIDIA GeForce RTX 2060Ti graphic processing unit (GPU) and 16 GB of random-access memory. The compiler is PyCharm, the programming language is Python, and the experimental framework is PyTorch.

In this paper, the adaptive momentum estimation (Adam) optimizer is chosen to update the parameters of the proposed architecture, CrossEntropy Loss is set as the Loss function, and the learning rate is 0.0001. The epochs are set as 60, and the batch size is set as 8. Based on our newly constructed dataset, the proportion of the training set and testing set is set as 8:2; that is, 1,068 fundus and gray scale images are used as the training set, and 268 fundus and gray scale images are used as the testing set.

Experimental Results and Discussion

Comparison of Reconstructed Visual Field Gray Scale Images

In this section, to prove the superiority of the visual field gray scale image being reconstructed at higher resolution proposed in this paper, we conduct experiments on ordinary gray images and newly reconstructed gray scale images based on the proposed architecture. Meanwhile, evaluation criteria are employed to present the whole performance of the proposed multimodal classification architecture. The results are listed in **Table 2**.

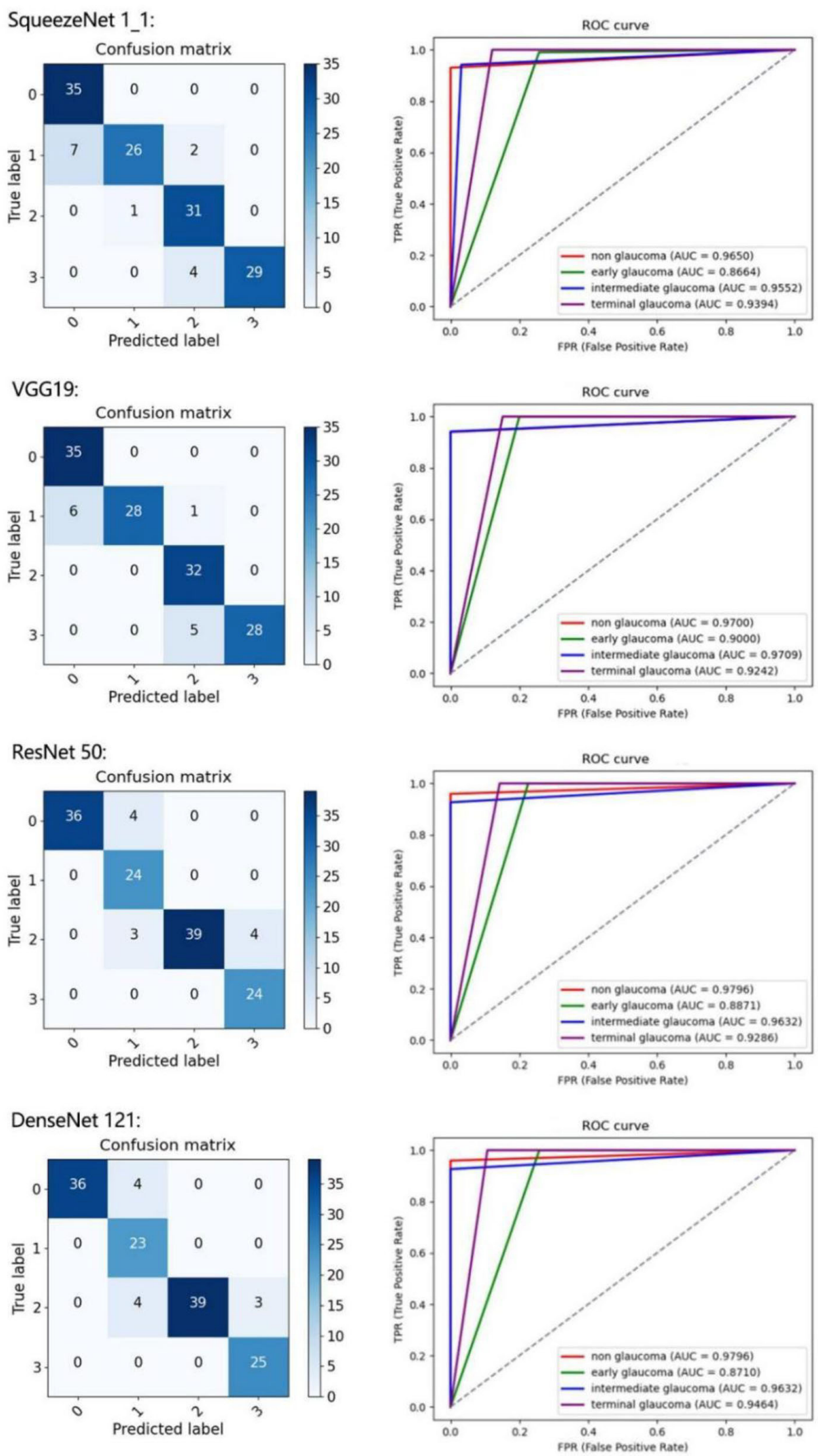
TABLE 5 | Ablation experiment of data augmentation.

	Augmentation	Accuracy	F1 score	Kappa	Jaccard	Recall
SqueezeNet 1_1	No	0.814	0.811	0.740	0.689	0.814
	Yes	0.896	0.895	0.862	0.812	0.896
Vgg 19	No	0.735	0.720	0.620	0.590	0.735
	Yes	0.911	0.910	0.881	0.836	0.911
ResNet 50	No	0.762	0.762	0.663	0.644	0.762
	Yes	0.918	0.919	0.889	0.852	0.918
DenseNet 121	No	0.812	0.812	0.736	0.699	0.812
	Yes	0.918	0.920	0.889	0.854	0.918

Table 2 indicates that the results of using the reconstructed gray scale image are more effective than the common gray scale image. The results of the proposed architecture are enhanced by 9.1, 9.6, 12.3, 13.3, and 9.1% in terms of average accuracy, F1 score, Kappa score, Jaccard score, and recall, respectively, compared with the results of common gray scale images. In particular, the accuracy of this task is enhanced by 12.1% by ResNet50. With these satisfying results, we draw the conclusion that the diagnostic architecture benefits from the reconstruction of the visual field gray scale image at higher resolution.

Comparison of Multimodal Fusion

In this section, two experiments are designed to present the effectiveness of multimodal fusion. The fundus image is first individually inputted to the proposed architecture, and then, the fundus image and the reconstructed gray scale image of the visual field are integrated into the multimodal fusion image and sent into the diagnostic architecture. The results are shown in



SqueezeNet 1_1:

	0	1	2	3
0	35	0	0	0
1	7	26	2	0
2	0	1	31	0
3	0	0	4	29

VGG19:

	0	1	2	3
0	35	0	0	0
1	6	28	1	0
2	0	0	32	0
3	0	0	5	28

ResNet 50:

	0	1	2	3
0	36	4	0	0
1	0	24	0	0
2	0	3	39	4
3	0	0	0	24

DenseNet 121:

	0	1	2	3
0	36	4	0	0
1	0	23	0	0
2	0	4	39	3
3	0	0	0	25

FIGURE 6 | Results of four classes on confusion matrix **(left)** and receiver operating characteristic (ROC) curves **(right)** for SqueezeNet1_1, VGG 19, ResNet 50, and DenseNet 121.

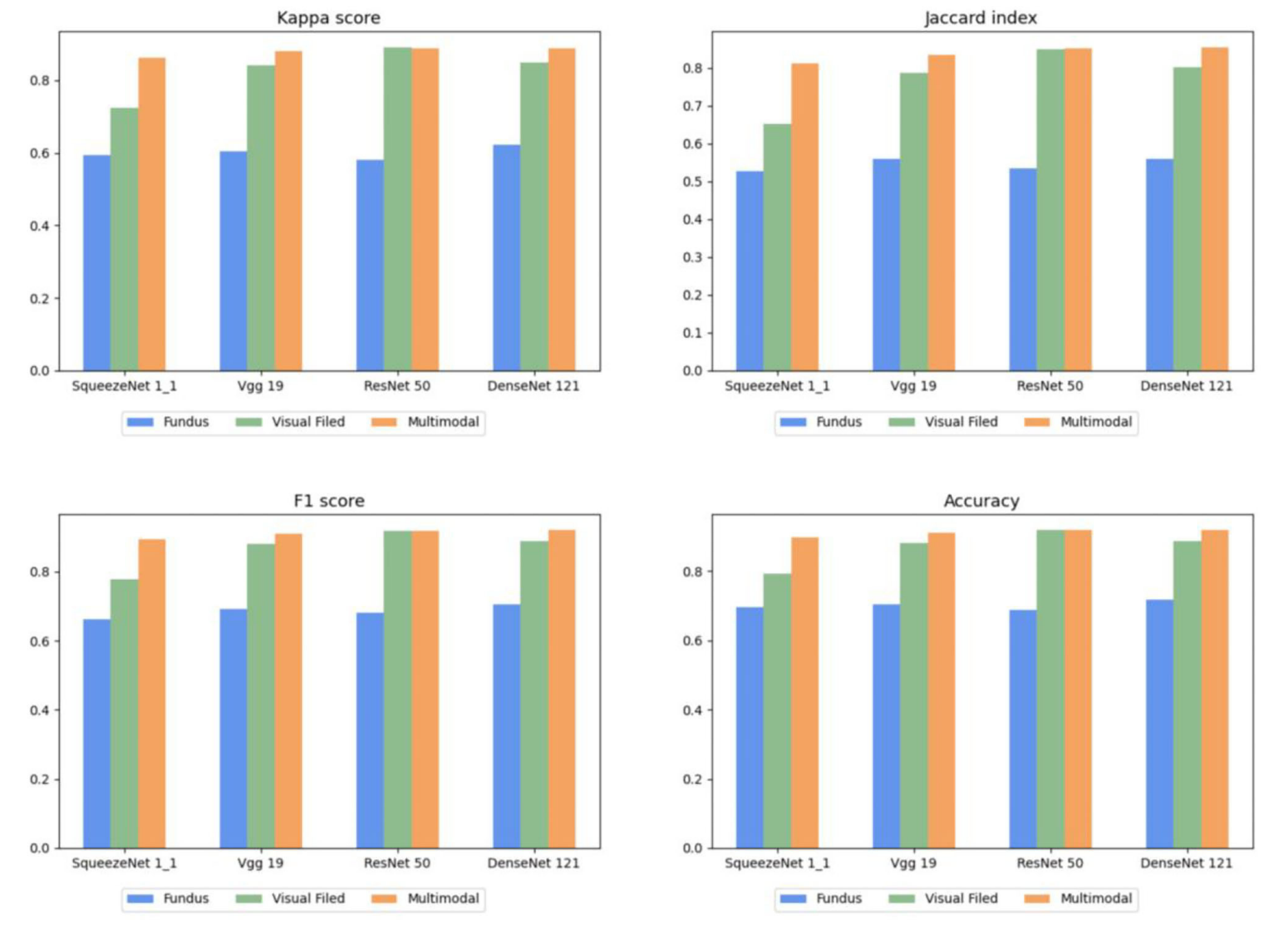


FIGURE 7 | Comparison of multimodal fusion and single path.

Tables 3, **4**. Finally, we compare **Tables 2–4** to verify the ability of multimodal fusion in the severity diagnosis of glaucoma.

By comparing **Tables 2–4**, the results of multimodal fusion data are better than single-path data: the accuracy of the above four CNNs achieves 89.6, 91.1, 91.8, and 91.8% in **Table 5**, and the average accuracy with 91.1% is higher than in **Table 2** (reconstructed gray scale image) with 87.0% and **Table 3** (fundus image) with 70.1%. The proposed architecture is enhanced by 4.5% in terms of the average F1 score compared with the results of the reconstructed gray scale image and 22.5% of the fundus image and improves by 5.4 and 28% in terms of the average kappa score. These results suggest that the proposed multimodal classification architecture is capable of superior diagnosis for glaucoma severity than a single type of data.

To further present the improvements of the proposed architecture, the classification results of each class are detailed in **Table 4**. We calculate the confusion matrix, AUC (**Figure 6**), and values for all the evaluation criteria including accuracy (Acc), sensitivity (Sen), specificity (Spec), Kappa score, and F1-score. Every CNN represents unique performance in the testing of glaucoma data. For instance, using DenseNet 121 led to the highest level of ordered pairs of (i) average accuracy and (ii) average F1-score of 91.8 and 91.2%, respectively, but its average AUC was lower than those of VGG 19 and ResNet 50.

TABLE 6 | Ablation experiment of proposed classifier.

	Accuracy	F1 score	Kappa	Jaccard	Recall
SqueezeNet 1_1	0.889	0.890	0.853	0.811	0.889
SqueezeNet 1_1+Classifier	0.901	0.900	0.868	0.820	0.901
Vgg 19	0.864	0.863	0.818	0.765	0.864
Vgg 19+Classifier	0.911	0.911	0.881	0.837	0.911
ResNet 50	0.882	0.883	0.851	0.847	0.882
ResNet 50+Classifier	0.924	0.924	0.897	0.862	0.924
DenseNet 121	0.913	0.911	0.886	0.844	0.913
DenseNet 121+Classifier	0.939	0.939	0.917	0.889	0.939

To describe this comparison more clearly, the histograms of **Tables 2–4** are shown in **Figure 7**, in which each evaluation metric of different CNNs (SqueezeNet1_1, VGG 19, ResNet 50,

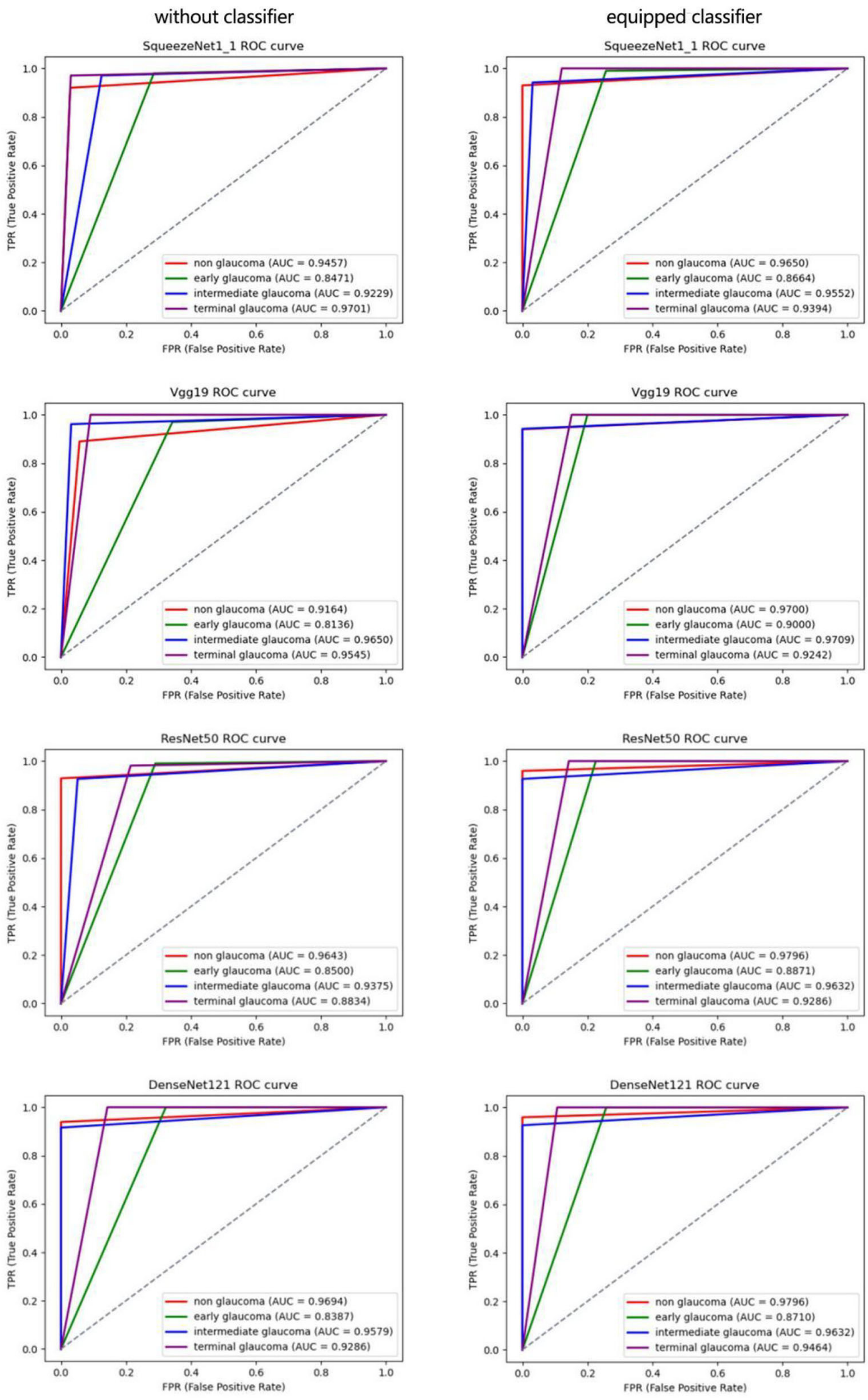


FIGURE 8 | Receiver operating characteristic curves of each subcategory for 4-category classification deep CNN.

TABLE 7 | Comparison of analogous approaches.

	Accuracy	AUC	Kappa	spec	Sen
Bizios et al. (2011)	0.9539	0.978	–	–	–
Chen et al. (2019)	0.9688	0.99	–	1.000	0.9167
Liu et al. (2014)	–	0.869	–	–	–
Ours	0.975	0.992	0.942	0.992	0.957

and DenseNet 121) is compared. Based on **Figure 7**, the same conclusion as above can be drawn.

Ablation Study

Ablation Study of Data Augmentation

In this section, we conduct an ablation experiment to prove the effectiveness of data augmentation technology. The results are shown in **Table 5**.

Table 5 compares the performance with or without data augmentation, and apparent improvements are obtained in all evaluation criteria. These results demonstrate that data augmentation technology has strong ability in the task of glaucoma classification.

Ablation Study of Proposed Classifier

In this section, we conduct an ablation experiment to prove the effectiveness of the proposed classifier, and the results are shown in **Table 6**.

In this section, 5-fold cross-validation is used to evaluate the performance of the proposed classifier in the above CNNs. **Table 6** lists the average results of the conducted experiments, which demonstrates that various evaluation metrics of these CNNs are improved to different degrees with the proposed classifier. Furthermore, our classifier can be flexibly plugged into common CNNs to integrate global features of images to enhance the performance in the diagnosis of glaucoma. The same conclusion can be drawn on the combination of multimodal classification architecture and the classifier.

To present the efficiency of the proposed classifier more clearly, we use the ROC curve to describe the results of each class in **Figure 8**. The AUC value can effectively measure the performance of the algorithm, which is defined as the area under the ROC curve. According to **Figure 8**, the AUC values of normal, early glaucoma, intermediate, and terminal glaucoma are improved to different degrees by each algorithm with the proposed classifier.

Comparison of Analogous Approaches

To prove the superiority of the proposed multimodal classification architecture over analogous approaches (Bizios et al., 2011; Chen et al., 2019), we compare the results for the same diagnosis task.

Table 7 shows that the proposed architecture achieves the best results with 0.975 in terms of average accuracy in the classification task of normal and glaucoma. This further demonstrates the advantage of the proposed multimodal classification architecture in glaucoma diagnosis.

CONCLUSION AND OUTLOOK

In this paper, we proposed a multimodal classification architecture based on deep learning for glaucoma severity diagnosis. The advantages of the framework are as follows: (1) More subtle gray scale units and corresponding gray scale images are reconstructed to address the limitation that the inferior resolution of common visual field gray scale images affects feature extraction in the task of glaucoma diagnosis. (2) Fundus images and reconstructed gray scale images of the visual field are fused as multimodal fusion images for the severity classification of glaucoma. Through experiments, we precisely distinguished the severity of glaucoma as normal, early, intermediate, and terminal by the proposed architecture, which yielded a significant contribution in clinical diagnosis. Meanwhile, we can see that the architecture based on the multimodal fusion image performs much better than the single-path architecture, which means that the multimodal fusion input improves the classification ability of the architecture. (3) We proposed a plug-and-play classifier to offset the CNNs' limitation of extracting global sequence information. This significantly improved the architecture's function of feature extraction. Experimental results demonstrated that with our classifier, regardless of what network is chosen as the architecture's backbone, the performance of the architecture is enhanced significantly.

There are many glaucoma patients worldwide, and the detection of the severity is very difficult, which results in a heavy burden and consumes considerable time for ophthalmologists. The proposed diagnosis architecture designed for the severity classification of glaucoma can be very convenient. In the future, we will collect more valid data such as OCT and try to integrate the retinal nerve fiber layer into our architecture to better classify the severity of glaucoma.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

This study was reviewed and approved by the Ethics Committee of the First Affiliated Hospital of Kunming Medical University, Kunming, China. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This work was supported by the National Natural Science Foundation of China (no. 81960176) and the National Natural Science Foundation of China (no. 82160347).

REFERENCES

Bajwa, M. N., Malik, M. I., Siddiqui, S. A., Dengel, A., Shafait, F., Neumeier, W., et al. (2020). Two-stage framework for optic disc localization and glaucoma classification in retinal fundus images using deep learning. *BMC Medical Inform. Decis. Mak.* 19, 1–16. doi: 10.1186/s12911-019-0842-8

Bizios, D., Heijl, A., and Bengtsson, B. (2011). Integration and fusion of standard automated perimetry and optical coherence tomography data for improved automated glaucoma diagnostics. *BMC Ophthalmol.* 11, 1–11. doi: 10.1186/1471-2415-11-20

Bowd, C., Belghith, A., Zangwill, L. M., Christopher, M., and Goldbaum, M. H. (2022). Deep learning image analysis of optical coherence tomography angiography measured vessel density improves classification of healthy and glaucoma eyes. *Am. J. Ophthalmol.* 236, 298–308. doi: 10.1016/j.ajo.2021.11.008

Chan, H. H., Ong, D. N., Kong, Y. X. G., O'Neill, E. C., Pandav, S. S., Coote, M. A., et al. (2014). Glaucomatous optic neuropathy evaluation (GONE) project: the effect of monoscopic versus stereoscopic viewing conditions on optic nerve evaluation. *Am. J. Ophthalmol.* 157, 936–944. doi: 10.1016/j.ajo.2014.01.024

Chen, Z., Zheng, X., Shen, H., Zeng, Z. and Liu, Q. (2019). Combination of enhanced depth imaging optical coherence tomography and fundus images for glaucoma screening. *J. Med. Syst.* 43, 1–12. doi: 10.1007/s10916-019-1303-8

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., and Houlsby, N. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. Haifa.

Escamez, C., Giral, E. M., Martinez, S. P., and Fernandez, N. T. (2021). High interpretable machine learning classifier for early glaucoma diagnosis. *Int. J. Ophthalmol.* 14, 393–398. doi: 10.18240/ijo.2021.03.10

Gour, N., and Khanna, P. (2020). Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network. *Biomed. Signal Process. Control* 66:102329. doi: 10.1016/j.bspc.2020.102329

Guo, F., Li, W., Tang, J., Zou, B., and Fan, Z. (2020). Automated glaucoma screening method based on image segmentation and feature extraction. *Med. Biol. Eng. Comput.* 58, 2567–2586. doi: 10.1007/s11517-020-02237-2

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern* (Las Vegas, NV: Recognition), 770–778.

Huang, G., Liu, Z., Laurens, V., and Weinberger, K. Q. (2016). *Densely Connected Convolutional Networks*. Las Vegas, NV: IEEE Computer Society.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., Keutzer, K., et al. (2017). "SqueezeNet: AlexNet-Level accuracy with 50x fewer parameters and <0.5MB model size," in *International Conference on Learning Representations (ICLR)* (Haifa). doi: 10.48550/arXiv.1602.07360

Ibrahim, M. H., Hacibeyoglu, M., Agaoglu, A., and Ucar, F. (2022). Glaucoma disease diagnosis with an artifcial algae-based deep learning algorithm. *Med. Biol. Eng. Comp*. 60, 785–796. doi: 10.1007/s11517-022-02510-6

Kang, H., Li, X., and Su, X. (2020). Cup-disc and retinal nerve fiber layer features fusion for diagnosis glaucoma. *Comp. Aided Diagn*. 11314, 945–953. doi: 10.1117/12.2548546

Kim, M., Janssens, O., Park, H. M., Zuallaert, J., Hoecke, S. V., and Neve, W. D. (2018). Web applicable computer-aided diagnosis of glaucoma using deep learning. doi: 10.1109/BIBM.2018.8621168

Krizhevsky, A., Sutskever, I., and Hinton, G. (2017). ImageNet classification with deep convolutional neural networks. *ACM* 60, 84–90. doi: 10.1145/3065386

Li, F., Wang, Z., Qu, G., Song, D., Yuan, Y., and Xu, Y. (2019). Correction to: automatic differentiation of Glaucoma visual field from non-glaucoma visual field using deep convolutional neural network. *BMC Med. Imaging* 19, 1. doi: 10.1186/s12880-019-0339-z

Liu, J., Xu, Y., Cheng, J., Zhang, Z., Wong, D. W. K., Yin, F., et al. (2014). Multiple modality fusion for glaucoma diagnosis. *IFMBE Proc.* 42, 5–8. doi: 10.1007/978-3-319-03005-0_2

Melas-Kyriazi, L. (2021). Do you even need attention? a stack of feed-forward layers does surprisingly well on imagenet. *arXiv [Preprint]*. 1–3. doi: 10.48550/arXiv.2105.02723

Mirzania, D., Thompson, A. C., and Muir, K. W. (2020). Applications of deep learning in detection of glaucoma: a systematic review. *Eur. J. Ophthalmol.* 31:112067212097734. doi: 10.1177/1120672120977346

Raja, H., Akram, M. U., Shaukat, A., Khan, S. A., and Nazir, N. (2020). Extraction of retinal layers through convolution neural network (CNN) in an OCTImage for glaucoma diagnosis. *J. Digit. Imaging*. 33, 1428–1442. doi: 10.1007/s10278-020-00383-5

Rajendrababu, S., Bansal, O., Shroff, S., Senthilkumar, V. A., and Uduman, M. S. (2021). Visual field-based grading of disease severity in newly diagnosed primary open angle glaucoma patients presenting to a tertiary eye care centre in India. *Int. Ophthalmol.* 41, 1–9. doi: 10.1007/s10792-021-01878-y

Simonyan, K., and Zisserman, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition. Computer Science*. Haifa.

Singh, L. K., Garg, H., Khanna, M., and Bhadoria, R. S. (2021). An enhanced deep image model for glaucoma diagnosis using feature-based detection in retinal fundus. *Med. Biol. Eng. Comp.* 59, 1–21. doi: 10.1007/s11517-020-02307-5

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst*. 5998–6008. doi: 10.48550/arXiv.1706.03762

Wang, J., Yang, L., Huo, Z., He, W., and Luo, J. (2020). Multi-label classification of fundus images with EfficientNet. *IEEEAccess* 8, 212499–212508. doi: 10.1109/ACCESS.2020.3040275

Wroblewski, D., Francis, B. A., Chopra, V., Kawji, A. S., Quiros, P., Dustin, L., et al. (2009). Glaucoma detection and evaluation through pattern recognition in standard automated perimetry data. *Graefes Arch. Clin. Exp. Ophthalmol.* 247, 1517. doi: 10.1007/s00417-009-1121-7

Yang, H. K., Kim, Y. J., Sung, J. Y., Dong, H. K., and Hwang, J. M. (2020). Efficacy for differentiating nonglaucomatous versus glaucomatous optic neuropathy using deep learning systems. *Am. J. Ophthalmol.* 216, 140–146. doi: 10.1016/j.ajo.2020.03.035

Yao, C., Hu, M., Zhai, G., and Zhang, X. P. (2021). *TransClaw U-Net: Claw U-Net with Transformers for Medical Image Segmentation*. Shanghai.

Zheng, C., Johnson, T. V., Garg, A., and Boland, M. V. (2019). Artificial intelligence in glaucoma. *Curr. Opin. Ophthalmol.* 30, 97–103. doi: 10.1097/ICU.0000000000000552

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.

frontiers | Frontiers in Neuroscience

ORIGINAL RESEARCH
published: 18 July 2022
doi: 10.3389/fnins.2022.937861

A Disentangled Representation Based Brain Image Fusion *via* Group Lasso Penalty

Anqi Wang [1,2], Xiaoqing Luo [1,2], Zhancheng Zhang [3] and Xiao-Jun Wu [1,2]*

[1] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China, [2] Advanced Technology and Research Institute, Jiangnan University, Wuxi, China, [3] School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou, China

Complementary and redundant relationships inherently exist between multi-modal medical images captured from the same brain. Fusion processes conducted on intermingled representations can cause information distortion and the loss of discriminative modality information. To fully exploit the interdependency between source images for better feature representation and improve the fusion accuracy, we present the multi-modal brain medical image fusion method in a disentangled pipeline under the deep learning framework. A three-branch auto-encoder with two complementary branches and a redundant branch is designed to extract the exclusive modality features and common structure features from input images. Especially, to promote the disentanglement of complement and redundancy, a complementary group lasso penalty is proposed to constrain the extracted feature maps. Then, based on the disentangled representations, different fusion strategies are adopted for complementary features and redundant features, respectively. The experiments demonstrate the superior performance of the proposed fusion method in terms of structure preservation, visual quality, and running efficiency.

Keywords: deep learning, image fusion, medical brain image, disentangled representation, group lasso penalty

OPEN ACCESS

Edited by:
Yu Liu,
Hefei University of Technology, China

Reviewed by:
Jun Wang,
Shanghai University, China
Xiangzhi Bai,
Beihang University, China
Jianping Gou,
Jiangsu University, China

****Correspondence:***
Xiaoqing Luo
xqluo@jiangnan.edu.cn

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *06 May 2022*
Accepted: *20 June 2022*
Published: *18 July 2022*

Citation:
Wang A, Luo X, Zhang Z and Wu X-J (2022) A Disentangled Representation Based Brain Image Fusion via Group Lasso Penalty. Front. Neurosci. 16:937861. doi: 10.3389/fnins.2022.937861

1. INTRODUCTION

Medical image fusion is an important branch of information fusion tasks. Typical types of medical images include Magnetic Resonance Imaging (MRI),Computed Tomography (CT), and Positron Emission Tomography (PET). MRI images are of high resolution and provide precise information about soft tissue, CT images provide dense structures like bones, and PET images assess the functions of organs and tissue. The objective of medical image fusion is to combine the complementary and redundant features from multi-modal medical images into one composite image with all the significant information, thus facilitating the process of clinical diagnosis. Image fusion methods can be generally divided into traditional ones and deep learning-based ones.

Traditional multi-scale transform (MST) based image fusion methods are popular in the community as the MST tools are able to simulate the human visual system to analyze the image, as well as to extract geometry structure and details of the image. Commonly adopted MST tools include discrete wavelet transform(DWT) (Ben et al., 2005), shift-invariant shearlet transform (Luo et al., 2016), and contourlet transform (Yang et al., 2010). Fused images with good quality can be obtained through the appropriate manual design of activity level measurements and fusion rules on the extracted features. However, to get better fusion performance, the manual design of fusion rules tends to become more and more complex, which results in higher computation costs.

Compared to the traditional methods, deep learning-based methods have been demonstrated with the great ability to automatically extract hierarchical and representative features of different abstraction levels. The typical deep learning model used for image fusion is Convolutional Neural Networks (CNN). Liu et al. (2017) applied CNN in image fusion, where the CNN predicts the importance of each pixel of source images. With the output decision map, source images are combined to get the fused image. Li et al. (2018) adopted the VGGNet pre-trained on the ImageNet dataset to extract the features from high frequency coefficients, which can effectively reflect the regions with abundant information. While these methods partially depend on the CNN and extra manual processes are required. To realize the end-to-end image fusion process, some unsupervised CNN-based methods and Generative adversarial Network (GAN) based methods are proposed subsequently (Huang et al., 2020; Ma et al., 2020; Xu and Ma, 2021; Guo et al., 2022; Xu et al., 2022). As an example for each category, Xu and Ma (2021) adopted both subjectively defined features and deep features to measure the activity level of input images, then adaptive weights can be assigned to loss functions to adjust the similarity between the fused image and each source image; Ma et al. (2020) proposed the DDcGAN which establishes the adversarial relationships between a generator and two discriminators to introduce abundant information from the source images of both modalities. Another popular pipeline for image fusion is to fuse the deep features extracted from an auto-encoder which has great feature extraction and image reconstruction abilities (Li and Wu, 2018; Li et al., 2020; Jian et al., 2021). Even though state-of-the-art performance has been achieved, the above methods leverage the same feature representation for different modalities to design the fusion rule or directly fuse the multi-modal features in an intermingled way, thus they cannot fully exploit the prior knowledge of complementary and redundant contained in multi-modal images. Redundant information is the common type of features such as structure and shape, while complementary information represents the most unique characteristics belonging to one specific modality, which is hierarchical and hard to represent by hand-crafted features. Thereby, fusion operations conducted on intermingled representations can cause the degradation of discriminative features and the introduction of distorted information.

The criteria for learning good representations discussed in Bengio et al. (2013) show that one of the important points is to disentangle the variable features for the explanatory factors. If exclusive representations can be obtained for multi-modal images to separate the complementary and redundant features, then the more interpretable representations can improve the accuracy of the fusion decision. Recently, some work has researched the disentanglement representations for image fusion. Xu et al. (2021) disentangled the features of infrared and visible images into attribute and scene modality, for each the weighted average fusion rule is adopted. Luo et al. (2021) believed that all kinds of paired source images share the private and common features, and proposed a general framework for image fusion that takes advantage of contrastive learning for better disentanglement. In the above two studies, the attribute and private features are exactly the complementary ones, while the scene and common features are the redundant ones. Both of them have alleviated the pressure of designing appropriate fusion strategies and achieved good fusion performance. However, there still exist some problems: (1) In Xu et al. (2021), the attribute modality is compressed into a vector, resulting in the loss of spatial information and lack of interpretability. Thereby, the weighted-average fusion rule on the attribute representation leads to blur results and information distortion. (2) Xu et al. (2021) force the infrared and visible attribute distribution close to a prior Gaussian distribution, while Luo et al. (2021) minimize the cosine similarity among private and common representations. Both of them lack the consideration of the importance of features in the local position of both source images, thus weakening the ability of disentangled representations to present the most meaningful information.

In order to achieve a more robust and controllable fusion decision, we aim to incorporate the explicit constraints on the deep feature maps extracted by the encoder. In the field of machine learning, feature selection is an important stage to reduce the data dimension and determine the relevant features for a specific learning task. Recently, sparsity-inducing regularization techniques are widely adopted in feature selection methods to filter out the irrelevant features from multiple heterogeneous feature descriptors (Zhao et al., 2015). Li et al. (2019) proposed an adaptive sparse group lasso penalty on the clustered genes to select the biologically significant genes. To control the attention response and restrain the noisy information, Wang and Guo (2021) applied sparse regularization on the computed attention maps. Considering the redundancy may exist among features, Wang et al. (2021) proposed using Group lasso to prevent the selection of redundant features which may have high correlations with other features. Inspired by these studies, we think the learning process of complementary and redundant representations can also be realized through the regularization techniques on the extracted feature maps to filter out the complementary features from the redundant ones.

Based on the above considerations, we propose a disentangled representation based brain image fusion method *via* group lasso penalty. A three-branch auto-encoder with two complementary branches and one redundant branch is designed to deal with the unique modality characteristics and common structure information inherent in the multi-modal source images. In the training stage, the auto-encoder should be able to reconstruct both source images conditioned on the extracted complementary features and redundant features. For effective disentangled representation learning, a complementary group lasso penalty is proposed to restrain the redundant information in the complementary features, promoting the complementary encoders to learn the most discriminative information. In the fusion stage, different fusion strategies are adopted for complementary and redundant features respectively. Then, the fused image can be obtained by reconstructing from the fused features. To sum up, the contributions of the proposed method are as follows:

- A disentangled representation based brain image fusion method is proposed to fully exploit the redundancy and complement prior relationships among multi-modal source images.
- A complementary group lasso penalty is designed to promote the disentanglement ability and ensure the complementary feature maps of significant modality information.
- Comparison experiments conducted on MRI-CT and MRI-PET fusion tasks with state-of-the-art deep learning-based methods demonstrate the superior fusion performance of the proposed method quantitatively and qualitatively.

The remaining part of the article is organized as follows. Section 2 briefly introduces the definition of group lasso penalty. Section 3 describes the proposed method in detail. The experiment results are shown in Section 4. The conclusion and an outlook of future study is presented in Section 5. The implementation code of the proposed model will be available on our project page.

2. GROUP LASSO PENALTY

In 2006, Yuan (2006) proposed the group lasso penalty in a linear model, which aims to select the grouped explanatory variables for the accurate prediction of a regression problem. Given a response variable $y \in R^N$, a feature matrix $X \in R^{N\times P}$, and a coefficient vector $\beta \in R^P$, where P is the number of feature variables and N is the number of observation values, the objective of the group lasso estimation model is defined as follows:

$$\arg\min_{\beta \in R^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l}\|\beta_l\|_2. \tag{1}$$

Here, the first term is the loss function and the second term is the group lasso penalty. P feature variables are further divided into L sub groups, each group contains p_l variables, β_l is the coefficient sub vector corresponding to the l^{th} group ($l = 1, 2, ..., L$), $\lambda \geq 0$ is a tuning parameter, $\|\cdot\|_2$ is the L2 norm. Group lasso penalty is able to exploit the group structure of variables and promote the selection of the most relevant feature variables, thus simplifying a model, avoiding overfitting, and enhancing the interpretability of a model.According to the context and requirements of a specific task, the loss function, grouping situation, and λ can be adjusted. Inspired by the effectiveness of the group lasso penalty in selecting significant features, we consider the feature vectors of different pixel positions in a feature map can be regarded as a feature waiting to be penalized, and a task like an image reconstruction can be regarded as the loss function in Equation 1. The difference is that the penalty in Equation 1 is imposed on the coefficients, while in this paper, the penalty is directly imposed on the extracted feature maps to filter out the redundant features from the complementary ones, thus promoting the accuracy of disentangled representations.

3. FRAMES AND METHODS

In this section, a detailed description of the disentangled representation based image fusion framework is given first. Then, the design of the loss functions and the adopted fusion strategies are described, respectively.

3.1. Overall Framework

The aim of the proposed method is to separate the complementary features from the redundant features for each modality, thus improving the interpretability of feature representation and the fusion accuracy. The overall framework of the proposed method is illustrated in **Figure 1**, which includes a training stage (**Figure 1A**) and a fusion stage (**Figure 1B**). The training stage is to train an auto-encoder to learn disentangled representation and image reconstruction ability, while the fusion stage is to get the fused image through fusing the disentangled representations. We denote that the input source images from two different modalities as I_1 and I_2, respectively. Since the complementary features contain the discriminative modality information and the redundant features contain the common structure information, two complementary encoders En_{C1} and En_{C2} is used to extract the unique information, respectively, and one shared redundant encoder En_R is designed to map the structure information into a common space.

In the training stage, I_1 and I_2 are encoded by the three encoders to get complementary and redundant features, respectively as follows:

$$\{C_*, R_*\} = \{En_{C*}(I_*), En_R(I_*)\}, * \in \{1, 2\}, \tag{2}$$

where C_* and R_* are the complementary and redundant features of I_*. Then, the input images should be able to be reconstructed from the combined features as follows:

$$\widetilde{I_*} = De_S(C_* + R_*), * \in \{1, 2\}, \tag{3}$$

where $\widetilde{I_*}$ is the reconstructed version of I_*. Besides, as the complementary features are expected to represent the most unique modality information and determine the appearance of an image, the output image should be as similar as possible to the input source image which provides the complementary features. The process is described as follows:

$$\begin{aligned} \widetilde{I_{1_2}} &= De_S(C_1 + R_2), \\ \widetilde{I_{2_1}} &= De_S(C_2 + R_1), \end{aligned} \tag{4}$$

where $\widetilde{I_{1_2}}$ is the reconstructed image I_1 conditioned on C_1 and R_2, while $\widetilde{I_{2_1}}$ has a similar definition. To achieve good image reconstruction ability, Mean Square Error (MSE) and Structural Similarity (SSIM) (Wang et al., 2004) are adopted as the image reconstruction loss. Only using the shared-weight strategy in En_R cannot guarantee the disentanglement, we adopted two kinds of constraints to improve the disentangled representation learning: a complementary group lasso penalty term and a redundant consistency constraint term, which are introduced in Section 3.2. The former is adopted to restrain the growth of redundant

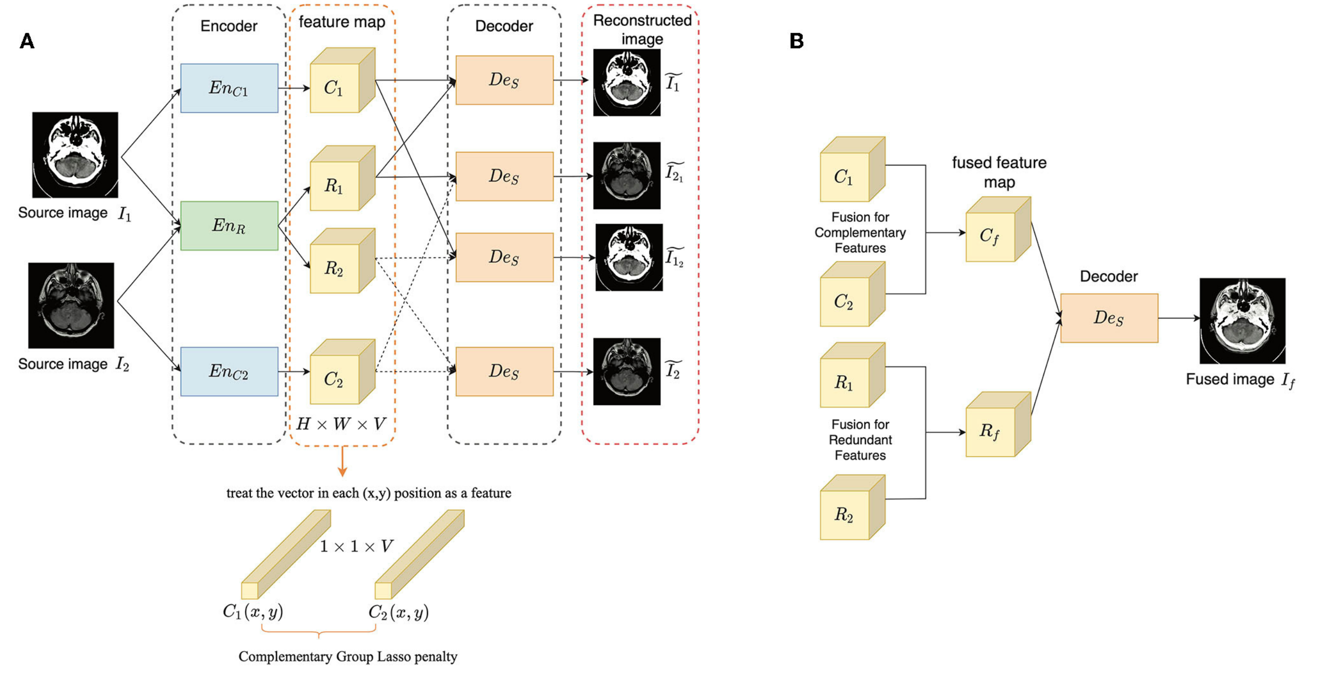


FIGURE 1 | The overview of the proposed method: **(A)** the training stage; **(B)** the fusion stage. The encoder-decoder architecture contains two complementary encoders En_{C1} and En_{C2}, one redundant encoder En_R and one shared decoder De_S. The extracted complementary and redundant features of the two source images are denoted as C_* and $R_*(* \in 1, 2)$, and each of them is of size $H \times W \times V$.

information in the extracted complementary feature maps, while the latter is designed based on the assumption that the multi-modal images captured in the same scene should share as much structure information as possible.

In the fusion stage, the complementary and redundant features are extracted from source images firstly as in the training stage, while before combining them, different fusion strategies (Section 3.3) are defined for them. After obtaining the fused complementary and redundant feature (C_f and R_f), they are added together and input to De_S to get the final fused image I_f as follow:

$$\widetilde{I_f} = De_S(C_f + R_f). \tag{5}$$

The input images are assumed as gray scale images. If the input is an *RGB* image, it is first converted into *YCbCr* color space, and the *Y*(luminance) component is used for fusion. After getting the gray scale fused image, it is combined with *Cb* and *Cr*(chrominance) components and inversely converted into the *RGB* fused image.

As for the network architecture, in each encoder, there are three 3 × 3 convolutional blocks with *ReLU* activation, except for the first one, each followed by a Batch Normalization layer. The weights of the first three layers in VGG-19 (Simonyan and Zisserman, 2015) are used to initialize the complementary and redundant encoders, as VGG-19 is a well-trained feature extractor that can relieve the training pressure. The architecture of the decoder is symmetric as the encoder, while in the output layer, Sigmoid is adopted as the activation function to constrain the value between [0,1]. Detailed information about the network is shown in **Table 1**.

3.2. Loss Function

1) Complementary group lasso penalty term: The extracted feature maps are considered with the size of $H \times W \times V$, where H, W, and V correspond to the height, width, and channel dimensions, respectively. Each $1 \times 1 \times V$ vector in position (x, y) is treated as a feature waiting to be penalized. We denote the features of I_1 and I_2 extracted by the complementary encoder in position (x, y) as $C_1(x, y)$ and $C_2(x, y)$. To determine the type of a feature, the similarity between $C_1(x, y)$ and $C_2(x, y)$ is computed by cosine similarity as follows:

$$r(x, y) = \frac{C_1(x, y) \cdot C_2(x, y)}{\|C_1(x, y)\|_2 \|C_2(x, y)\|_2}, \tag{6}$$

The high similarity means the information is redundant, on the contrary, complementary. The importance ϕ_* of a feature is measured based on the $L1$ norm and average operator in a local block around $C_*(x, y)$ as follow:

$$\phi_*(x, y) = \frac{\sum_{i=-r}^{r} \sum_{j=-r}^{r} \hat{C}_*(x+i, y+j)}{(2r+1)^2}, \tag{7}$$

where $\hat{C}_*(x, y)$ is the $L1$ norm of $C_*(x, y)$ computed as follows:

$$\hat{C}_*(x, y) = \|C_*(x, y)\|_1. \tag{8}$$

TABLE 1 | The architecture of encoder and decoder.

	Layer	Size	Stride	Channel (input)	Channel (output)	Activation	Normalization
Encoder	Conv1	3 x 3	1	1	64	ReLU	/
	Conv2	3 x 3	1	64	64	ReLU	Batch
	Conv3	3 x 3	1	64	128	ReLU	Batch
Decoder	Conv1	3 x 3	1	128	64	ReLU	Batch
	Conv2	3 x 3	1	64	64	ReLU	Batch
	Conv3	3 x 3	1	64	1	Sigmoid	/

Conv means the convolutional block with activation and normalization layer.

Then, a complementary Group lasso penalty L_c is proposed to restrain the redundancy and promote complement in C_1 and C_2:

$$L_c = \sum_{i=1}^{W \times H} (\omega_1 \| C_1(x,y) \|_2 + \omega_2 \| C_2(x,y) \|_2), \tag{9}$$

where ω_1 and ω_2 are defined in the form of a Sigmoid function as follows:

$$\begin{aligned} \omega_1 &= \frac{1}{1 + exp(k(\phi_2(x,y) - \phi_1(x,y)))}, \\ \omega_2 &= 1 - \omega_1. \end{aligned} \tag{10}$$

In Equation (10), k is the parameter that controls the shape of the function and is defined based on the similarity:

$$k = \frac{1}{r^2(x,y)}. \tag{11}$$

The smaller the similarity between $C_1(x,y)$ and $C_2(x,y)$ is, the larger the k is. Then, the shape of the sigmoid function becomes steeper.

Figure 2 shows the function of ω_1 in Equation 10. The smaller the similarity is, the closer the weight assignment is to choose-max, on the contrary, close to average-weighting. Then, the weight value is further determined by the $\phi_*(x,y)$. When Equation 9 is going to be minimized in an iteration if $\phi_1(x,y)$ is much larger than $\phi_2(x,y)$, which means $C_1(x,y)$ is much more important than $C_2(x,y)$. At this time, $\phi_1(x,y) - \phi_2(x,y)$ is a positive value, and ω_1 tends to become zero. Then, less penalty is imposed on $C_1(x,y)$, while $C_2(x,y)$ is greatly penalized and filtered out from the complementary feature maps. On the contrary, $C_1(x,y)$ is greatly penalized. If $C_1(x,y)$ is similar to $C_2(x,y)$, it means they share a lot of redundant information, and $\phi_1(x,y) - \phi_2(x,y)$ becomes close to zero. Thereby both of them are equally penalized and gradually pushed into $R_1(x,y)$ and $R_2(x,y)$. Finally, the complementary feature maps should contain the most significant modality characteristics.

2) Redundant consistency constraint term: As the multi-modal medical images are captured from the same brain, they must contain redundant information like structure and shape. It is expected that both R_1 and R_2 maintain a similar information. However, the multi-modal medical images provide an unequal amount of information, and they show their own biases toward some specific parts of the brain. Moreover, a shared En_R is adopted to extract the redundant feature, thus R_1 and R_2 cannot be the same. Compared to constraining the similarity of the extracted features, the redundant consistency constraint term L_r is conducted on the reconstructed results of R_1 and R_2 as follows:

$$L_r = \| De_S(R_1) - De_S(R_2) \|_1, \tag{12}$$

3) Image reconstruction loss: The image reconstruction loss is to enforce the output images to have high reconstructed precision with the input images, thus ensuring that the auto-encoder has both good feature extraction and image reconstruction ability. The image reconstruction loss L_{rec} is defined based on pixel loss L_{MSE} and SSIM (Wang et al., 2004) L_{SSIM} is as follows:

$$\begin{aligned} L_{MSE} &= \| I_1 - \widetilde{I_1} \|_2 + \| I_1 - \widetilde{I_{1_2}} \|_2 + \| I_2 - \widetilde{I_2} \|_2 + \| I_2 - \widetilde{I_{2_1}} \|_2, \\ L_{SSIM} &= (1 - SSIM(I_1, \widetilde{I_1})) + (1 - SSIM(I_1, \widetilde{I_{1_2}})) \\ &\quad + (1 - SSIM(I_2, \widetilde{I_2})) + (1 - SSIM(I_2, \widetilde{I_{2_1}})), \\ L_{rec} &= \lambda_{SSIM} L_{SSIM} + L_{MSE}, \end{aligned} \tag{13}$$

where λ_{SSIM} is the parameter to balance the pixel loss and SSIM loss.

Thus, the overall loss is defined as follows:

$$L = L_{rec} + \lambda_r L_r + \lambda_c L_c, \tag{14}$$

where λ_r and λ_c are the parameters to control the tradeoff of L_r and L_c.

3.3. Fusion Strategy

The complementary features are exclusive for each modality, here, three kinds of fusion strategies are considered, including the addition strategy, max-selection strategy, and $L1$-norm strategy. Their impact on the results is compared in Section 4. The addition strategy is formulated as follows:

$$C_f(x,y) = C_1(x,y) + C_2(x,y), \tag{15}$$

The max-selection strategy preserves the features of higher magnitude and is formulated as follows:

$$C_f(x,y) = \begin{cases} C_1(x,y), & C_1(x,y) \geq C_2(x,y), \\ C_2(x,y), & C_2(x,y) < C_1(x,y). \end{cases} \tag{16}$$

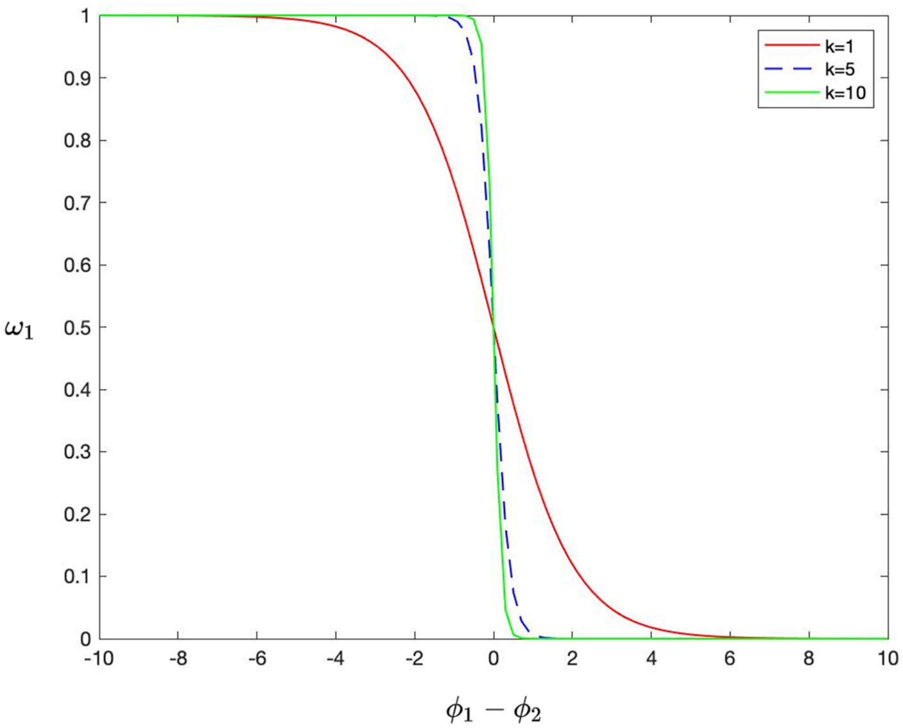


FIGURE 2 | The shape of the function of ω_1.

The $L1$-norm strategy is designed based on the importance of each pixel position to adjust the information preservation degree of each source image. The L1-norm of complementary feature maps is computed as Equation 8 and is treated as the activity level measurement $A_*(x, y), * \in \{1, 2\}$, then, the $L1$-norm strategy is formulated as follows:

$$C_f(x, y) = \mu_1 \times C_1(x, y) + \mu_2 \times C_2(x, y), \tag{17}$$

where

$$\begin{aligned} \mu_1 &= \frac{A_1(x, y)}{A_1(x, y) + A_2(x, y)}, \\ \mu_2 &= 1 - \mu_2. \end{aligned} \tag{18}$$

The redundant information is mapped to the same space, thereby, a simple average strategy is adopted as follows:

$$R_f = \frac{R_1(x, y) + R_2(x, y)}{2}. \tag{19}$$

The final fused image is reconstructed by decoding the added C_f and R_f.

4. EXPERIMENTS AND ANALYSES

In this section, we compare the proposed method with several typical deep learning-based image fusion methods on MRI-CT and MRI-PET image fusion tasks. First, the ablation study is conducted on the proposed complementary group lasso penalty term to verify its effectiveness. Then, the comparative study is conducted qualitatively and quantitatively. Finally, the time cost comparison of different methods is also conducted.

4.1. Experimental Settings

The training and testing dataset is built on the Harvard medical dataset (Summers, 2003), providing a brain image with a size 256×256. The slices with effective information are selected and there are a total of 180 pairs of MRI-CT images and 260 pairs of MRI-PET images. Considering that the number of the image is limited, when in the training phase, 10-fold verification experiments are performed and all the input images are randomly cropped into image patches of size 120 × 120, as well as randomly flipped and rotated. The setting of parameters are as follows: the batchsize is 8, the learning rate is 1e-4, and the size of a local block to measure the importance of a feature is 3, thus r is defined as 1. The other parameters like λ_{SSIM}, λ_r, and λ_c are set as 1,000, 10, and 10. The proposed method was implemented in Pytorch, and all experiments are conducted on a platform with Intel Core i7-6850K CPU and GeForce GTX 1080Ti GPU.

In the testing phase, the proposed method is compared with 6 deep learning-based methods, including CNN based methods EMFusion (Xu and Ma, 2021), U2Fusion (Xu et al., 2022), GAN based method DDcGAN (Ma et al., 2020), auto-encoder based method IFSR (Luo et al., 2021), DRF (Xu et al., 2021), and SEDR (Jian et al., 2021). All the code of the comparison methods are publicly available and the parameter settings are set according to the reference paper. Besides, the proposed method takes three different fusion strategies for the complementary features and they are also compared, which are denoted as proposed-add, proposed-max, and proposed-l1, respectively. For the proposed method, the average results of the quantitative evaluations and their corresponding variances of the 10 groups of the multi-fold verification experiments are presented in the table. The other comparison methods are also tested on the 10 groups respectively and the average values and variances of the 10 groups are computed.

4.2. Objective Metrics

Eleven objective metrics are adopted to conduct a comprehensive evaluation, including standard deviation (*SD*), spatial frequency (*SF*) (Ma et al., 2019), normalized mutual information (Q_{MI}) (Hossny et al., 2008), nonlinear correlation information entropy (Q_{NCIE}) (Qiang et al., 2005), gradient-based fusion performance (Q_G) (Xydeas and Pv, 2000), a multiscale scheme based metric (Q_M) (Wang and Liu, 2008), Piella's Metric (Q_S) (Piella and Heijmans, 2003), multi-scale structural similarity (*MSSSIM*) (Ma et al., 2015),the sum of the correlations of differences (*SCD*) (Aslantas and Bendes, 2015), Chen-Blum Metric (Q_{CB}) (Chen and Blum, 2009), and visual information fidelity based method (*VIFF*) (Han et al., 2013). Among them, *SD* reveals the distribution of gray levels and reflects the contrast of an image. *SF* measures the vertical and horizontal gradients, reflecting the changes in texture. Q_{MI} measures the amount of information transferred from source images to the fused images. Q_{NCIE} reveals the nonlinear correlation between source images and fused images. Q_G measures the amount of edge information transferred from source images to the fused images, while Q_M measures the amount of multi-scale edges. Both Q_S and *MSSSIM* reflect the structural similarity between source images and fused image, as well as quantifying the perceived distortion, while the former is edge-dependent and the latter is conducted based on multi-scale decomposition. *SCD* reveals how the complementary information is obtained by the fused image from source images. Q_{CB} and *VIFF* are human perception inspired metrics. Q_{CB} measures the similarity between source images and fused images based on the characteristics of a human visual system such as contrast and masking phenomenon, while *VIFF* measures the effective visual information contained in the fused image based on the natural scene statistics theory. A larger value of all the mentioned metrics corresponds to a good fusion performance.

4.3. Ablation Study

In this section, we verify the effectiveness of the complementary group lasso penalty term L_c. The proposed method trained without L_c is denoted as the proposed method without L_c, and the fusion evaluation is conducted based on the addition strategy. In **Figure 3**, the extracted feature maps of one MRI-CT sample and one MRI-PET sample is presented. It can be seen that the proposed method without L_c provides the redundant and complementary features (**Figures 3B,C**) quite similar to the source images, but with different pixel intensity, which means a relatively weak disentanglement ability. On the contrary, L_c is able to promote the disentanglement and extract the complementary features with sharper details (**Figure 3E**). From the fused results in **Figures 3F–I**, the edge and texture of (**Figures 3F,H**) are a bit blur, and (**Figure 3H**) loses a lot of MRI information. We also present the corresponding quantitative evaluation in **Tables 2**, **3**. L_c is able to improve the performance on almost all the metrics. In the MRI-PET task, proposed without L_c achieves the best Q_{CB}, which reveals that the fused results should have good visual contrast, while the results of the rest metrics show that there is much loss of details and structural information. The ablation study demonstrates the function of L_c to better exploit the complementary and redundant relationships among multi-modal images.

4.4. Qualitative Evaluation

Two typical pairs of MRI-CT images and two typical pairs of MRI-PET images are presented in **Figures 4**, **5**, respectively. MRI images depict accurate and abundant soft tissue, CT images provide dense structures with less distortion, and PET images provide a detailed function of focus of infection and metabolism information. From the visual results, it can be seen that the fused images of DDcGAN show a lot of distorted information in **Figures 4C,N**, and it almost loses all the MRI information in **Figures 5C,N**. This is caused by the instability of GAN, and it is inappropriate for the adopted loss function to represent the information of MRI as gradients only. The fused images of IFSR, U2Fusion, and SEDR lose much saliency of soft tissue and dense structures and present a low contrast on the whole. DRF provides relatively blurred results and loses a lot of sharp details. Besides, the color of the PET image is severely distorted in its fused results. Among these methods, U2Fusion measures the amount of gradient in each source image to assign the weights of the loss function, realizing the adaptive control of similarity between fused images and source images. However, such assignments are conducted evenly on the whole image, thus leading to the degradation of image contrast. SEDR maps both source images into the same space, ignoring the unique modality information. Fusion operations on such features can lead to loss of significance. IFSR and DRF all take into account the disentanglement, however, they lose the consideration of the corresponding relationship of source images in different positions. Moreover, DRF compresses the modality information into a vector, which can cause the distortion of spatial information. On the whole, EMFusion and the proposed method taking different fusion strategies can all provide the fused image with abundant details and clear edges. EMFusion is able to enhance the PET information with MRI details, while the proposed method can show the CT and MRI information with higher brightness.

4.5. Quantitative Evaluation

The quantitative results of MRI-CT and MRI-PET fusion tasks are presented in **Tables 4**, **5**. From the average values, DDcGAN obtains the best *SD* and *SF* in the MRI-CT task, which means

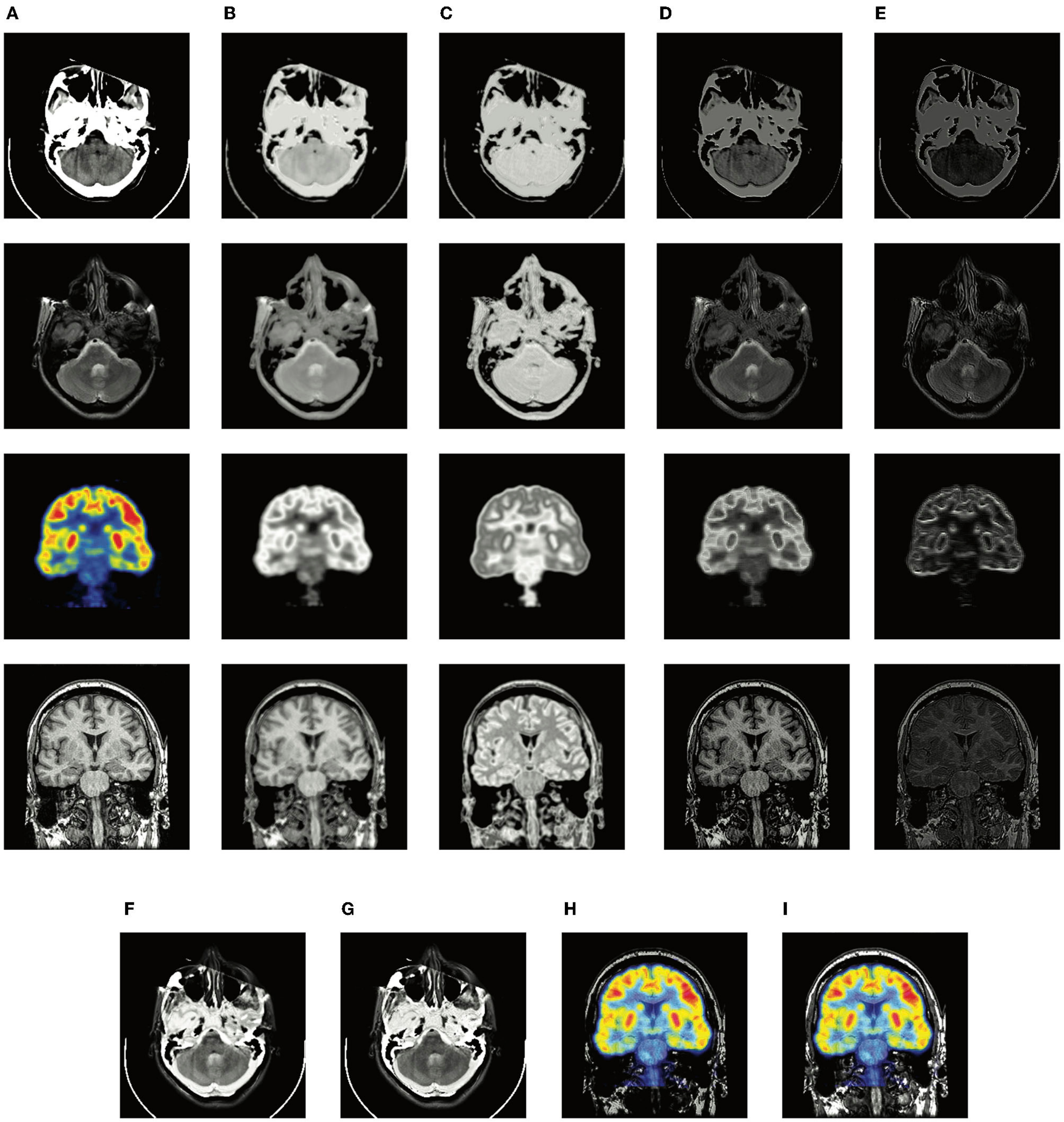


FIGURE 3 | The illustration of visualized feature maps and fused results without and with complementary group lasso penalty term L_c. **(A)** source images; **(B)** the redundant features of the proposed method without L_c; **(C)** the redundant features of the proposed method; **(D)** the complementary features of the proposed method without L_c; **(E)** the complementary features of the proposed method; **(F,H)** the fused results of the proposed-add without L_c; **(G,I)** the fused results of the proposed-add.

the fused results a higher dispersion degree of the gray value and many details of high frequency. However, SD and *SF* can only reflect the quality of the fused image itself and fails to measure the information transferred from source images, meanwhile, **Figures 4C,N** contains much distorted information. SEDR is able to achieve the best Q_{MI} and Q_{NCIE} in the MRI-CT fusion task, which reveals the fused image shows a higher correlation with both source images. But it shows weaker performance on MRI-CT tasks, as **Figures 4H,S** shows that the image contrast is degraded. EMFusion shows the best performance on Q_G, Q_S,

TABLE 2 | The quantitative evaluation of the proposed method (addition strategy) without and with L_c on the MRI-CT dataset.

Methods	Objective metrics										
	SD	***SF***	***Q_{MI}***	***Q_{NCIE}***	***Q_G***	***Q_M***	***Q_S***	***MSSSIM***	***SCD***	***Q_{CB}***	***VIFF***
proposed-add without Lc	80.38 ± 2.12	25.41 ± 1.30	0.78 ± 0.02	0.81 ± 0.00	0.68 ± 0.02	0.14 ± 0.01	0.65 ± 0.23	0.91 ± 0.01	1.12 ± 0.09	0.54 ± 0.19	0.43 ± 0.01
proposed-add	84.92 ± 2.64	27.23 ± 0.27	0.80 ± 0.01	0.81 ± 0.00	0.72 ± 0.01	0.16 ± 0.02	0.82 ± 0.01	0.91 ± 0.01	1.37 ± 0.10	0.66 ± 0.01	0.45 ± 0.03

The results of the proposed method are shown as average ± variance of 10-fold verification experiments.

TABLE 3 | The quantitative evaluation of the proposed method (addition strategy) without and with L_c on the MRI-PET dataset.

Methods	Objective metrics										
	SD	***SF***	***Q_{MI}***	***Q_{NCIE}***	***Q_G***	***Q_M***	***Q_S***	***MSSSIM***	***SCD***	***Q_{CB}***	***VIFF***
proposed-add without Lc	80.20 ± 1.75	26.19 ± 1.78	0.65 ± 0.01	0.81 ± 0.00	0.62 ± 0.02	0.17 ± 0.02	0.76 ± 0.02	0.91 ± 0.01	1.36 ± 0.05	0.58 ± 0.01	0.52 ± 0.01
proposed-add	89.28 ± 1.21	34.41 ± 0.55	0.76 ± 0.01	0.81 ± 0.00	0.77 ± 0.00	0.51 ± 0.08	0.80 ± 0.01	0.94 ± 0.00	1.65 ± 0.03	0.50 ± 0.01	0.59 ± 0.00

The results of the proposed method are shown as average ± variance of 10-fold verification experiments.

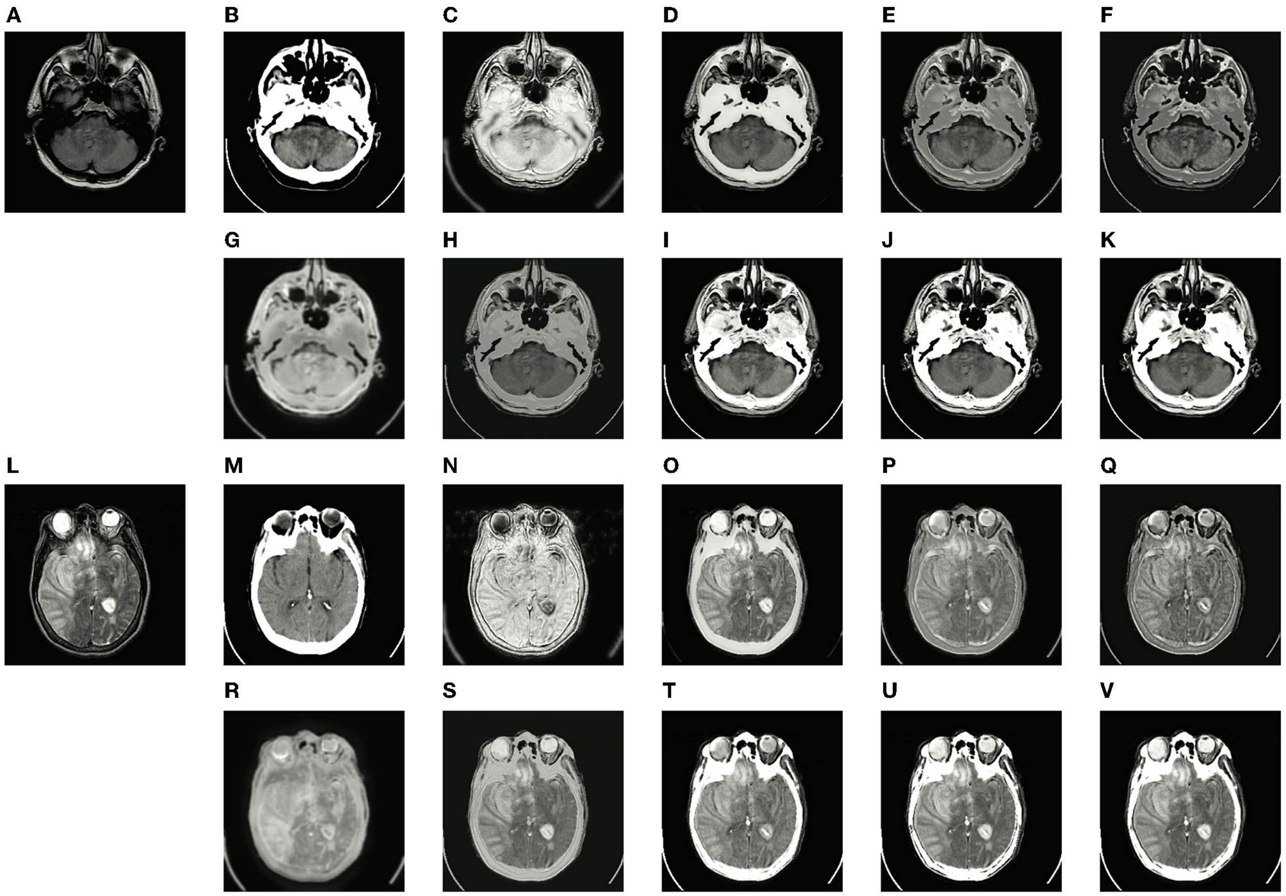


FIGURE 4 | Experiments results of the proposed method with six deep learning-based methods on two typical MRI and CT image pairs. **(A,L)** MRI images; **(B,M)** CT images; **(C,N)** fused results of DDcGAN; **(D,O)** fused results of EMFusion; **(E,P)** fused results of IFSR; **(F,Q)** fused results of U2Fusion; **(G,R)** fused results of DRF; **(H,S)** fused results of SEDR; **(I,T)** fused results of proposed-add; **(J,U)** fused results of proposed-max; and **(K,V)** fused results of proposed-l1.

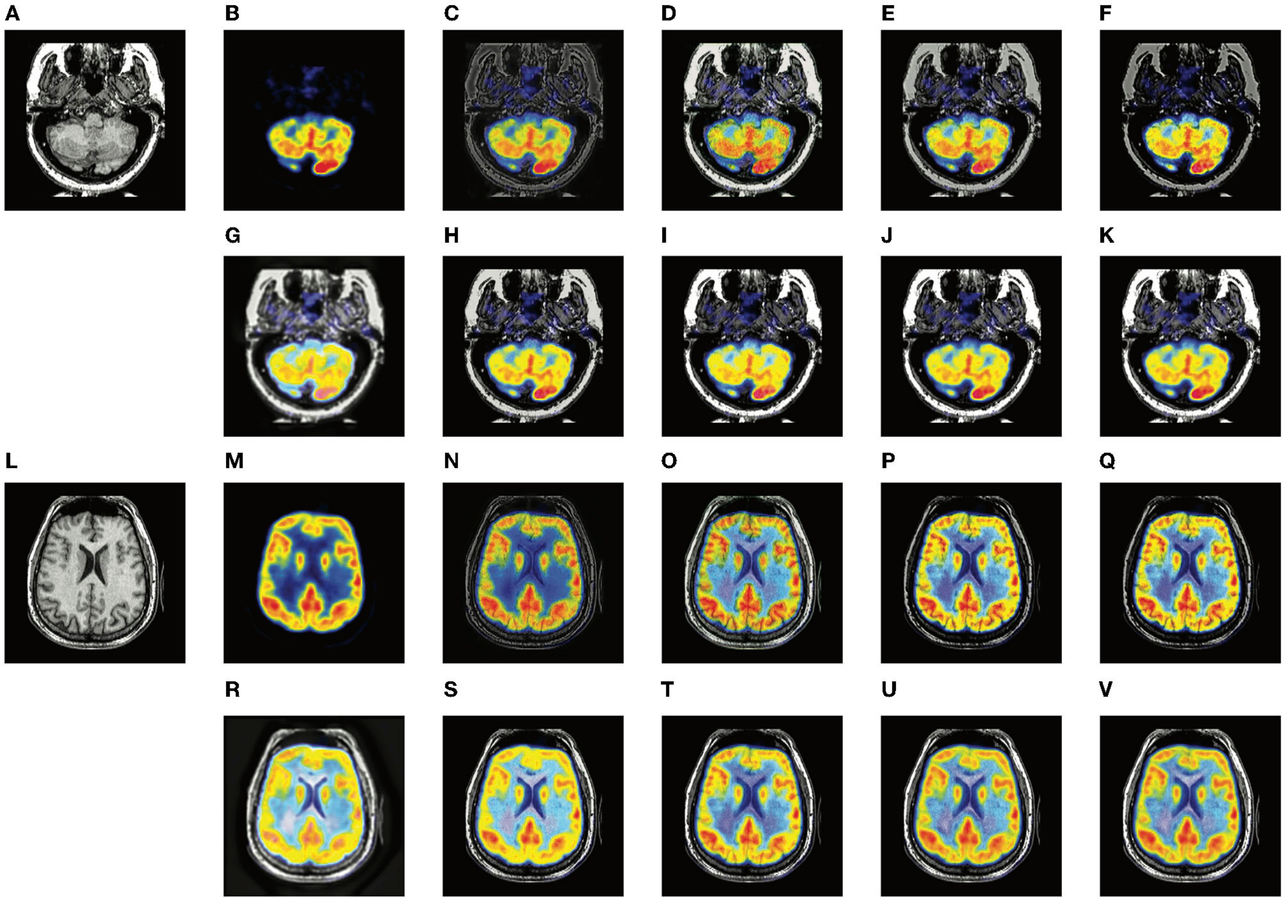


FIGURE 5 | Experiments results of proposed method with six deep learning-based methods on two typical MRI and PET image pairs. **(A,L)** MRI images; **(B,M)** PET images; **(C,N)** fused results of DDcGAN; **(D,O)** fused results of EMFusion; **(E,P)** fused results of IFSR; **(F,Q)** fused results of U2Fusion; **(G,R)** fused results of DRF; **(H,S)** fused results of SEDR; **(I,T)** fused results of proposed-add; **(J,U)** fused results of proposed-max; **(K,V)** fused results of proposed-l1.

TABLE 4 | The quantitative evaluation of different comparison methods on the MRI-CT dataset.

Methods	**Objective metrics**										
	SD	***SF***	Q_{MI}	Q_{NCIE}	Q_G	Q_M	Q_S	***MSSSIM***	***SCD***	Q_{CB}	***VIFF***
DDcGAN	88.13 ± 0.89	32.40 ± 0.57	0.58 ± 0.01	0.80 ± 0.00	0.57 ± 0.01	0.17 ± 0.00	0.25 ± 0.01	0.71 ± 0.00	1.24 ± 0.02	0.23 ± 0.01	0.25 ± 0.00
EMFusion	80.36 ± 0.59	20.76 ± 0.30	0.81 ± 0.01	0.81 ± 0.00	0.72 ± 0.01	0.16 ± 0.00	0.81 ± 0.00	0.89 ± 0.00	1.20 ± 0.05	0.67 ± 0.02	0.42 ± 0.01
IFSR	68.91 ± 0.46	19.81 ± 0.42	0.67 ± 0.01	0.81 ± 0.01	0.54 ± 0.01	0.11 ± 0.01	0.62 ± 0.00	0.89 ± 0.01	1.01 ± 0.03	0.34 ± 0.00	0.40 ± 0.00
U2Fusion	58.77 ± 0.38	21.06 ± 0.29	0.68 ± 0.01	0.81 ± 0.00	0.67 ± 0.01	0.13 ± 0.00	0.34 ± 0.00	0.89 ± 0.01	0.76 ± 0.03	0.28 ± 0.00	0.35 ± 0.01
DRF	75.42 ± 0.70	9.58 ± 0.14	0.51 ± 0.01	0.80 ± 0.00	0.22 ± 0.01	0.11 ± 0.00	0.22 ± 0.00	0.74 ± 0.00	1.19 ± 0.08	0.18 ± 0.00	0.31 ± 0.01
SEDR	63.78 ± 0.68	20.75 ± 0.35	0.81 ± 0.01	0.81 ± 0.00	0.56 ± 0.02	0.14 ± 0.00	0.32 ± 0.01	0.87 ± 0.00	0.85 ± 0.05	0.24 ± 0.00	0.36 ± 0.01
Proposed-add	84.92 ± 2.64	27.23 ± 0.27	0.80 ± 0.01	0.81 ± 0.00	0.72 ± 0.01	0.16 ± 0.02	0.82 ± 0.01	0.91 ± 0.01	1.37 ± 0.10	0.66 ± 0.01	0.45 ± 0.03
Proposed-max	80.48 ± 2.62	30.23 ± 1.05	0.80 ± 0.01	0.81 ± 0.00	0.74 ± 0.01	0.21 ± 0.03	0.82 ± 0.01	0.88 ± 0.01	1.16 ± 0.07	0.68 ± 0.01	0.41 ± 0.04
Proposed-l1	81.59 ± 2.49	29.28 ± 1.43	0.81 ± 0.02	0.81 ± 0.00	0.75 ± 0.01	0.22 ± 0.01	0.82 ± 0.01	0.88 ± 0.01	1.20 ± 0.10	0.68 ± 0.00	0.41 ± 0.04

The evaluation is shown as average ± variance of testing results on 10-group image pairs, which the proposed method adopts for 10-fold verification experiments.

TABLE 5 | The quantitative evaluation of different comparison methods on the MRI-PET dataset.

Methods	Objective metrics										
	SD	***SF***	Q_{MI}	Q_{NCIE}	Q_G	Q_M	Q_S	***MSSSIM***	***SCD***	Q_{CB}	***VIFF***
DDcGAN	57.93 ± 0.23	22.93 ± 0.14	0.53 ± 0.00	0.81 ± 0.00	0.53 ± 0.01	0.16 ± 0.01	0.57 ± 0.00	0.80 ± 0.00	0.67 ± 0.00	0.34 ± 0.00	0.35 ± 0.00
EMFusion	75.97 ± 0.11	32.03 ± 0.06	0.68 ± 0.00	0.81 ± 0.00	0.77 ± 0.00	0.42 ± 0.02	0.91 ± 0.00	0.91 ± 0.00	1.02 ± 0.01	0.62 ± 0.00	0.46 ± 0.00
IFSR	69.21 ± 0.15	25.29 ± 0.09	0.60 ± 0.00	0.81 ± 0.00	0.63 ± 0.00	0.15 ± 0.01	0.80 ± 0.00	0.92 ± 0.00	1.17 ± 0.02	0.55 ± 0.00	0.51 ± 0.00
U2Fusion	72.90 ± 0.07	26.05 ± 0.05	0.64 ± 0.00	0.80 ± 0.00	0.63 ± 0.00	0.17 ± 0.00	0.70 ± 0.01	0.89 ± 0.00	1.29 ± 0.00	0.60 ± 0.01	0.51 ± 0.00
DRF	71.40 ± 0.65	12.76 ± 0.09	0.46 ± 0.00	0.80 ± 0.01	0.34 ± 0.00	0.10 ± 0.00	0.46 ± 0.00	0.74 ± 0.00	0.82 ± 0.02	0.38 ± 0.00	0.36 ± 0.00
SEDR	84.47 ± 0.10	31.01 ± 0.06	0.76 ± 0.00	0.81 ± 0.00	0.73 ± 0.00	0.34 ± 0.01	0.84 ± 0.00	0.92 ± 0.00	1.52 ± 0.02	0.57 ± 0.00	0.55 ± 0.00
Proposed-add	89.28 ± 1.21	34.41 ± 0.27	0.76 ± 0.01	0.81 ± 0.00	0.77 ± 0.00	0.51 ± 0.08	0.80 ± 0.01	0.94 ± 0.00	1.65 ± 0.03	0.50 ± 0.01	0.59 ± 0.00
Proposed-max	86.84 ± 1.16	35.24 ± 0.49	0.72 ± 0.12	0.81 ± 0.00	0.70 ± 0.11	0.39 ± 0.20	0.77 ± 0.13	0.87 ± 0.08	1.31 ± 0.21	0.54 ± 0.02	0.49 ± 0.07
Proposed-l1	88.72 ± 2.17	36.05 ± 1.43	0.72 ± 0.15	0.81 ± 0.00	0.70 ± 0.14	0.38 ± 0.09	0.74 ± 0.06	0.86 ± 0.07	1.36 ± 0.17	0.52 ± 0.04	0.49 ± 0.06

The evaluation is shown as average ± variance of testing results on 10-group image pairs, which the proposed method adopts for 10-fold verification experiments.

TABLE 6 | Time cost comparison.

Methods	DDcGAN	EMFusion	IFSR	U2Fusion	DRF	SEDR	Proposed
Image size	256 × 256	256 × 256	256 × 256	256 × 256	256 × 256	256 × 256	256 × 256
Time cost	0.589s	0.448s	2.499s	0.086s	1.176s	0.900s	0.037s

and Q_{CB} in the MRI-PET task. Compared to other methods, EMFusion makes use of the MRI images to enhance the details of chrominance channels in PET images, instead of only fusing the luminance channel separately from the chrominance channels. Thus, EMFusion is capable of presenting high-quality color information with clear gradients. The proposed method which adopts different fusion strategies is able to achieve the best results on Q_M, *MSSSIM*, *SCD*, and *VIFF* in both tasks, and it also shows second-best performance on the most of the rest metrics. By comparing the three different fusion strategies on complementary features quantitatively and qualitatively, the addition strategy is good at showing more texture details as it directly combines all the information together. Thereby, it can also maintain the integral structure in both MRI-CT and MRI-PET tasks. The L1-selection strategy shows better performance in MRI-CT as it can adaptively assign the fusion weights. Max-selection can preserve the position with strong pixel intensity, however, it cannot avoid the loss of information to some degree. From the variance, the proposed method shows larger fluctuation than other methods on SD and SF. We assume this is because the content of training images in different folds of dataset can affect the generalization ability of a neural network to some degree. Besides, SD and SF evaluate the image quality by measuring the statistical features of the fused image, without considering the source image. To make a comprehensive assessment, the two metrics should be combined with the rest metrics which reflect the transfer ability of the fusion methods. In general, the proposed method presents a good ability in transferring edge details and preserving structural information, able to provide images with good visual quality. Such advantage is attributed to the disentanglement of redundant and complementary features, which makes the fusion process more accurate.

4.6. Time Cost Comparison

The running efficiency of a method is an important index to measure the performance as well. The average running time of different methods on all the test MRI-CT and MRI-PET image pairs is presented in **Table 6**. All methods are conducted on the same platform with Intel Core i7-6850K CPU and GeForce GTX 1080Ti GPU. From the time cost comparison, the proposed method is the most efficient than other comparison methods.

5. CONCLUSION

In this article, a disentangled representation based brain image fusion method is proposed. A three-branch auto-encoder architecture is designed to fully explore the significant features and correlations benefit of image fusion tasks, dealing with the unique modality characteristics. Based on the prior knowledge of complementary and redundant relationships, a complementary group lasso penalty is proposed for effective disentangled representation learning, which is able to separate the discriminative modality information from the structure information. The disentangled representations show better interpretability to allow simple fusion strategies and improve the precision of fusion results. The experiments on MRI-CT and MRI-PET fusion tasks demonstrate the effectiveness of the proposed method in retaining structure and details, as well as presenting good visual quality.

Nevertheless, the proposed method only focuses on the fusion of gray-scale images, and the chrominance channels of PET images are kept and directly combined with the fused gray-scale images, which leads to the degradation of texture information. In the future, how to embed the chrominance channels into a disentangled framework should be considered.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: http://www.med.harvard.edu/aanlib/. Code and pre-trained models are available at https://github.com/qqchong/A-Disentangled-Representationbased-Brain-Image-Fusion-via-Group-Lasso-Penalty.

AUTHOR CONTRIBUTIONS

AW and ZZ conceived the study. AW and XL designed the specific method. ZZ, XL, and X-JW analyzed the experiment data. AW wrote the draft. All authors gave critical revision and consent for this submission.

FUNDING

This study was supported in part by the National Natural Science Foundation of China under Grant No. 61772237, and the Six Talent Peaks Project in Jiangsu Province under Grant XYDXX-030.

REFERENCES

Aslantas, V., and Bendes, E. (2015). A new image quality metric for image fusion: the sum of the correlations of differences. *AEU-Int. J. Electron. Commun*. 69, 1890–1896. doi: 10.1016/j.aeue.2015.09.004

Ben, H. A., Yun, H., Hamid, K., and Alan, W. (2005). A multiscale approach to pixel-level image fusion. *Integrated Comput. Aided Eng*. 12, 135–146. doi: 10.3233/ICA-2005-12201

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell*. 35, 1798–1828. doi: 10.1109/TPAMI.2013.50

Chen, Y., and Blum, R. S. (2009). A new automated quality assessment algorithm for image fusion. *Image Vis. Comput*. 27, 1421–1432. doi: 10.1016/j.imavis.2007.12.002

Guo, K., Hu, X., and Li, X. (2022). MMFGAN: a novel multimodal brain medical image fusion based on the improvement of generative adversarial network. *Multimed Tools Appl*. 81, 5889–5927. doi: 10.1007/s11042-021-11822-y

Han, Y., Cai, Y., Cao, Y., and Xu, X. (2013). A new image fusion performance metric based on visual information fidelity. *Inform. Fusion* 14, 127–135. doi: 10.1016/j.inffus.2011.08.002

Hossny, M., Nahavandi, S., and Creighton, D. (2008). Comments on 'information measure for performance of image fusion'. *Electron. Lett*. 44, 1066–1067. doi: 10.1049/el:20081754

Huang, J., Le, Z., Ma, Y., Fan, F., Zhang, H., and Yang, L. (2020). MGMDcGAN: Medical image fusion using multi-generator multi-discriminator conditional generative adversarial network. *IEEE Access* 8, 55145–55157. doi: 10.1109/ACCESS.2020.2982016

Jian, L., Yang, X., Liu, Z., Jeon, G., Gao, M., and Chisholm, D. (2021). Sedrfuse: a symmetric encoder-decoder with residual block network for infrared and visible image fusion. *IEEE Trans. Instrum Meas*. 70, 1–15. doi: 10.1109/TIM.2020.3022438

Li, H., and Wu, X.-J. (2018). Densefuse: a fusion approach to infrared and visible images. *IEEE Trans. Image Process*. 28, 2614–2623. doi: 10.1109/TIP.2018.2887342

Li, H., Wu, X.-J., and Durrani, T. (2020). Nestfuse: an infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Trans. Instrum. Meas*. 69, 9645–9656. doi: 10.1109/TIM.2020.3005230

Li, H., Wu, X.-J., and Kittler, J. (2018). "Infrared and visible image fusion using a deep learning framework," in *2018 24th International Conference on Pattern Recognition (ICPR)* (Beijing: IEEE), 2705–2710.

Li, J., Wang, Y., Xiao, H., and Xu, C. (2019). Gene selection of rat hepatocyte proliferation using adaptive sparse group lasso with weighted gene co-expression network analysis. *Comput. Biol. Chem*. 80, 364–373. doi: 10.1016/j.compbiolchem.2019.04.010

Liu, Y., Chen, X., Peng, H., and Wang, Z. (2017). Multi-focus image fusion with a deep convolutional neural network. *Inform. Fusion* 36, 191–207. doi: 10.1007/978-3-319-42999-1

Luo, X., Zhang, Z., and Wu, X. (2016). A novel algorithm of remote sensing image fusion based on shift-invariant shearlet transform and regional selection. *AEU-Int. J. Electron. Commun*. 70, 186–197. doi: 10.1016/j.aeue.2015.11.004

Luo, X., Gao, Y., Wang, A., Zhang, Z., and Wu, X.-J. (2021). IFSepR: a general framework for image fusion based on separate representation learning. *IEEE Trans. Multimedia* 1–16. doi: 10.1109/TMM.2021.3129354

Ma, J., Ma, Y., and Li, C. (2019). Infrared and visible image fusion methods and applications: a survey. *Inform. Fusion* 45, 153–178. doi: 10.1016/j.inffus.2018.02.004

Ma, J., Xu, H., Jiang, J., Mei, X., and Zhang, X.-P. (2020). DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans. Image Process*. 29, 4980–4995. doi: 10.1109/TIP.2020.2977573

Ma, K., Zeng, K., and Wang, Z. (2015). Perceptual quality assessment for multi-exposure image fusion. *IEEE Trans. Image Process*. 24, 3345–3356. doi: 10.1109/TIP.2015.2442920

Piella, G., and Heijmans, H. (2003). "A new quality metric for image fusion," in *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429), Vol. 3* (Barcelona), III-173.

Qiang, W., Yi, S., and Jian, Q. Z. (2005). A nonlinear correlation measure for multivariable data set. *Physica D* 200, 287–295. doi: 10.1016/j.physd.2004.11.001

Simonyan, K., and Zisserman, A. (2015). "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, eds eds Y. Bengio and Y. LeCun (San Diego, CA).

Summers, D. (2003). Harvard whole brain atlas, www.med.harvard.edu/aanlib/home.html. *J. Neurol. Neurosurg. Psychiatry* 74, 288–288. doi: 10.1136/jnnp.74.3.288

Wang, J., Zhang, H., Wang, J., Pu, Y., and Pal, N. R. (2021). Feature selection using a neural network with group lasso regularization and controlled redundancy. *IEEE Trans. Neural Netw. Learn. Syst*. 32, 1110–1123. doi: 10.1109/TNNLS.2020.2980383

Wang, Q., and Guo, G. (2021). DSA-Face: diverse and sparse attentions for face recognition robust to pose variation and occlusion. *IEEE Trans. Inform. Forensics Security* 16, 4534–4543. doi: 10.1109/TIFS.2021.3109463

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process*. 13, 600–612. doi: 10.1109/TIP.2003.819861

Wang, P., and Liu, B. (2008). "A novel image fusion metric based on multi-scale analysis," in *2008 9th International Conference on Signal Processing* (Beijing), 965–968.

Xu, H., and Ma, J. (2021). EMFusion: an unsupervised enhanced medical image fusion network. *Inform. Fusion* 76:177–186. doi: 10.1016/j.inffus.2021.06.001

Xu, H., Ma, J., Jiang, J., Guo, X., and Ling, H. (2022). U2Fusion: a unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 502–518. doi: 10.1109/TPAMI.2020.3012548

Xu, H., Wang, X., and Ma, J. (2021). DRF: Disentangled representation for visible and infrared image fusion. *IEEE Trans. Instrum Meas.* 70, 1–13. doi: 10.1109/TIM.2021.3056645

Xydeas, C. S., and Pv, V. (2000). Objective image fusion performance measure. *Military Techn. Courier* 56, 181–193. doi: 10.1049/el:20000267

Yang, S., Min, W., Jiao, L., Wu, R., and Wang, Z. (2010). Image fusion based on a new contourlet packet. *Inform. Fusion* 11, 78–84. doi: 10.1016/j.inffus.2009.05.001

Yuan, M. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc.* 68, 49–67. doi: 10.1111/j.1467-9868.2005.00532.x

Zhao, L., Hu, Q., and Wang, W. (2015). Heterogeneous feature selection with multi-modal deep neural networks and sparse group lasso. *IEEE Trans. Multimedia* 17, 1–1. doi: 10.1109/TMM.2015.2477058

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.

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 09 August 2022
DOI 10.3389/fnins.2022.986153

OPEN ACCESS

EDITED BY
Jiayi Ma,
Wuhan University, China

REVIEWED BY
Hancheng Zhu,
China University of Mining and Technology, China
Pengfei Chen,
Xidian University, China

*CORRESPONDENCE
Chuangeng Tian
tianchuangeng@163.com

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 04 July 2022
ACCEPTED 13 July 2022
PUBLISHED 09 August 2022

CITATION
Tang L, Hui Y, Yang H, Zhao Y and Tian C (2022) Medical image fusion quality assessment based on conditional generative adversarial network.
Front. Neurosci. 16:986153.
doi: 10.3389/fnins.2022.986153

Medical image fusion quality assessment based on conditional generative adversarial network

Lu Tang[1], Yu Hui[1], Hang Yang[1], Yinghong Zhao[1] and Chuangeng Tian[2]*

[1]School of Medical Imaging, Xuzhou Medical University, Xuzhou, China, [2]School of Information and Electrical Engineering, Xuzhou University of Technology, Xuzhou, China

Multimodal medical image fusion (MMIF) has been proven to effectively improve the efficiency of disease diagnosis and treatment. However, few works have explored dedicated evaluation methods for MMIF. This paper proposes a novel quality assessment method for MMIF based on the conditional generative adversarial networks. First, with the mean opinion scores (MOS) as the guiding condition, the feature information of the two source images is extracted separately through the dual channel encoder-decoder. The features of different levels in the encoder-decoder are hierarchically input into the self-attention feature block, which is a fusion strategy for self-identifying favorable features. Then, the discriminator is used to improve the fusion objective of the generator. Finally, we calculate the structural similarity index between the *fake* image and the *true* image, and the MOS corresponding to the maximum result will be used as the final assessment result of the fused image quality. Based on the established MMIF database, the proposed method achieves the state-of-the-art performance among the comparison methods, with excellent agreement with subjective evaluations, indicating that the method is effective in the quality assessment of medical fusion images.

KEYWORDS

attention mechanism, conditional, generative adversarial networks, image quality assessment, medical image fusion

Introduction

As the population aging becomes familiar, and the vulnerability of the human brain to physical, chemical, and viral attacks, the incidence of brain diseases such as intracranial tumors, intracranial infectious diseases, and cerebrovascular diseases is gradually increasing, which has seriously threatened human health and wellbeing (Chen et al., 2022; Gottesman and Seshadri, 2022). There are many medical imaging modalities for clinical diagnosis and treatment of brain diseases, including computed tomography

(CT), magnetic resonance imaging (MRI), positron emission tomography (PET), and so on. Different imaging methods always have their unique advantages in attracting clinicians to choose (Liu et al., 2019; Cauley et al., 2021; Preethi and Aishwarya, 2021). For example, CT could superbly display the histological structure of the skull and the density changes in the brain parenchyma, while MRI could faithfully restore the essential features of the nervous or soft tissue. Generally, it is difficult for medical experts to identify the necessary information from a single modality of brain images to ensure the reliability of clinical diagnosis (Townsend, 2008). Additionally, some early work found that radiologists could effectively improve the diagnostic accuracy if they can analyze imaging results of more than two modalities at the same time (Li and Zhu, 2020). From a technical point of view, multimodal medical image fusion (MMIF) just meets this clinical need. Therefore, recently, MMIF has received attention and extensive exploration by researchers (Li et al., 2020; Ma et al., 2020; Liu et al., 2021).

The purpose of MMIF is to complement the image in different modalities to obtain better image expression, quality, and information perception experience (Azam et al., 2022; Liu et al., 2022). The fused images may contain both anatomical structure and tissue metabolism information (e.g., image fusion of CT and MRI), which improves the applicability of image-based diagnosis or assessment of diseases, thereby simplifying diagnosis. At present, many high-quality MMIF methods have been proposed (Arif and Wang, 2020; Wang K. P. et al., 2020; Duan et al., 2021; Ma et al., 2022; Xu et al., 2022). Madanala and Rani (2016) proposed a two-stage fusion framework based on the cascade of discrete wavelet transform (DWT) and non-subsampled contour transform (NSCT) domains, realizing the combination of spatial domain and transform domain. Inspired by the Tchebichef moments' ability to effectively capture edge features, Tang et al. (2017) used the Tchebichef moments energy to characterize the image shape, and thus designed an MMIF method based on the pulse coupled neural network (PCNN). However, the performance evaluation of these MMIF models and fused images has not been fully explored.

Normally, the higher the image quality, the more features and information human observers can receive or perceive through the image. As the ultimate observers and beneficiaries of the fused images, medical experts, although they subjectively evaluate the fused images as the most direct and reliable solution, it will be a very time-consuming and labor-intensive task, and it is not very useful in practical applications. Hence, objective image quality assessment (IQA) is very necessary (Liu et al., 2018; Shen et al., 2020; Wang W. C. et al., 2020). Some existing objective quality assessment studies include deblocking images, screen content images, multiple distorted images, and noisy images, etc. (Gao et al., 2008; Min et al., 2019; Liu et al., 2020; Meng et al., 2020). For instance, in early work, Wang et al. (2004) developed structural similarity (SSIM) index based on the subjective perception of image structure information, which achieved a breakthrough in the objective evaluation of image quality. Kang et al. (2014) used deep learning techniques to accurately predict the quality of images without reference images, and their method greatly improved the performance and robustness of the algorithm. On the premise of highlighting the important detection objects, Lei et al. (2022) fuses multiple features of the images at the pixel level and designed an IQA method of main target region extraction and multi-feature fusion. However, among these IQA methods, they are proposed for general use in the field of image fusion, not specifically for MMIF. Note that the quality assessment of medical fusion images includes information fidelity, contrast, grayscale tolerance, and region of interest (ROI). In clinical practice, the ROI usually refers to the lesion area. And, the ROI has a great influence on the results of IQA, which is the most different from the natural image (Du et al., 2016a; Cai et al., 2020; Chabert et al., 2021). As a result, there is an urgent need for a dedicated objective IQA method for medical fusion images.

We discussed with radiologists and found that the quality of a medical fusion image mainly depends on its impact on disease diagnosis. That is, the medical fusion image retains disease-relevant information in the ROI, it will be acceptable and will be given a higher subjective evaluation score. To this end, we propose a novel medical fusion image quality assessment method that uses the radiologist's mean opinion scores (MOS) as the constraint on conditional generative adversarial networks (GANs). Concretely, the method firstly extracts the feature of different depths from MOS and two input source images with the aid of dual-channel encoder-decoder. Next, under the supervision of the attention mechanism, we fuse the feature information hierarchically, and generate the fused image through the up-sampling algorithm. Then, the discriminator (***D***) differentiates the source of the fused images to improve the generator (***G***) performance. Finally, we calculate the SSIM of the *fake* image and *true* image, and the constrain value corresponding to the maximum value of SSIM as the evaluation result. The experimental results show that the proposed method is superior to the previous IQA algorithms, and the objective results obtained are more consistent with the subjective evaluation of radiologists.

The content of this paper is arranged as follows. In see section "Methodology," the proposed method is mainly introduced from four aspects: Encoder-Decoder, ***G***, ***D***, and objective function. The details of the experiments are presented in see section "Experiments." See section "Discussion and conclusion" contains the discussion and conclusion of this paper.

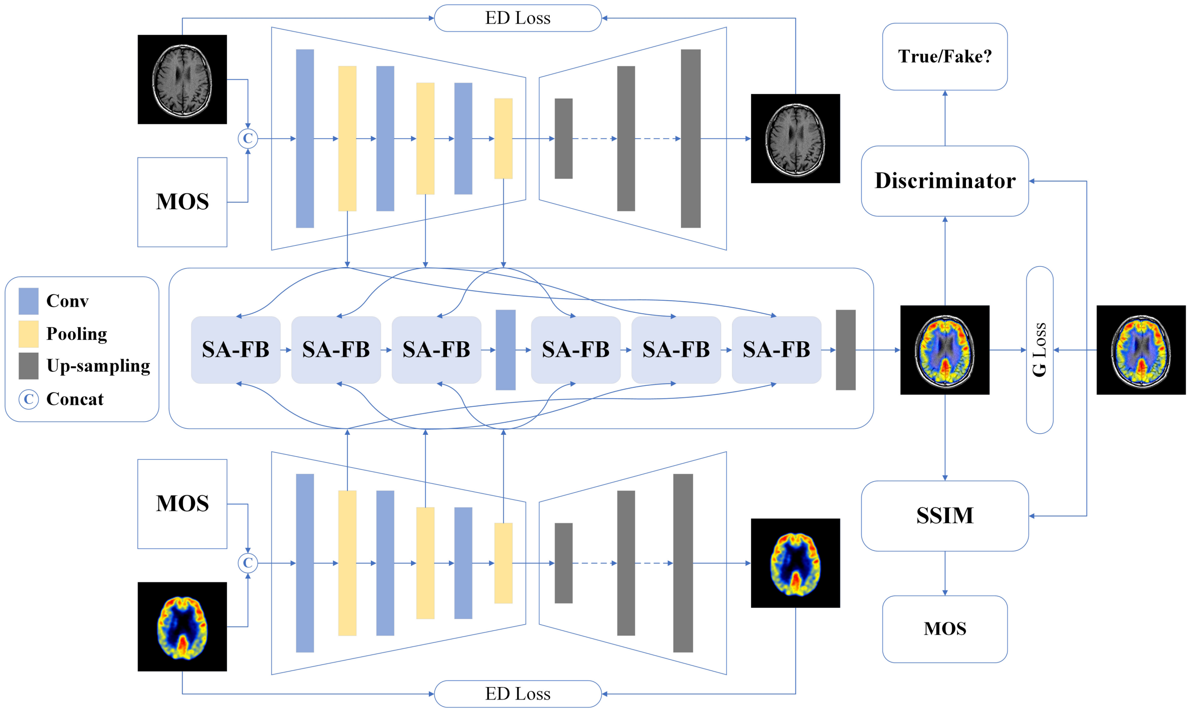


FIGURE 1
The overall architecture of our proposed method.

Methodology

The structure of our proposed model based on conditional generative adversarial network is shown in **Figure 1**, and the details are described below.

Dual-channel encoder-decoder

Among the existing multimodal medical images, each image has its unique imaging method and the advantage of displaying different human tissue. Therefore, accurately extracting the latent and deep key features of each modality image will be extremely conducive the image fusion (Ma et al., 2019). Besides, we also hope that MOS, the gold standard for image quality assessment, can participate in the feature extraction process of model learning images, in other words, learning the non-linear mapping relationship between MOS and fused images. To achieve this vision, we develop a dual-channel encoder-decoder structure.

First of all, we encapsulate three convolutional blocks, each of which contains two sets of convolutional layers, batch normalization (BN) layers, and activation layers. Specifically, the filter, stride, and padding of each convolutional layer are 3 × 3, 1, and 1, respectively. BN operation can effectively accelerate the network training as well as alleviate the problem of over-fitting. Thus, we append such operation after each convolutional layer. Considering that the image encoding process is important to learn image features and image fusion, we use a more comprehensive activation algorithm: Lleaky Rectified Line Unit (LeakyReLU). Then, we added max pooling operation instead of average pooling operation after each convolutional block. The reason is that the model should perform some specific feature selection under the constraints of MOS to learn more recognizable features. Each feature map output through the pooling operation is fed to the self-attention fusion block (SA-FB) separately, and more details will be explained in the next section. For the decoder, seven groups of deconvolution layer, BN layer, and Rectified Line Unit (ReLU) activation function layer complete the up-sampling operation of the feature maps. Finally, a reconstructed image of size 128 × 128 is obtained. It is worth noting that during the decoding operation, there is no feature map as output.

Perform the concatenating operation on the image of two different modalities (MI_i, $i = 1, 2$) and the corresponding MOS of their fused image, and the result is named MI_{imos}, and then input into two encoder-decoders, respectively. The feature map after the pooling layer is represented as F_{ij}, then the j-th feature map for the i-th modality can be marked as:

$$F_{ij} = ConvB(MI_{imos})_j \tag{1}$$

where $ConvB(\bullet)$ means the operation process of the j-th convolution block. The integer value range of j is one to three as only three convolution blocks are established in the encoding process. Here, sum of absolute difference is employed as the loss function for single modality image restoration, as defined by the following equation:

$$L_{ED} = \sum_{i} \sum \left| MI_i - \widehat{MI_i} \right|, i = 1, 2 \tag{2}$$

Where $\widehat{MI_i}$ refers to the original modal image restored by the decoder, and i represent the two modal images input to the dual-channel encoder-decoder, respectively.

Generator architecture

It is generally known that image fusion is the operation of synthesizing two or more images into one image, preserving the most representative features of each modality. To avoid the impact on image feature learning, independent of the dual-channel encoder-decoders, we design a feature fusion method based on the self-attention (SA) mechanism, as shown in **Figure 1**. Different levels of features contain different image information, for example, shallow features mean contour information while deep features represent texture information. For the three-level of feature F_{ij} yielded in the encoder, we develop the SA-FB to complete the fusion hierarchically. The structure diagram of SA-FB is shown in **Figure 2**.

In particular, the first SA-FB has only two inputs (i.e., F_{ij}), and the fusion feature F_{sa} is null. We do not carry out any feature selection operations (such as taking extreme values) during inputting, but directly feed the initial features F_{1j} and F_{2j} to SA after concatenating, and SA will sign weights to the features. Such setting can replace the manual feature selection algorithm, thus avoiding the loss of important information. SA is a variant of the attention mechanism from Sergey and Nikos (2017). It could coarsely estimate the foreground region to find prominent features that are in favor of later search. At the same time, it also reduces the dependence on external information, and is better at capturing the internal relevance of features. Immediately after, we adopt a convolution layer at the end of the SA. The convolution kernel size is set to 1 × 1 with stride 1 for adapt the output feature map weights. The output of this convolutional layer is concatenated with F_{sa}, and further input to a new convolution layer with a filter size of 3 × 3, and stride 1. In the end, a feature output F_{sa+1} that has undergone a complete SA-FB is obtained, and can be expressed as:

$$F_{sa+1} = safb(F_{ij}, F_{sa}), (i = 1, 2, j = 1, 2, 3) \tag{3}$$

where $safb(\bullet)$ is a series of operations of SA-FB. It should be mentioned that each convolution layer in the first three SA-FB is followed by BN layer and LeakyReLU as an activation function, which is similar to the encoder. The max pooling operation also appends after each SA-FB. The SA-FB in the up-sampling stage eliminates the pooling operation and changes the activation function to ReLU. On the basis of MOS as the condition to extract two modal image features, the ***G*** generates a fused image with 128 × 128. The parameters of the ***G*** are only renewed by the following loss function:

$$L_{fusion} = \frac{1}{N} \sum_{n=1}^{N} \left| y_{true} - \hat{y} \right|_1 \tag{4}$$

where y_{true} means the fused image with the corresponding MOS and the $\hat{y}$ represents the fused image produced by the ***G***. N is the total number of generations, and n represents the n-th generation. When training ***G***, minimize the following objective function:

$$L_G = V_G^{mos}(G, D) = E_{MI_1, MI_2 \sim P_{\mathrm{data}M}}$$
$$[\log(1 - D(MI_1, MI_2, (G(MI_1, MI_2 \,|mos))))] + \alpha L_{fusion} \tag{5}$$

where P_{dataM} represents the distribution of MI_1 and MI_2, respectively, and $E_{MI_1, MI_2 \sim P_{dataM}}$ represents the expectation of $G(MI_1, MI_2 \,|mos)$. α is a weight hyperparameter and is set to 100 during training.

To sum up, we restrict the generator based on MOS conditional information, and achieve the goal of generating image content. This is similar to that the generator analyzes the fused image by simulating the human visual system (HVS) and learns the non-linear mapping relationship between MOS and image. That is, the generator simulates a radiologist to assess the quality of the fused image, there by producing a fused image that matches the quality of MOS (i.e., ***G*** has learned the evaluation experience of radiologist).

To evaluate the quality of the fused image FI_{12}, first of all its original two modal images FI_1 and FI_2 should be input and then generate the fusion image FI_{fake} by ***G***. Where *1* and *2* represent two modal images, respectively. We have created five *fake* MOS ($MOS_k = 0.2k$, $k \in [1, 5]$, $k \in \mathbb{Z}$) as the conditional constraints ***G***, so the FI_{fake} can be renewed to $FI_{fake,k}$, which represents the fused image generated under the five constraints. Finally, the SSIM between FI_{12} and $FI_{fake,k}$ is calculated, and the MOS corresponding to the optimal value is taken as the assessment result, as follows:

$$Q = \max SSIM(FI_{12}, G(FI_1, FI_2 \,|MOS_k)) \tag{6}$$

Discriminator architecture

The discriminator needs to determine whether the generated image conforms to the real data distribution, so its structure is much simpler than the generator. In the proposed method, the input of the ***D*** is the generated fusion image or the original fusion image, all of which are 128 × 128 in size, and down-sampling is implemented using the discriminator

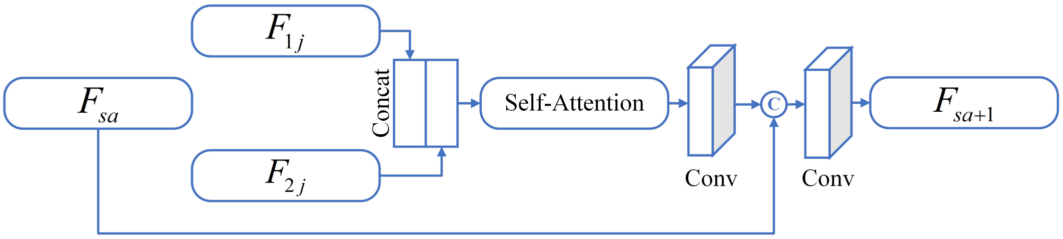


FIGURE 2
The diagrammatic sketch of SA-FB.

block (DB). Each DB consists of a convolution layer with a filter size of 3 × 3, stride of 2 and padding of 1, and followed by BN processing. The LeakyReLU is used as the activation function for each block. The image passes through four DB in sequence, and after each DB, the size of the feature map becomes a quarter of that before input. An independent convolutional layer with convolution kernel 3 × 3 and stride 1 is appended to the last DB, and the final obtained feature map is 6 × 6. At last, the discriminator will judge the authenticity of the result. We apply mean square error (MSE) as the loss function to optimizing the parameters of the ***D***. Further, the objective function of ***D*** can be reformulated as:

$$L_D = V_D^{mos}(G, D) = E_{y_{\mathrm{true}} \sim P_{data}}[\log D(y_{true} \\ |mos)]E_{MI_1, MI_2 \sim P_{dataM}}[\log(1 - D(G(MI_1, MI_2\,|mos)))] \quad (7)$$

where P_{data} represents the distribution of y_{true} and $E_{y_{true} \sim P_{data}}$ represents the expectation of y_{true} .

Total objective loss function

As shown in **Figure 1**, we use MOS as a condition to limit the content of the image generated by ***G***, and ***D*** determines whether the distribution of the generated fused image is true or false. ***G*** and ***D*** are trained against each other, and finally achieve the goal of Nash Equilibrium. Therefore, the optimization process of the whole network can be expressed by Eq. 8:

$$L_{all} = \min_G \max_D V(G, D) + \beta L_{ED} \quad (8)$$

where $V(G, D)$ can be obtained by Eqs. 5 and 7, respectively. β is a weight hyperparameter and is set to 20 in this experiment.

Experiments

Dataset

Image quality assessment has been developed in full swing in many fields and has made substantial progress. But, in the past period, the short-lived time of the MMIF algorithm has resulted in few research dedicated to the quality assessment of medical fusion images. In order to enable the medical image fusion algorithm to restore the brain structure more accurately and reflect tissue metabolic information more objectively, meeting the needs of clinical diagnosis, based on our previous work (Tang et al., 2020), we construct a special multimodal medical image fusion image database (MMIFID) with subjective evaluation of radiologists. Particularly, this work uses brain images from the AANLIB dataset, provided by Harvard Medical School and accessible online. The image size is 256 × 256, which can be browsed directly on the online web page. Most importantly, since image registration is completed for each combination of different modal images, it is one of the most widely used datasets. We selected 120 pairs of images in the AANLIB dataset and fused the images through ten image fusion algorithms. **Figure 3** shows examples of results generated by ten fusion algorithms. Consistent with our previous work (Tang et al., 2020), radiologists subjectively evaluated the quality of the fused image and gave a score (1 is the lowest and 5 is the highest), and finally obtained the MOS.

Evaluation metrics

To comprehensively evaluate the performance of the proposed method, that is, the consistency of the model's assessment of the fused image quality with the MOS score, we adopted four commonly used performance metrics: Spearman Rank-order Correlation Coefficient (SRCC), Kendall Rank-order Correlation Coefficient (KRCC), Pearson Linear Correlation Coefficient (PLCC), and Root Mean Square Error (RMSE). To sum up, the higher SRCC, KRCC and PLCC value and lower RMES value mean better model performance. Note, the model is evaluated at the end of each training epoch, and the final model is the checkpoint model with the best evaluation performance within 200 epochs.

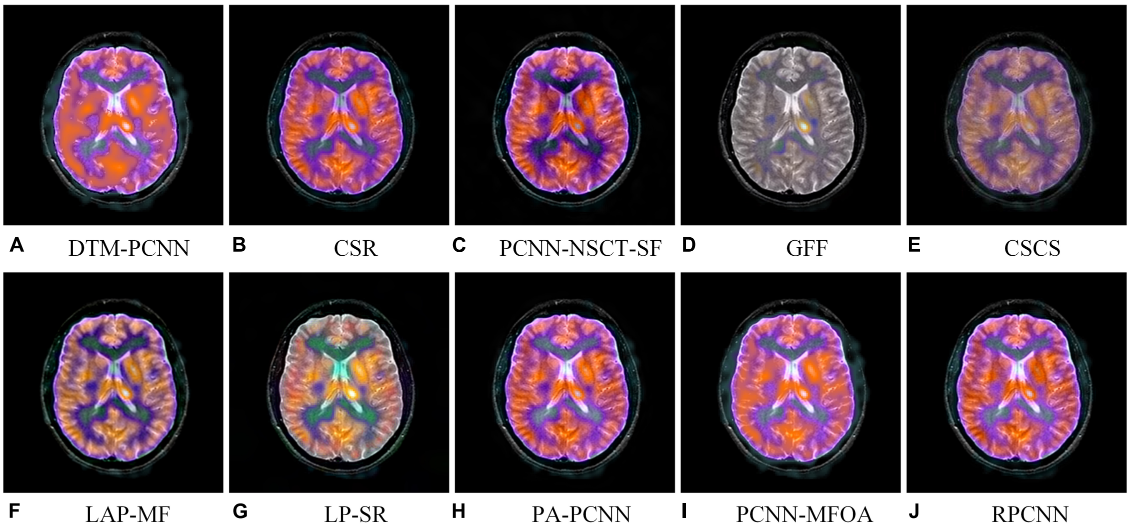


FIGURE 3
An example of fused images generated by ten different MMIF algorithms. Algorithms include **(A)** discrete Tchebichef moments and pulse coupled neural network (DTM-PCNN) (Min et al., 2019), **(B)** convolutional sparse representation (CSR) (Liu et al., 2016), **(C)** pulse-coupled neural network with modified spatial frequency based on non-subsampled contourlet transform (PCNN-NSCT-SF) (Das and Kundu, 2012), **(D)** guided filtering (GFF) (Li et al., 2013), **(E)** cross-scale coefficient selection (CSCS) (Shen et al., 2013), **(F)** union Laplacian pyramid with multiple features (LAP-MF) (Du et al., 2016b), **(G)** Laplacian pyramid and sparse representation (LP-SR) (Liu et al., 2015), **(H)** parameter-adaptive pulse-coupled neural network (PA-PCNN) (Yin et al., 2019), **(I)** pulse coupled neural network using the multi-swarm fruit fly optimization algorithm (PCNN-MFOA) (Tang et al., 2019), and **(J)** reduced pulse-coupled neural network (RPCNN) (Das and Kundu, 2013).

Comparison methods

The results are compared with those of the state-of-the-art (SOTA) image fusion quality metrics, which are listed as follows:

Mutual Information (Q_{MI}) (Hossny et al., 2008): As an objective method for evaluating image fusion performance, this method can measure the features and visual information from the input initial image and the fused image. The MI method we adopted is optimized by Hossny et al. (2008).

Non-linear Correlation Information Entropy (Q_{NCIE}) (Wang and Shen, 2004): Wang et al. propose a method based on non-linear correlation measures. This method evaluates the performance of image fusion algorithms by analyzing the general relationship between the source image and the fused image.

Gradient based fusion metric (Q_G) (Xydeas and Petrovic, 2000): This performance metric measures the amount of visual information transmitted from the source image to the fused image.

Ratio of spatial frequency error (Q_{rSFe}) (Zheng et al., 2007): This is a new metric based on extended spatial frequencies, and its original intention is to guide the algorithm to obtain a better fusion image.

The metric proposed by Yang et al. (2008) (Q_Y): According to the structural similarity between the source image and the fused image, this method treats redundant regions and complementary / conflicting regions, respectively.

A metrics based on edge preservation (Q_{EP}) (Wang and Liu, 2008): An image fusion metric method is proposed based on the perspective of edge information preservation.

A metric based on an absolute image feature measurement (Q_P) (Zhao et al., 2007): Based on phase congruency and its moments, a pixel-level image fusion performance metric is defined, which provides an absolute measure of image features.

Table 1 shows the performance of the above methods on our MMIFID, and the last row is the performance of the method proposed in this paper. Generally, SRCC, KRCC, and PLCC

TABLE 1 Comparison of quality assessment performance of different models.

Methods	SRCC	KRCC	PLCC	RMSE
Q_{MI}	0.2545	0.3604	0.2772	0.3804
Q_{NCIE}	0.2647	0.3608	0.2920	0.4093
Q_G	0.2488	0.3322	0.2444	0.2791
Q_{rSFe}	0.1801	0.2076	0.3126	0.2872
Q_Y	0.1884	0.2400	0.2503	0.4002
Q_{EP}	0.0960	0.1275	0.2235	0.2970
Q_P	0.1093	0.1216	0.0803	0.3007
Proposed	**0.8259**	**0.7426**	**0.8197**	**0.1709**

The bold values are the results of our proposed method, which achieves the best performance.

can measure the agreement between MOS and the objective scores, while RMSE can calculate its absolute error. Thus, the higher the SRCC, KRCC, and PLCC values, the better the quality evaluation metrics. The smaller the RMSE, the higher accuracy of the assessment. From Table 1, we can observe that the proposed method outperforms all SOAT methods. Furthermore, it can also be noticed that our proposed metrics are obviously better than these methods, which especially highlights that the quality assessment methods for medical images differ from natural images. Therefore, it is necessary to explore the special indicators for the quality evaluation of medical fusion images.

Ablation experiment

As we know, image fusion can be divided into two categories: early fusion and late fusion. The early fusion fuses the image directly together and then carries on the process of feature extraction and selection, while the late fusion allows the images to go through the process of feature extraction and selection, respectively, and then perform image feature fusion. Therefore, our two ablation experiments are to downgrade the proposed method to the early fusion and late fusion model, named Early-FM and Late-FM, respectively. Specifically, Early-FM first concatenates F_1, F_2 and MOS, and then completes feature learning through the single-channel encoder-decoder structure (e.g., we use the single-channel encoder-decoder to replace dual-channel encoder-decoder). The features output by the third convolutional block will be used to generate the fused image. Different from Early-FM, the Late-FM first concatenates the images of the two modalities and their respective MOS, and then inputs them to the dual-channel encoder-decoder, respectively, to complete feature learning. The third convolution block of the two channels outputs features, and the fused features are obtained by fusion operation. Finally, $\boldsymbol{G}$ generates the fused image. For the third ablation experiment, we eliminated the SA mechanism in SA-FB, and the rest of the structure is consistent with the proposed method, which is marked as proposed w/o SA. We train the Early-FM, Late-FM and the proposed w/o SA based on the same method applied in the proposed method and tabulate their test performances in Table 2.

Two main conclusions can be drawn from the experimental results. First, the performance results of both Early-FM and Late-FM are worse than those of the hierarchical fusion strategy we designed (i.e., the proposed method without or with SA). More concretely, the results comparison between Early-FM and proposed method are notably improved by 11.82% for SRCC, 12.18% for KRCC, and 14.18% for PLCC, while the RMSE decreased by 7.16%. For Late-FM, the proposed method also improves SRCC, KRCC, and PLCC by 9.71, 9.99, and 13.64%, respectively, while reducing RMSE by 7.08%. It is conceivable that the unnecessary noise in the early fusion will affect the quality of the fused image, and the late fusion may lose important details of the image. Thus, the obtained results are not pleasing. Second, the performance of the proposed method with SA as guidance is better than that without SA, which means that with the assistance of the SA mechanism, the process of model learning features is superior.

TABLE 2 Comparative results of ablation experiments.

Methods	SRCC	KRCC	PLCC	RMSE
Early-FM	0.7077	0.6208	0.6779	0.2425
Late-FM	0.7288	0.6427	0.6833	0.2417
Proposed w/o SA	0.7825	0.7113	0.7867	0.2020
Proposed w SA	**0.8259**	**0.7426**	**0.8197**	**0.1709**

The bold values are the results of our proposed method, which achieves the best performance.

Discussion and conclusion

Multimodal medical image fusion, as a way to express multimodal diagnostic information at the same time, has gradually gained attention in the field of medical imaging. However, the diagnostic information that a radiologist can perceive is *not only* related to the amount of initial image information contained in the fused image, *but also* to the quality of the fused image. Therefore, the quality assessment of MMIF plays an increasingly important role in the field of image processing and medical imaging diagnosis. At the same time, it has also aroused the interest of many scholars in the industry.

As MMIF is gradually gaining recognition in the medical field, quality assessment of fused images has also developed vigorously as an emerging field. An excellent objective assessment method can *not only* achieve the purpose of image quality control, *but also* guide the optimization of image fusion algorithms, so as to find the best algorithm for image fusion of different modalities. For instance, a certain algorithm can achieve very good results for image of MRI and CT, but it is not suitable for image fusion of MRI and SPECT, and maybe another algorithm should be more suitable. Unfortunately, most of existing IQA research methods are based on natural images, and it is difficult to achieve satisfactory performance for medical fusion images (see section "Comparison methods"). On the basis of previous work, we augmented the medical image database, MMIFID, which takes the doctor's MOS as the gold standard for subjective evaluation. The image content generated by $\boldsymbol{G}$ is constrained by MOS as a condition, and the non-linear mapping relationship between subjective evaluation and fused image is learned. The experimental results show that the objective evaluation results obtained from the model can match the subjective evaluation values well. In addition, compared with other IQA algorithms, we found that the proposed method outperforms the SOTA methods. Finally, we enumerate the potential limitations of this work as follows: (1) Although the database we built, as far as we know, is the largest multimodal medical image fusion database with MOS. However, it may still be a challenge for training GANs. In the future, we will continue to work on expanding the database. (2) Currently, the images contained in MMIFID are brain data, and we hope to add other body parts to the database in the future. (3) This work uses SSIM to calculate and obtain the final fusion image quality evaluation results, which may affect the accuracy of assessment to a certain extent. It would be better if the final evaluation result could also be directly assigned by GANs. Future, we will continue to explore the impact of fusing two modalities image through different methods, and design another novel IQA algorithm based on the idea of no reference.

Data availability statement

The original contributions presented in this study are included in the article; further inquiries can be directed to the corresponding author. The brain images are accessible online: https://www.med.harvard.edu/aanlib/home.html.

Author contributions

LT and HY wrote the main manuscript and contributed to the final version of the manuscript. CT and YH implemented the algorithm and conducted the experiments. YZ supervised the project and collected the data. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (82001912), Project Funded by China Postdoctoral Science Foundation (2018M642325), and QingLan Project of Jiangsu Province (2021).

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

Arif, M., and Wang, G. J. (2020). Fast curvelet transform through genetic algorithm for multimodal medical image fusion. *Soft Comput.* 24, 1815–1836. doi: 10.1007/s00500-019-04011-5

Azam, M. A., Khan, K. B., Salahuddin, S., Rehman, E., Ali Khan, S., Attique Khan, M., et al. (2022). A review on multimodal medical image fusion: compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Comput. Biol. Med.* 144:105253. doi: 10.1016/j.compbiomed.2022.105253

Cai, C. L., Chen, L., Zhang, X. Y., and Gao, Z. (2020). End-to-End optimized ROI image compression. *IEEE Trans. Image Process.* 29, 3442–3457. doi: 10.1109/TIP.2019.2960869

Cauley, K. A., Hu, Y., and Fielden, S. W. (2021). Head CT: toward making full use of the information the X-rays give. *Am. J. Neuroradiol.* 42, 1362–1369. doi: 10.3174/ajnr.a7153

Chabert, S., Castro, J. S., Munoz, L., Cox, P., Riveros, R., Vielma, J., et al. (2021). Image quality assessment to emulate experts' perception in lumbar MRI using machine learning. *Appl. Sci. Basel* 11:6616. doi: 10.3390/app11146616

Chen, S., Zhao, S., and Lan, Q. (2022). Residual block based nested U-type architecture for multi-modal brain tumor image segmentation. *Front. Neurosci.* 16:832824. doi: 10.3389/fnins.2022.832824

Das, S., and Kundu, M. K. (2012). NSCT-based multimodal medical image fusion using pulse-coupled neural network and modified spatial frequency. *Med. Biol. Eng. Comput.* 50, 1105–1114. doi: 10.1007/s11517-012-0943-3

Das, S., and Kundu, M. K. (2013). A neuro-fuzzy approach for medical image fusion. *IEEE Trans. Biomed. Eng.* 60, 3347–3353. doi: 10.1109/TBME.2013.2282461

Du, J., Li, W. S., Lu, K., and Xino, B. (2016a). An overview of multi-modal medical image fusion. *Neurocomputing* 215, 3–20. doi: 10.1016/j.neucom.2015.07.160

Du, J., Li, W., Xiao, B., and Nawaz, Q. (2016b). Union Laplacian pyramid with multiple features for medical image fusion. *Neurocomputing* 194, 326–339. doi: 10.1016/j.neucom.2016.02.047

Duan, J. W., Mao, S. Q., Jin, J. W., Zhou, Z., Chen, L., and Chen, C. L. P. (2021). A novel GA-based optimized approach for regional multimodal medical image fusion with Superixel segmentation. *IEEE Access.* 9, 96353–96366. doi: 10.1109/ACCESS.2021.3094972

Gao, X. B., Lu, W., Li, X. L., and Tao, G. (2008). Wavelet-based contourlet in quality evaluation of digital images. *Neurocomputing* 72, 378–385. doi: 10.1016/j.neucom.2007.12.031

Gottesman, R. F., and Seshadri, S. (2022). Risk factors, lifestyle behaviors, and vascular brain health. *Stroke* 53, 394–403. doi: 10.1161/strokeaha.121.032610

Hossny, M., Nahavandi, S., and Creighton, D. (2008). Comments on 'Information measure for performance of image fusion. *Electron. Lett.* 44, 1066–1067. doi: 10.1049/el:20081754

Kang, L., Ye, P., Li, Y., and Doermann, D. (2014). "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Columbus, OH), doi: 10.1109/CVPR.2014.224

Lei, F., Li, S., Xie, S., and Liu, J. (2022). Subjective and objective quality assessment of swimming pool images. *Front. Neurosci.* 15:766762. doi: 10.3389/fnins.2021.766762

Li, C. X., and Zhu, A. (2020). Application of image fusion in diagnosis and treatment of liver cancer. *Appl. Sci.* 10:1171. doi: 10.3390/app10031171

Li, S., Kang, X., and Hu, J. (2013). Image fusion with guided filtering. *IEEE Trans. Image Process.* 22, 2864–2875. doi: 10.1109/TIP.2013.2244222

Li, X. X., Guo, X. P., Han, P. F., Wang, X., Li, H., and Luo, T. (2020). Laplacian rede composition for multimodal medical image fusion. *IEEE Trans. Instr. Meas.* 69, 6880–6890. doi: 10.1109/TIM.2020.2975405

Liu, R. S., Liu, J. Y., Jiang, Z. Y., Fan, X., and Luo, Z. (2021). A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion. *IEEE Trans. Image Process.* 30, 1261–1274. doi: 10.1109/TIP.2020.3043125

Liu, Y., Chen, X., Wang, Z., Wang, Z. J., Ward, R. K., and Wang, X. (2018). Deep learning for pixel-level image fusion: recent advances and future prospects. *Inform. Fusion* 42, 158–173. doi: 10.1016/j.inffus.2017.10.007

Liu, Y., Chen, X., Ward, R. K., and Wang, Z. J. (2016). Image fusion with convolutional sparse representation. *IEEE Signal Process. Lett.* 23, 1882–1886. doi: 10.1109/LSP.2016.2618776

Liu, Y., Chen, X., Ward, R. K., and Wang, Z. J. (2019). Medical image fusion via convolutional sparsity based morphological component analysis. *IEEE Signal Process. Lett.* 26, 485–489. doi: 10.1109/LSP.2019.2895749

Liu, Y., Liu, S., and Wang, Z. (2015). A general framework for image fusion based on multi-scale transform and sparse representation. *Inform. Fusion* 24, 147–164. doi: 10.1016/j.inffus.2014.09.004

Liu, Y., Shi, Y., Mu, F., Cheng, J., Li, C., and Chen, X. (2022). Multimodal MRI volumetric data fusion with convolutional neural networks. *IEEE Trans. Inst. Meas.* 71, 1–15. doi: 10.1109/TIM.2022.3184360

Liu, Y., Wang, L., Cheng, J., Li, C., and Chen, X. (2020). Multi-focus image fusion: a Survey of the state of the art. *Inform. Fusion* 64, 71–91. doi: 10.1016/j.inffus.2020.06.013

Ma, J. Y., Xu, H., Jiang, J. J., Mei, X., and Zhan, X.-P. (2020). DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans. Image Process.* 29, 4980–4995. doi: 10.1109/TIP.2020.2977573

Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., and Ma, Y. (2022). Swinfusion: cross-domain long-range learning for general image fusion via Swin transformer. *IEEE CAA J. Autom. Sin.* 9, 1200–1217. doi: 10.1109/JAS.2022.105686

Ma, J., Yu, W., Liang, P., Li, C., and Jiang, J. (2019). FusionGAN: a generative adversarial network for infrared and visible image fusion. *Inform. Fusion* 48, 11–26. doi: 10.1016/j.inffus.2018.09.004

Madanala, S., and Rani, K. J. (2016). "PCA-DWT based medical image fusion using non sub-sampled contourlet transform," in *Proceedings of the 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, (Paralakhemundi), doi: 10.1109/SCOPES.2016.7955608

Meng, C. L., An, P., Huang, X. P., Yang, C., and Liu, D. (2020). Full reference light field image quality evaluation based on angular-spatial characteristic. *IEEE Signal Process. Lett.* 27, 525–529. doi: 10.1109/LSP.2020.2982060

Min, X. K., Zhai, G. T., Gu, K., Yang, X., and Guan, X. (2019). Objective quality evaluation of dehazed images. *IEEE Trans. Intell. Transp. Syst.* 20, 2879–2892. doi: 10.1109/TITS.2018.2868771

Preethi, S., and Aishwarya, P. (2021). An efficient wavelet-based image fusion for brain tumor detection and segmentation over PET and MRI image. *Multimedia Tools Appl.* 80, 14789–14806. doi: 10.1007/s11042-021-10538-3

Sergey, Z., and Nikos, K. (2017). "Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer," in *Proceeding of the International Conference on Learning Representations (ICLR)*, (Paris).

Shen, R., Cheng, I., and Basu, A. (2013). Cross-scale coefficient selection for volumetric medical image fusion. *IEEE Trans. Biomed. Eng.* 60, 1069–1079. doi: 10.1109/TBME.2012.2211017

Shen, Y. X., Sheng, B., Fang, R. G., Li, G., Dai, L., Stolte, S., et al. (2020). Domain-invariant interpretable fundus image quality assessment. *Med. Image Anal.* 61:101654. doi: 10.1016/j.media.2020.101654

Tang, L., Qian, J., Li, L., Hu, J., and Wu, X. (2017). Multimodal medical image fusion based on discrete Tchebichef moments and pulse coupled neural network. *Int. J. Imaging Syst. Technol.* 27, 57–65. doi: 10.1002/ima.22210

Tang, L., Tian, C., and Xu, K. (2019). Exploiting quality-guided adaptive optimization for fusing multimodal medical images. *IEEE Access* 7, 96048–96059. doi: 10.1109/ACCESS.2019.2926833

Tang, L., Tian, C., Li, L., Hu, B., Yu, W., and Xu, K. (2020). Perceptual quality assessment for multimodal medical image fusion. *Signal Process.* 85:115852. doi: 10.1016/j.image.2020.115852

Townsend, D. W. (2008). Dual-modality imaging: combining anatomy and function. *J. Nuclear Med.* 49, 938–955. doi: 10.2967/jnumed.108.051276

Wang, K. P., Zheng, M. Y., Wei, H. Y., Qi, G., and Li, Y. (2020). Multi-modality medical image fusion using convolutional neural network and contrast pyramid. *Sensors* 20:2169. doi: 10.3390/s20082169

Wang, P., and Liu, B. (2008). "A novel image fusion metric based on multi-scale analysis," in *Processing of the International Conference on Signal Processing (ICSP)*, (Beijing). doi: 10.1109/TIP.2017.2745202

Wang, Q., and Shen, Y. (2004). "Performances evaluation of image fusion techniques based on nonlinear correlation measurement," in *Proceedings of the IEEE Instrumentation and Measurement Technology Conference*, (Como), doi: 10.1109/IMTC.2004.1351091

Wang, W. C., Wu, X. J., Yuan, X. H., and Gao, Z. (2020). An experiment-based review of low-light image enhancement methods. *IEEE Access* 8, 87884–87917. doi: 10.1109/ACCESS.2020.2992749

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 600–612. doi: 10.1109/TIP.2003.819861

Xu, H., Ma, J., Jiang, J., Guo, X., and Ling, H. (2022). U2Fusion: a unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 502–518. doi: 10.1109/TPAMI.2020.3012548

Xydeas, C. S., and Petrovic, V. S. (2000). "Objective pixel-level image fusion performance measure," in *Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE) 4051, Sensor Fusion: Architectures, Algorithms, and Applications IV*, (Orlando, FL), doi: 10.1117/12.381668

Yang, C., Zhang, J. Q., Wang, X. R., and Liu, X. (2008). A novel similarity based quality metric for image fusion. *Inform. Fusion* 9, 156–160. doi: 10.1016/j.inffus.2006.09.001

Yin, M., Liu, X., Liu, Y., and Chen, X. (2019). Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled Shearlet transform domain. *IEEE Trans. Instr. Meas.* 68, 49–64. doi: 10.1109/TIM.2018.2838778

Zhao, J. Y., Laganiere, R., and Liu, Z. (2007). "Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement," in *Processing of the International Conference on Innovative Computing, Information and Control (ICICIC)*, (Ottawa, ON).

Zheng, Y., Essock, E. A., Hansen, B. C., and Haun, A. M. (2007). A new metric based on extended spatial frequency and its application to DWT based fusion algorithms. *Inform. Fusion* 8, 177–192. doi: 10.1016/j.inffus.2005.04.003

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 14 September 2022
DOI 10.3389/fnins.2022.1000587

OPEN ACCESS

EDITED BY
Xiangzhi Bai,
Beihang University, China

REVIEWED BY
Zhaoying Liu,
Beijing University of Technology, China
Junzhang Chen,
Beihang University, China

*CORRESPONDENCE
Xun Chen
xunchen@ustc.edu.cn

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 22 July 2022
ACCEPTED 18 August 2022
PUBLISHED 14 September 2022

CITATION
Liu Y, Mu F, Shi Y, Cheng J, Li C and Chen X (2022) Brain tumor segmentation in multimodal MRI via pixel-level and feature-level image fusion. *Front. Neurosci.* 16:1000587. doi: 10.3389/fnins.2022.1000587

Brain tumor segmentation in multimodal MRI *via* pixel-level and feature-level image fusion

Yu Liu[1,2], Fuhao Mu[1], Yu Shi[1], Juan Cheng[1,2], Chang Li[1,2] and Xun Chen[3]*

[1]Department of Biomedical Engineering, Hefei University of Technology, Hefei, China, [2]Anhui Province Key Laboratory of Measuring Theory and Precision Instrument, Hefei University of Technology, Hefei, China, [3]Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China

Brain tumor segmentation in multimodal MRI volumes is of great significance to disease diagnosis, treatment planning, survival prediction and other relevant tasks. However, most existing brain tumor segmentation methods fail to make sufficient use of multimodal information. The most common way is to simply stack the original multimodal images or their low-level features as the model input, and many methods treat each modality data with equal importance to a given segmentation target. In this paper, we introduce multimodal image fusion technique including both pixel-level fusion and feature-level fusion for brain tumor segmentation, aiming to achieve more sufficient and finer utilization of multimodal information. At the pixel level, we present a convolutional network named PIF-Net for 3D MR image fusion to enrich the input modalities of the segmentation model. The fused modalities can strengthen the association among different types of pathological information captured by multiple source modalities, leading to a modality enhancement effect. At the feature level, we design an attention-based modality selection feature fusion (MSFF) module for multimodal feature refinement to address the difference among multiple modalities for a given segmentation target. A two-stage brain tumor segmentation framework is accordingly proposed based on the above components and the popular V-Net model. Experiments are conducted on the BraTS 2019 and BraTS 2020 benchmarks. The results demonstrate that the proposed components on both pixel-level and feature-level fusion can effectively improve the segmentation accuracy of brain tumors.

KEYWORDS

brain tumor segmentation, medical image fusion, pixel-level fusion, feature-level fusion, convolutional neural networks

1. Introduction

Automatically and accurately segmenting brain tumor areas from multimodal magnetic resonance imaging (MRI) scans can provide crucial information about tumors including shape, volume, and localization. Based on these information, quantitative assessment of lesions can be carried out, which is of great significance to disease diagnosis, treatment planning, survival prediction, and other relevant tasks. Most existing brain tumor segmentation studies are concentrating on gliomas since they are the most common brain tumors in adults. However, due to the factors like the variety of tumor size, shape and position, the fuzzy boundaries, and the difference in intensity distribution of MRI data obtained by different devices, the accurate segmentation of brain tumors is always a very challenging task (Zhao et al., 2018).

Owing to the good ability in capturing high-resolution anatomic structure of tissues, MRI is mostly used in brain tumor segmentation. Commonly-used MRI modalities for brain tumor segmentation include T1-weighted (T1), contrast-enhanced T1-weighted (T1c), T2-weighted (T2), and fluid attenuated inversion recovery (Flair). Figure 1 gives an example of multimodal MRI volumes for brain tumor segmentation, which comes from the dataset released by the Brain Tumor Segmentation (BraTS) challenge (Menze et al., 2015), an annual event held by the Medical Image Computing and Computer Assisted Intervention (MICCAI). The segmentation label (i.e., ground truth) provided by physicians is also shown in Figure 1. The green, red, and yellow regions indicate edema (ED), necrosis and non-enhancing tumor (NCR/NET), and enhancing tumor (ET), respectively. In the BraTS challenge, the segmentation performance is evaluated on three partially overlapping sub-regions of tumors, namely, whole tumor (WT), tumor core (TC), and enhancing tumor (ET). The WT is the union of ED, NCR/NET, and ET, while the TC includes NCR/NET and ET. We can see from Figure 1 that different pathological features of tumors are captured by MRI data of different modalities.

In recent years, various brain tumor segmentation methods have been proposed. Traditional image segmentation methods based on threshold, region, and pixel clustering are difficult to achieve good results in this task due to its high complexity as mentioned above (Liu et al., 2014). The performance of machine learning approaches based on hand-crafted features and classifiers like support vector machines and random forests is still limited in most cases. In the last few years, deep learning-based methods have emerged as the trend in this field due to their obvious advantages on segmentation accuracy (Bakas et al., 2018). Some methods adopt a 2D or 3D patch-based manner, in which convolutional networks are applied to predict the class of the center voxel (Havaei et al., 2017; Kamnitsas et al., 2017; Zhao et al., 2018). However, these methods tend to ignore the correlation among different patches within a large receptive field. To better address the global contextual information, the encoder-decoder architectures represented by U-Net (Ronneberger et al., 2015) and V-Net (Milletari et al., 2016) have become more and more popular in brain tumor segmentation (Wang et al., 2017; Li et al., 2019a; Zhang et al., 2020a; Zhou et al., 2020).

As brain tumor segmentation in MRI is essentially a multimodal image segmentation problem, the joint utilization of multimodal information plays a critical role in this task (Zhang et al., 2022). However, we argue that most existing methods do not pay enough attention to this issue and the utilization of multimodal information is not sufficient. In existing brain tumor segmentation methods, the most common way of using multimodal MR images is to simply stack them or their low-level features as the model input (Cao et al., 2021; Chen et al., 2021; Valanarasu et al., 2021; Wang et al., 2021; Zhang et al., 2021b). In addition, as mentioned above, MR images with different modalities reflect different pathological features (Chen et al., 2021; Wang et al., 2021), so their importance to a given segmentation target should be different. However, many methods fail to take this difference into consideration in their segmentation models and there is a lack of refinement for multimodal features, which will have an adverse effect on the segmentation performance.

In this paper, we address the above problems *via* the multimodal image fusion technique at both the pixel level and the feature level. For one thing, we adopt pixel-level image fusion to enrich the input modalities of the segmentation model and the fused modalities can strengthen the association among different types of pathological information captured by multiple source modalities. For another, we embed an attention-based feature fusion module into the segmentation network to refine multimodal features for better segmentation performance.

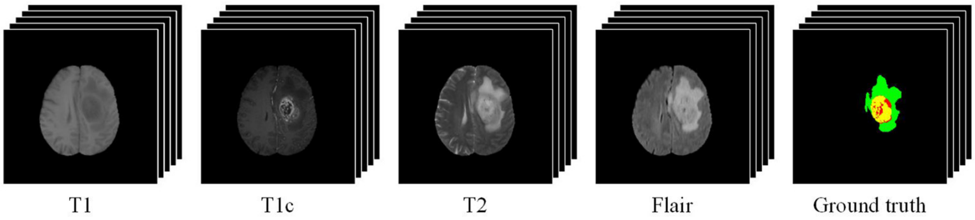


FIGURE 1
An example of multimodal MRI volumes for brain tumor segmentation. The green, red, and yellow regions in the ground truth indicate edema (ED), non-enhancing tumor and necrosis (NCR/NET), and enhancing tumor (ET), respectively.

Specifically, the main contributions of this work are summarized into four points:

1. To make use of multimodal information more sufficiently for brain tumor segmentation, we introduce the multimodal image fusion technique including both pixel-level fusion and feature-level fusion into the segmentation task.
2. We present a pixel-level image fusion network (PIF-Net) to fuse 3D multimodal MR images, aiming to enrich the input modalities of the segmentation model. This is actually a modality enhancement approach since the fused modalities obtained by the PIF-Net can effectively combine the pathological information from multiple source modalities.
3. To address the difference among multiple modalities for a given segmentation target, we design an attention-based modality selection feature fusion (MSFF) module for multimodal feature refinement and it is embedded into the segmentation network for performance improvement.
4. We propose a two-stage brain tumor segmentation framework based on the PIF-Net, the MSFF module and the V-Net. Experimental results on the BraTS 2019 and BraTS 2020 benchmarks demonstrate the effectiveness of the proposed pixel-level and feature-level fusion approaches for brain tumor segmentation.

The rest of this paper is organized as follows. Section 2 introduces the related works. In Section 3, the proposed method is presented in detail. The experimental results and discussion are given in Section 4. Finally, we conclude the paper in Section 5.

2. Related work

2.1. Brain tumor segmentation

Many automatic brain tumor segmentation methods have been proposed in recent years. They can be roughly divided into two categories (Havaei et al., 2017): the generative model-based methods and the discriminative model-based methods. The generative model-based methods require domain-specific prior knowledge about the appearance characteristics of tumorous and healthy tissues, but they are challenging to characterize due to the complexity of brain tissues. The discriminative model-based methods treat brain tumor segmentation as a pattern classification problem for the voxels in MRI volumes and they have become the mainstream in this field owing to the rapid development of machine learning techniques. Popular hand-crafted features used in brain tumor segmentation include local histograms (Goetz et al., 2014), structure tensor eigenvalues (Kleesiek et al., 2014), texture features (Subbanna et al., 2013), and so on, while typical shallow learning models such as support vector machines and random forests are frequently adopted in brain tumor segmentation (Bauer et al., 2011; Meier et al., 2014; Pinto et al., 2015).

In the last few years, deep learning has rapidly achieved the dominance in brain tumor segmentation owing to the significantly improved performance. Some early methods adopt a patch-based classification manner by utilizing convolutional networks to predict the class of the center voxel of a 2D or 3D image patch. Havaei et al. (2017) proposed a two-pathway architecture to extract features with 2D convolutional kernels of different sizes. They also explored three cascade architectures in which the output of the first network with larger input size is supplemented as an additional source for the second network to extract information of multiple scales simultaneously. The DeepMedic (Kamnitsas et al., 2017), a well-known 3D brain tumor segmentation model proposed by Kamnitsas et al., also adopts a dual pathway architecture that uses patches of different sizes as the network input, aiming to incorporate both local and larger contextual information. In addition, the dense training scheme is employed in Kamnitsas et al. (2017) to address the relationship among neighboring patches. Zhao et al. (2018) integrated fully convolutional neural networks (FCNNs) and the conditional random field (CRF) into a unified framework for brain tumor segmentation. In their method, features are also extracted from receptive fields of different sizes.

The above patch-based classification methods can't fully consider the correlation among neighboring patches and the range of the receptive field is always limited, although some improved strategies are adopted. To address this problem, the encoder-decoder semantic segmentation architectures such as U-Net (Ronneberger et al., 2015), 3D U-Net (Çiçek et al., 2016), and V-Net (Milletari et al., 2016) have become more and more popular in brain tumor segmentation. Myronenko (2018) proposed a segmentation method that won the first place in the BraTS 2018 challenge by adding an variational auto-encoder (VAE) branch into an encoder-decoder architecture to obtain an additional regularization to the encoder part. To alleviate the issue of class imbalance, some methods apply a cascaded architecture to decompose the original multi-label segmentation problem into multiple binary segmentation sub-problems. Wang et al. (2017) cascaded three CNNs to realize the segmentation of three tumor areas including WT, TC and ET. Zhang et al. (2020a) proposed a task-structured brain tumor segmentation network to address the task-modality and task-task relationship simultaneously. Zhou et al. (2020) proposed a one-pass multi-task network with cross-task guided attention for brain tumor segmentation, which integrates the multiple segmentation sub-tasks into one deep model. Li et al. (2019a) proposed a multi-step cascaded network that takes the hierarchical topology of the brain tumor sub-structures into account and segments the sub-structures from coarse to fine.

However, it is worth noting that current study on brain tumor segmentation does not pay enough attention to the joint utilization of multimodal MR images, which is in fact a key issue in this multimodal image segmentation task (Zhang et al., 2022). The most common way of using multimodal MR images is to simply stack them or their low-level features as the model input (Cao et al., 2021; Chen et al., 2021; Valanarasu et al., 2021; Wang et al., 2021; Zhang et al., 2021b). In addition, many methods treat each modality data with equal importance to a given segmentation target (Chen et al., 2021; Wang et al., 2021). These factors motivate us to introduce image fusion technique including both pixel-level fusion and feature-level fusion into the brain tumor segmentation framework for better performance.

2.2. Pixel-level medical image fusion

The purpose of pixel-level medical image fusion is to integrate the complementary information contained in multimodal medical images by generating a composite fused image, which is expected to be more suitable for human or machine perception. A variety of medical image fusion methods have been proposed over the past few decades and most of them are developed under a "decomposition-fusion-reconstruction" three-phase framework (Li et al., 2017; Liu et al., 2020b). Specifically, the source images are first decomposed into a transform domain and the decomposed coefficients from different source images are then fused. The fused image is finally reconstructed based on the fused coefficients. Multi-scale transform (MST) and sparse representation (SR) are two main categories of image decomposition that are widely used in medical image fusion (Liu et al., 2015, 2016, 2019, 2021; Du et al., 2016; Yang et al., 2016; Li et al., 2017; Zhang et al., 2018; Zhu et al., 2018; Yin et al., 2019).

However, most previous works in medical image fusion focus on the 2D image fusion problem, while methods for 3D image fusion were rarely studied (Yin, 2018). Using 2D fusion methods to tackle 3D medical images slice by slice independently neglects the correlation among adjacent slices and thereby tends to lose spatial contextual information of volumetric data. Wang et al. (2014) proposed a 3D multimodal medical image fusion method based on the 3D discrete shearlet transform (3D-DST) and designed a global-to-local strategy to fuse the decomposed coefficients. Yin (2018) introduced the tensor sparse representation (TSR), which is a high-dimensional extension of 2D SR, for 3D medical image fusion. Nevertheless, in these methods, the source images are treated equally in the fusion framework with identical decomposition approach and isotropic fusion strategy. As a result, the characteristics of different source modalities are not fully considered, leading to the loss of important modality information.

Recently, deep learning has emerged as an active direction in the field of image fusion (Liu et al., 2018; Zhang et al., 2021a) and some medical image fusion methods based on deep learning models like CNNs and generative adversarial networks (GANs) have been proposed (Liu et al., 2017, 2022; Liang et al., 2019; Ma et al., 2020a, 2022; Zhang et al., 2020b; Tang et al., 2021; Xu and Ma, 2021; Xu et al., 2022). By optimizing the loss functions that are specially designed based on the characteristics of source modalities, the deep learning-based methods have advantages over conventional MST-based and SR-based fusion methods on preserving modality information. However, the above deep learning-based methods are generally developed for 2D image fusion. In this work, we present a CNN-based 3D medical image fusion approach and introduce it for brain tumor segmentation by enriching the input modalities. In fact, current study on pixel-level medical image fusion is mostly devoted to pursuing good visual quality for physician observation and high evaluation results on objective metrics of image fusion, while very few study focuses on the application of image fusion to some specific clinical machine vision problems such as classification, detection and segmentation. Therefore, this work is also of high significance from the viewpoint of medical image fusion.

3. The proposed method

3.1. Overview

Figure 2 shows the schematic diagram of the proposed brain tumor segmentation framework. It consists of two stages to achieve the segmentation result of WT, TC, and ET areas. The two stages share a similar architecture that is composed of a PIF-Net to enrich the input modalities of the segmentation model *via* pixel-level fusion, an MSFF module to refine the mutlimodal features *via* feature-level fusion, and a V-Net (Milletari et al., 2016) with the encoder-decoder structure to obtain the segmentation result. The target of the first stage is to segment the WT area, while the second stage aims to identify the TC and ET areas. Since the TC and ET areas are included in the WT area, the segmentation result of the first stage is used to locate the input region of the second stage, which is helpful to alleviate the class imbalance issue. The sliding window-based approach introduced in Lyu and Shu (2020) is adopted to determine the input region of the second stage, namely, the window that contains the maximum number of tumor voxels is selected. In addition, considering that the peritumoral edema are mainly highlighted in T2 and Flair modalities, we only use T2 and Flair as the input source modalities in the first stage. The PIF-Net is used to generate their fused modality, which is denoted as T2-Flair. These three modalities (i.e., T2, Flair and T2-Flair) are fed together to the subsequent MSFF module in the first stage. In the second stage, all the four source modalities (i.e., T1, T1c, T2, and Flair) are adopted as the original input. The PIF-Net is applied to obtain two additional fused modalities, which are the fusion of T1c and T2 (denoted as T1c-T2), and the fusion of T1c and Flair (denoted as T1c-Flair). We mainly choose the T1c modality for

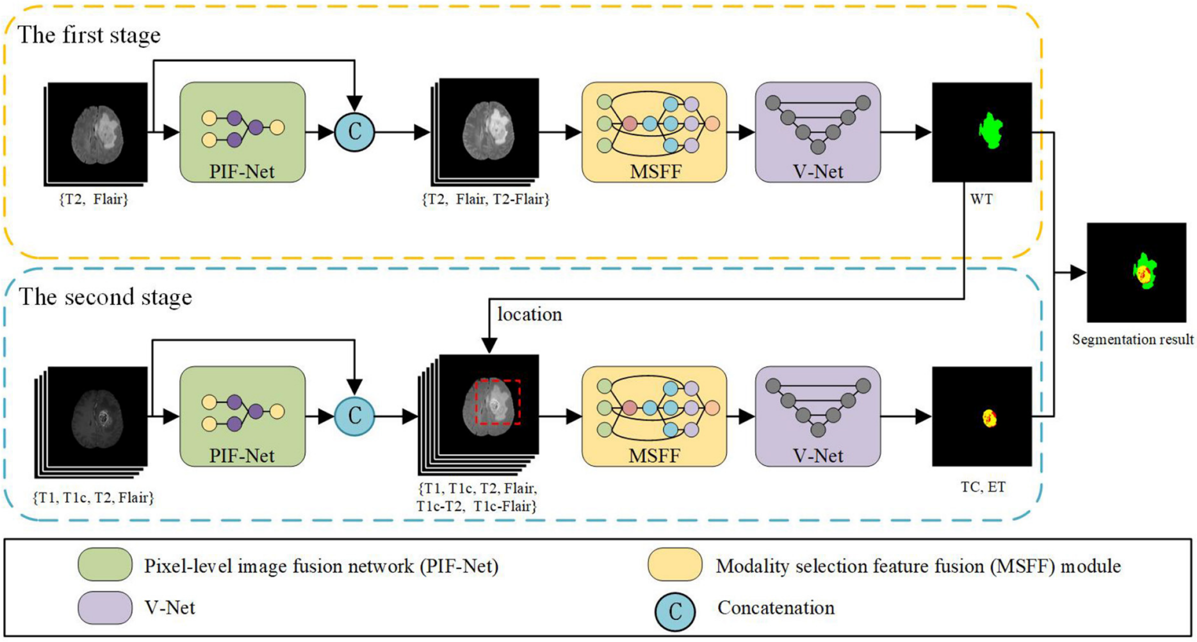


FIGURE 2
The schematic diagram of the proposed brain tumor segmentation framework.

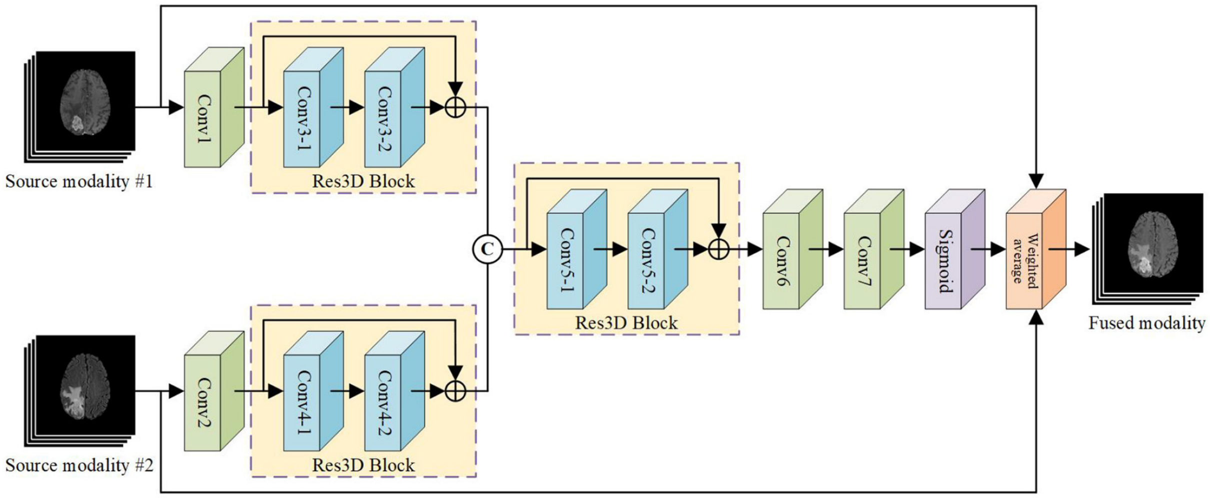


FIGURE 3
The architecture of our PIF-Net for 3D multimodal MR image fusion.

fusion because it is known to be very effective in detecting the TC and ET areas. By contrast, the T1 modality provides relatively less information for segmenting brain tumors and it generally plays an auxiliary role in this task (Bakas et al., 2018; Ma and Yang, 2018). Thus, the input of the MSFF module in the second stage contains six modalities in total. The final segmentation result is achieved by combining the results obtained at two stages together.

3.2. PIF-Net

Considering the high computational cost and memory usage of 3D convolutional networks, we design a relatively plain network architecture as shown in Figure 3 for 3D pixel-level image fusion. Note that this is likely to be the first work on CNN-based 3D medical image fusion to our knowledge, as mentioned in Section 2.2. The PIF-Net contains two branches for feature

TABLE 1 Detailed parameter configuration of the PIF-Net.

Layer	K_s	S_s	P_s	I_c	O_c	A
Conv1	3 × 3 × 3	1	1	1	32	ReLU
Conv2	3 × 3 × 3	1	1	1	32	ReLU
Conv3-1	3 × 3 × 3	1	1	32	32	ReLU
Conv3-2	3 × 3 × 3	1	1	32	32	/
Addition	/	/	/	32	32	ReLU
Conv4-1	3 × 3 × 3	1	1	32	32	ReLU
Conv4-2	3 × 3 × 3	1	1	32	32	/
Addition	/	/	/	32	32	ReLU
Conv5-1	3 × 3 × 3	1	1	64	64	ReLU
Conv5-2	3 × 3 × 3	1	1	64	64	/
Addition	/	/	/	64	64	ReLU
Conv6	3 × 3 × 3	1	1	64	32	/
Conv7	3 × 3 × 3	1	1	32	1	/
Sigmoid	/	/	/	1	1	/
Weighted average	/	/	/	1	1	/

K_s, S_s, P_s, I_c, O_c, and A denote the kernel size, stride, padding size, number of input channels, number of output channels, and activation operation, respectively.

extraction from two source modalities. Each branch is composed of a 3 × 3 × 3 convolutional layer and a 3D residual (denoted as Res3D) block that contains two 3 × 3 × 3 convolutional layers using the skip connection. The feature maps obtained from two branches are then concatenated and fed to another Res3D block. Two 3 × 3 × 3 convolutional layers are further applied to reduce the number of channels to 1 and a sigmoid operation is conducted to reconstruct a weight mask. Finally, the fused modality is reconstructed by performing the weighted average calculation based on the mask and source images. It is worth noting that the fused image can also be reconstructed directly from the fused feature maps without using a weight mask. However, since the voxels in the meaningless background regions have zero-valued intensity in each source modality, a direct regression tends to cause inappropriate non-zero predictions in these regions, which will affect the fusion quality. The voxel-wise weighted average strategy adopted can effectively avoid this problem and we experimentally found that it can produce good fusion results. The detailed parameter configuration of the network architecture is given in Table 1.

The definition of loss function is a key issue in deep learning-based image fusion methods as it determines the preservation of modality information from source images. In this work, the loss function of our PIF-Net is formulated as

$$L_{pif} = L_{pixel} + \alpha L_{ssim}, \tag{1}$$

where L_{pixel} and L_{ssim} indicate the pixel loss and the structural similarity loss, respectively. α is the regularization parameter that balances these two terms, and it is experimentally set to 450 in our method. The pixel loss is designed to preserve the intensity information, which is often related to the lesions(e.g., edema) that have very high or low intensity in some MRI modalities. It is defined as

$$L_{pixel} = ||\mathbf{F} - \mathbf{S}_1||_F^2 + \beta||\mathbf{F} - \mathbf{S}_2||_F^2, \tag{2}$$

where $\mathbf{S}_1$ and $\mathbf{S}_2$ denote the source images, and $\mathbf{F}$ denotes the fused image. β is the trade-off parameter and $||\cdot||_F^2$ denotes the tensor Frobenius norm. The structural similarity loss is adopted to extract anatomic structure information from source images and it is defined as

$$L_{ssim} = \gamma(1 - \mathrm{SSIM}(\mathbf{F}, \mathbf{S}_1)) + (1 - \mathrm{SSIM}(\mathbf{F}, \mathbf{S}_2)), \tag{3}$$

where $\mathrm{SSIM}(\cdot,\cdot)$ represents the 3D structural similarity measure and γ is the trade-off parameter.

The parameters β and γ are set according to the specific characteristics of fusion problems. In the first stage, for the fusion of T2 and Flair images, β and γ are both set to 1 since these two modalities have relatively similar pathological and structural information. In the second stage, let $\mathbf{S}_1$ and $\mathbf{S}_2$ denote the T1c and T2/Flair images, respectively. Considering that the T2/Flair image contains more lesion information regarding the edema area, we increase the weight of T2/Flair images in L_{pixel}. Meanwhile, since the T1c image captures more tissue structures in the TC and ET areas, a larger weight is assigned to the T1c image in L_{ssim}. In our method, we set both β and γ to 2 for the fusion of T1c and T2/Flair images.

The PIF-Net is trained based on the training set released by the BraTS challenge 2019. The training set contains 335 cases of multimodal MRI volumes and four modalities (i.e., T1, T1c, T2, and Flair) are provided in each case. The original volumes of size 155 × 240 × 240 are cropped into patches of size 80 × 80 × 80 by the sliding window technique to enlarge the scale of the training set. The learning rate is fixed as 10^{-4} during the training process and the Adam optimizer is adopted to train the network. Figure 4 shows an example of fusion results obtained by the PIF-Net. The results of two representative 3D medical image fusion methods 3D-DST (Wang et al., 2014) and TSR (Yin, 2018) are also provided for comparison. The results of T2 and Flair fusion and T1c and Flair fusion are given at the first and second rows in Figure 4, respectively. It can be seen that the PIF-Net achieves higher fusion quality than the other two methods on the tumor areas, especially for the T1c and Flair fusion, in which the 3D-DST and TSR methods fail in preserving the edema information contained in the Flair images well, while the PIF-Net simultaneously preserve important modality information from both two source images.

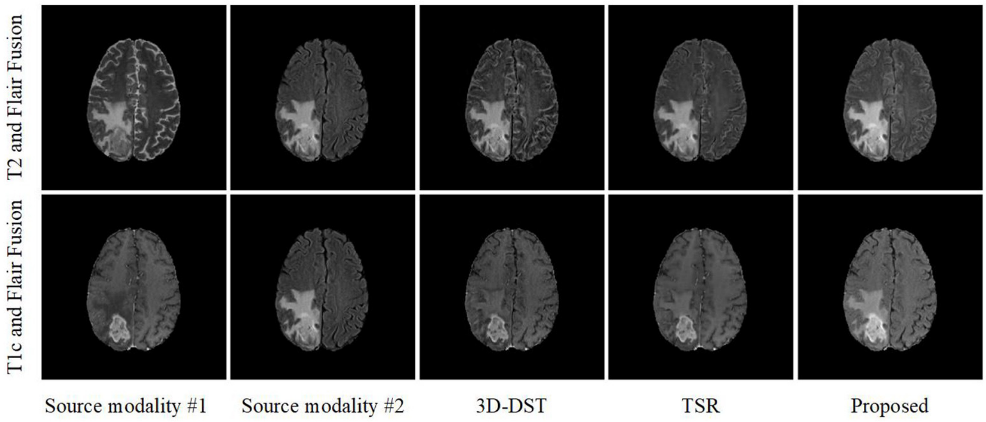


FIGURE 4
An example of fusion results obtained by different 3D medical image fusion methods.

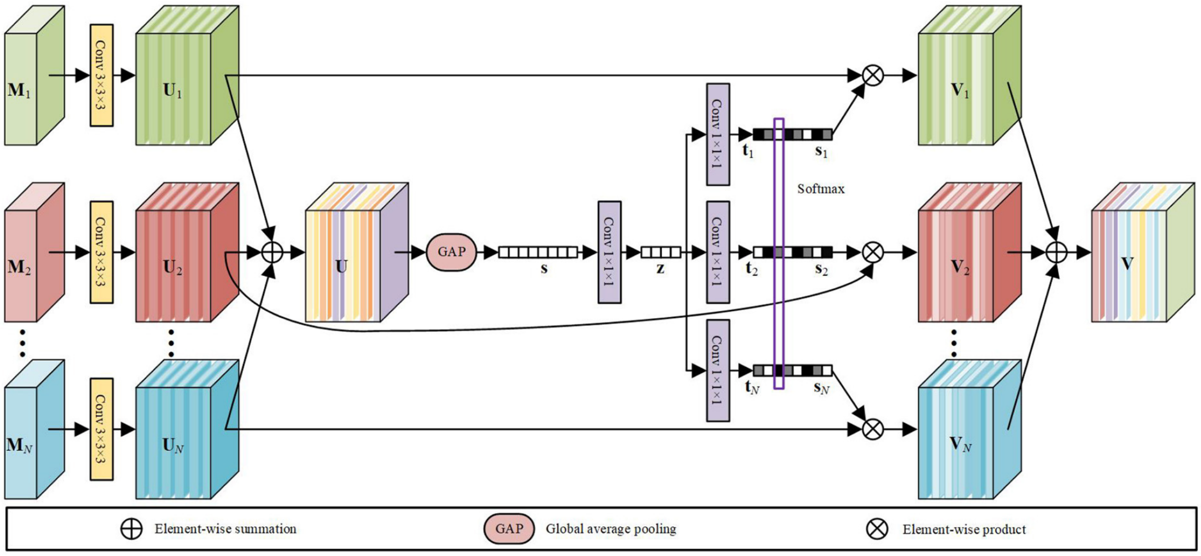


FIGURE 5
The architecture of our MSFF module for multimodal feature refinement.

3.3. MSFF module

The MSFF module is designed to refine the features extracted from multimodal MRI volumes for subsequent segmentation. Inspired the selective kernel network (SKNet) for multi-scale feature extraction (Li et al., 2019b), an attention-based feature fusion module is presented to adaptively adjust the weights of the features from different modalities. The architecture of our MSFF module is shown in Figure 5. Let $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_N \in \mathbb{R}^{L\times H\times W\times 1}$ denote the input multimodal MRI volumes that involve both the original source modalities and the fused modalities obtained by the PIF-Net, where N is total number of input modalities. A $3 \times 3 \times 3$ convolutional layer is firstly performed on each input volume for feature extraction. The obtained features are denoted as $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_N \in \mathbb{R}^{L\times H\times W\times C}$, where $L \times H \times W$ denotes the size of the 3D feature map and C denotes the number of feature maps. In our method, C is set to 16. The features from different sources are firstly merged *via* an element-wise summation as

$$\mathbf{U} = \sum_{i=1}^{N} \mathbf{U}_i. \tag{4}$$

Then, we embed the global information by a channel-wise global average pooling (GAP) operation to get a feature vector

$\mathbf{s} \in \mathbb{R}^{1\times1\times1\times C}$. Specifically, the c-th element of $\mathbf{s}$ is calculated as

$$s_c = \Phi_{GAP}(\mathbf{U}_c) = \frac{1}{L \times H \times W} \sum_{i=1}^{L} \sum_{j=1}^{H} \sum_{k=1}^{W} \mathbf{U}_c(i,j,k). \tag{5}$$

Further, a compact feature $\mathbf{z} \in \mathbb{R}^{1\times1\times1\times C/r}$ is generated by a $1 \times 1 \times 1$ convolutional layer for channel reduction, which is actually equivalent to a fully connected layer. The ratio factor r is set to 4 in our model. Next, we adopt N parallel channel up-scaling convolutions with kernel size of $1 \times 1 \times 1$ to reconstruct N C-dimensional vectors $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N \in \mathbb{R}^{1\times1\times1\times C}$. This is actually the excitation operation used in the SENet (Hu et al., 2018). Subsequently, a channel-wise softmax calculation is performed on each element across all the N vectors (indicated by the purple frame) to obtain the attention vectors $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N \in \mathbb{R}^{1\times1\times1\times C}$. Specifically, the c-th element of $\mathbf{s}_i$ is calculated as

$$s_{i,c} = \frac{e^{t_{i,c}}}{\sum_{j=1}^{N} e^{t_{j,c}}}, \tag{6}$$

where $t_{i,c}$ denotes the c-th element of $\mathbf{t}_i$, $i \in \{1, 2, \ldots, N\}$, $c \in \{1, 2, \ldots, C\}$.

Finally, the fused feature $\mathbf{V} \in \mathbb{R}^{L\times H\times W\times C}$ is calculated by a channel-wise weighted average over the source features using the attention weights as

$$\mathbf{V} = \sum_{i=1}^{N} \mathbf{s}_i \cdot \mathbf{U}_i. \tag{7}$$

According to a recent survey on attention mechanism (Guo et al., 2022), the attention mechanism used in our MSFF module belongs to the branch attention, which can be viewed as a dynamic branch selection mechanism and typically used in a multi-branch architecture. In the proposed method, to be more specific, the attention mechanism can be regarded as a kind of modality attention, aiming to extract features from multimodal MR images more effectively.

3.4. Segmentation loss

The loss function used for training the segmentation model is defined as

$$L_{seg} = L_{dice} + \lambda L_{bce}, \tag{8}$$

where L_{dice} and L_{bce} denote the dice loss and the binary cross entropy (BCE) loss, respectively, as

$$L_{dice} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2 + \varepsilon}, \tag{9}$$

$$L_{bce} = -\frac{1}{N} \sum_{i=1}^{N} [g_i \log p_i + (1 - g_i) \log(1 - p_i)], \tag{10}$$

where $g_i \in G$ is the ground truth binary volume, $p_i \in P$ is the network prediction, and N denotes the number of voxels. The parameter ε is a small constant to avoid dividing by 0. The Dice loss is known to be capable of alleviating the class imbalance issue (Milletari et al., 2016), while the BCE is the mostly used loss function for binary classification or segmentation. In brain tumor segmentation, the union of these two losses is a common way as it can combine their complementary advantages. The parameter λ controls the trade-off between these two losses and it is experimentally set to 0.5 in our method.

4. Experimental results and discussion

4.1. Data and implementation details

The BraTS 2019 and BraTS 2020 benchmarks (Menze et al., 2015) are adopted to demonstrate the effectiveness of the proposed method. The multimodal MRI data in a BraTS benchmark is divided into three parts: a training set, a validation set and a testing set. Only the training set releases the segmentation label (i.e., ground truth) annotated by experts to the public. The validation set is used to adjust model training and the MRI data is available, but the label is not provided. Users must upload their segmentation results to the organizer's sever at https://ipp.cbica.upenn.edu/ to obtain the evaluation results. Both data and label in the testing set are not available to users. In our experiments, just as most previous studies in this field, we adopt the training set for model training and validation, while use the validation set for performance evaluation. In particular, the BraTS training set is further divided into two parts: 80% samples are used for network training and the remaining 20% samples are used as a validation set to guide the training process. The BraTS 2019 training dataset includes 335 cases, while BraTS 2020 has a larger one comprising 369 cases. These multimodal MRI data have been skull-striped, re-sampled, and co-registered. Each case contains MRI data of four modalities (i.e., T1, T1c, T2, and Flair) and each volume is of size 155 × 240 × 240.

For data pre-processing and augmentation, the popular z-score normalization approach is applied to each MRI volume, namely, the data is subtracted by the mean and divided by the standard deviation of the non-zero region. The training volume is randomly cropped into patches of size 128 × 192 × 160 before fed to the network in the first stage. For each volume, the patch of size 128×128×128 that contains maximum tumor voxels is used for training in the second stage. Moreover, in both two stages, the intensity of each volume is randomly shifted by a value in $[-0.1\sigma, 0.1\sigma]$ (σ denotes the standard deviation) and randomly

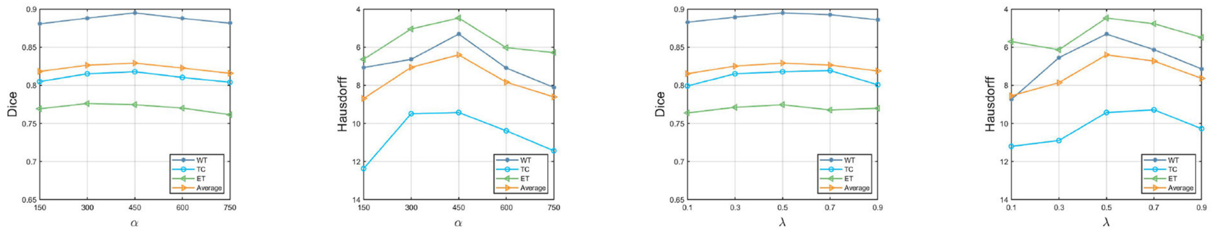


FIGURE 6
Impact of the parameters α and λ on the model performance.

scaled by a factor in [0.9, 1.1]. In addition, a random flipping along each axis is applied with a probability of 50%.

Our network is implemented in PyTorch and trained on two NVIDIA TITAN RTX GPUs. The Adam optimizer is used for updating weights. The learning rate is progressively decayed using the following rule:

$$l = l_0 \times (1 - \frac{i}{N})^{0.9}, \tag{11}$$

where l_0 is the initial learning rate, i is an epoch counter and N is the total number of the epochs. We experimentally set l_0 to 10^{-4} and N to 300.

The labels provided by the BraTS benchmark include the ED, NCR/NET and ET, while the evaluation of segmentation accuracy is performed on three partially overlapping regions: WT (ET + NCR/NET + ED), TC (ET + NCR/NET) and ET, as mentioned in Section 1. In our experiments, we adopt the region-based training strategy, which directly optimizes these three sub-regions instead of individual labels, since its effectiveness has been widely verified in brain tumor segmentation (Isensee et al., 2020). For post-processing, we also adopt a frequently-used approach that the ET is replaced by the NCR/NET when its volume is less than 500 voxels to remove possible false predictions on ET (Isensee et al., 2020; Lyu and Shu, 2020; Zhang et al., 2020a). Two popular objective metrics including the Dice score and the Hausdorff distance (%95) are used to evaluate the segmentation accuracy.

4.2. Parameter analysis

The loss functions in our method contain several trade-off parameters such as α, β, γ, and λ. The principle for determining the values of β and γ has been detailed in Section 3.2. In this subsection, we analyze the effect of the parameters α and λ on the segmentation performance of the proposed method. The parameter α is used to balance the pixel loss and the structural similarity loss, and these two terms should have relatively close values so that both of them can have sufficient contribution. Based on the experimental observations, we set α to 150, 300, 450, 600, and 750 to study its impact. The corresponding results are shown in the first two sub-figures in Figure 6. It can be seen that the proposed method can obtain relatively stable performance when α is set between 150 and 750, and in particular between 300 and 600. Based on these results, we set α to 450 by default in our experiments. The parameter λ controls the balance between the dice loss and the BCE loss in the segmentation model. Similarly, we set λ to 0.1, 0.3, 0.5, 0.7, 0.9 to analyze its effect on the model performance. The corresponding results are given in the last two sub-figures in Figure 6. We can see that the setting of 0.5 can result in the best performance in most cases, so the parameter λ is set to 0.5 by default in our method.

4.3. Ablation study of the proposed method

In this subsection, an ablation study is conducted to evaluate the effectiveness of our PIF-Net and MSFF module in the proposed method. Specifically, the following four models are considered in this study:

- **OURS w/o PIF-Net&MSFF**: Removing the PIF-Net and the MSFF module simultaneously from the proposed brain tumor segmentation framework. In each stage, only the V-Net is remained for segmentation. This is the original baseline for our method.
- **OURS w/o PIF-Net**: Removing the PIF-Net from the proposed brain tumor segmentation framework. The MSFF module is embedded before the V-Net to realize multimodal feature refinement for segmentation in both stages.
- **OURS w/o MSFF**: Removing the MSFF module from the proposed brain tumor segmentation framework. The PIF-Net is used to generate the fused modalities as the additional input of the segmentation model in both stages.
- **OURS**: The complete model proposed in this work.

The evaluation results on the BraTS 2019 and BraTS 2020 benchmarks are listed in Tables 2, 3, respectively. method

TABLE 2 Objective evaluation results for the ablation study of the proposed method on the BraTS 2019 validation sets.

Tumor region	Metrics	OURS w/o PIFnet & MSFF	OURS w/o PIFnet	OURS w/o MSFF	OURS
WT	Dice	0.8635	0.8771	0.8832	**0.8942**
	Hausdorff	7.1211	7.7784	7.1654	**5.3490**
TC	Dice	0.7788	0.8065	0.8045	**0.8142**
	Hausdorff	15.7345	**10.1822**	14.4599	10.8988
ET	Dice	0.7682	0.7698	0.7692	**0.7710**
	Hausdorff	9.1385	**5.3155**	6.4719	5.8548
Average	Dice	0.8035	0.8178	0.8190	**0.8265**
	Hausdorff	10.6647	7.7587	9.3657	**7.3675**

Bold values indicate the best-performing scores on each metric (each row in the tables) among all the four models.

TABLE 3 Objective evaluation results for the ablation study of the proposed method on the BraTS 2020 validation sets.

Tumor region	Metrics	OURS w/o PIFnet & MSFF	OURS w/o PIFnet	OURS w/o MSFF	OURS
WT	Dice	0.8678	0.8725	0.8878	**0.8950**
	Hausdorff	11.5732	9.6274	7.8896	**5.3117**
TC	Dice	0.8025	0.8153	0.8139	**0.8178**
	Hausdorff	11.6728	10.4340	10.9337	**9.4285**
ET	Dice	0.7631	0.7730	0.7678	**0.7745**
	Hausdorff	6.9469	5.9442	7.1674	**4.4715**
Average	Dice	0.8111	0.8203	0.8232	**0.8291**
	Hausdorff	10.0643	8.6685	8.6636	**6.4039**

Bold values indicate the best-performing scores on each metric (each row in the tables) among all the four models.

generally has a better a slightly better performance for BraTS 2020 than performance for BraTS 2020 than BraTS 2019, which is mainly because the BraTS 2020 benchmark contains more training samples in the training set, with additional 34 samples in comparison to the BraTS 2019 benchmark. The comparison between **OURS** and **OURS w/o PIFnet&MSFF** demonstrates that the utilization of our PIF-Net and MSFF module can significantly improve the performance (1.8% to 2.3% in terms of the mean Dice score, and 3.3 to 3.7 in terms of the mean Hausdorff distance) over the baseline model. The comparison between **OURS w/o MSFF** and **OURS w/o PIF-Net&MSFF** (as well as the comparison between **OURS** and **OURS w/o PIF-Net**) verifies the effectiveness of the PIF-Net in improving the segmentation accuracy. The comparison between **OURS w/o PIF-Net** and **OURS w/o PIF-Net&MSFF** (as well as the comparison between **OURS** and **OURS w/o MSFF**) shows that the MSFF module is also beneficial for the segmentation performance. Some segmentation results obtained by **OURS w/o PIF-Net&MSFF**, **OURS w/o PIF-Net**, **OURS w/o MSFF**, and **OURS** are visualized in Figure 7. It can be seen that the complete model can generally obtain more accurate segmentation results than the baseline methods when compared to the ground truth.

An interesting observation we can obtain from Tables 2, 3 are that the improvements achieved by the PIF-Net and the MSFF module have their characteristics on different sub-regions. Specifically, for the WT area, the PIF-Net is more effective in improving the segmentation accuracy than the MSFF module. On the other hand, for the TC and ET areas, the MSFF module is more helpful in comparison to the PIF-Net. This phenomenon can be observed from the comparison between **OURS w/o PIF-Net** and **OURS w/o MSFF**. The results shown in Figure 7 also verify this point. By referring to the ground truth, we can see that **OURS w/o MSFF** generally obtains more accurate results for the ED area (shown in green) than **OURS w/o PIF-Net**, while **OURS w/o PIF-Net** performs better for the NCR/NET and ET areas (shown in red and yellow). We provide an explanation to this observation as follows. The

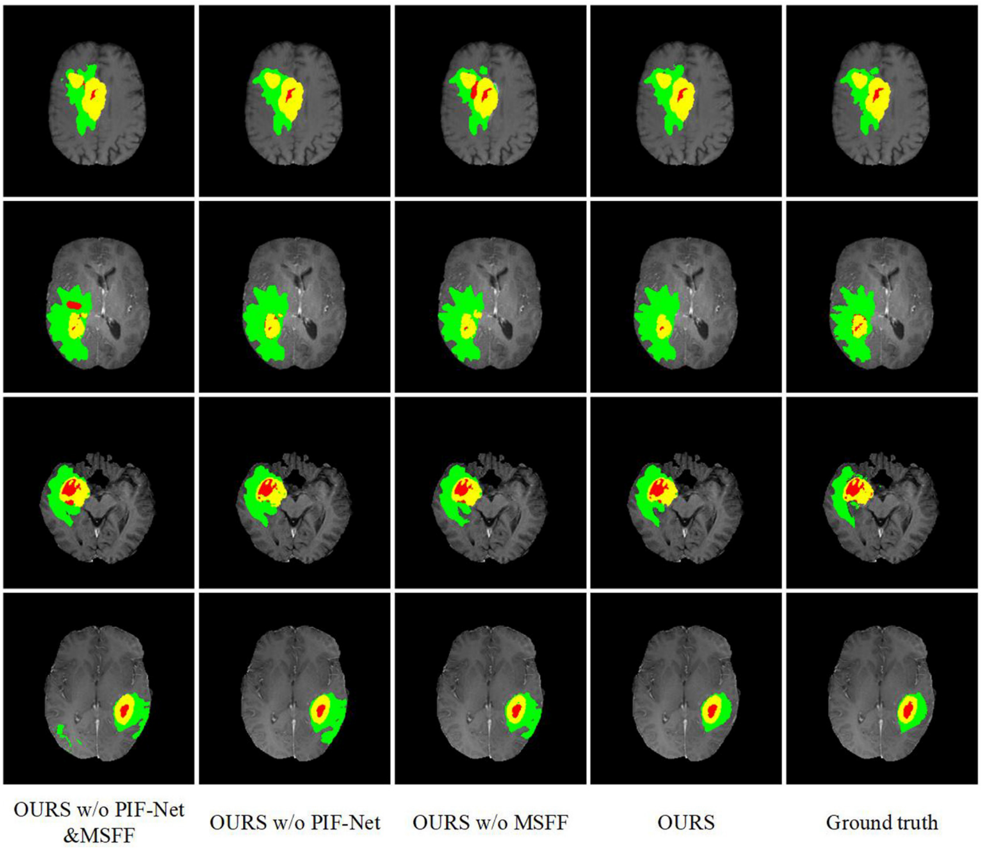


FIGURE 7
Examples of brain tumor segmentation results obtained by di3fferent methods in the ablation study. The green, red, and yellow regions indicate edema (ED), non-enhancing tumor and necrosis (NCR/NET), and enhancing tumor (ET), respectively.

segmentation of WT is mainly based on the ED area that can be effectively captured in the T2 and Flair volumes. The modality characteristics on the ED area in T2 and Flair volumes are generally close, so the requirement of multimodal feature fusion or selection is not very urgent. By contrast, the pixel-level image fusion achieved by the PIF-Net can enrich the input modalities for the segmentation model and this modality enhancement approach can also be viewed as a data augmentation method to some extent, which tends to be relatively more effective for WT segmentation as only two source modalities are used. In comparison to WT, the segmentation of TC and ET is more difficult due to the factors like smaller size, more irregular shape, etc. As a result, more modalities are typically required in TC and ET segmentation. In such a situation, the refinement of multimodal features achieved by the MSFF module is of higher significance. Therefore, the segmentation of TC and ET benefits more from the MSFF module. Nevertheless, it is worth noting that our PIF-Net and MSFF module both improve the segmentation accuracy of all the three sub-regions, just with different extents.

4.4. Comparison with other methods

In this subsection, we compare the proposed method with some existing brain tumor segmentation methods, which are mainly included in the proceedings of BraTS 2019-2021 challenges and generally have good performance. Tables 4, 5 report the evaluation results of different methods on BraTS 2019 and BraTS 2020 validation sets, respectively. For the comparison methods, the results reported in the original publications are adopted since the benchmarks used are exactly the same. In addition, the results obtained by a single model instead of multi-model ensemble are used for the sake of fair comparison. In each case, the best score is indicated in bold and the second best score is underlined. We can observe from Tables 4, 5 that the proposed method achieves very competitive performance among all the methods. For WT and TC regions, the proposed method obtains the highest Dice scores on both BraTS 2019 and BraTS 2020 validation sets. Our method achieves 0.8265 and 0.8291 in terms of the mean Dice score on these two datasets, which are both in the second place among all the methods. It is worth mentioning

TABLE 4 Objective evaluation results of different brain tumor segmentation methods on the BraTS 2019 validation sets.

References	WT		TC		ET		Average	
	Dice	Hausdorff	Dice	Hausdorff	Dice	Hausdorff	Dice	Hausdorff
Xu et al. (2019)	0.8930	6.9640	0.8070	7.6630	0.7590	**4.1930**	0.8197	**6.2733**
Baid et al. (2019)	0.8700	13.3600	0.7700	12.7100	0.7000	6.4500	0.7800	10.8400
González et al. (2019)	0.8882	8.1231	0.7833	7.5618	0.7231	4.9132	0.7982	6.8660
Lorenzo et al. (2019)	0.8904	-	0.7511	-	0.6634	-	0.7683	-
Ahmad et al. (2019)	0.8518	9.0083	0.7576	10.6744	0.6230	8.4683	0.7441	9.3837
Abraham and Khan (2019)	0.8605	-	0.7108	-	0.6323	-	0.7345	-
Bhalerao and Thakur (2019)	0.8527	8.0793	0.7091	9.5708	0.6668	7.2700	0.7429	8.3067
Yan et al. (2019)	0.8600	40.3100	0.7300	10.4000	0.6600	18.5300	0.7500	23.0800
Iantsen et al. (2019)	0.8700	8.3500	0.7900	9.5800	0.6700	7.8200	0.7767	8.5833
Astaraki et al. (2019)	0.8700	5.9000	0.8100	**7.1600**	0.7100	6.0200	0.7967	6.3600
Cao et al. (2021)	0.8938	7.5050	0.7875	9.2600	**0.7849**	6.9250	0.8221	7.8967
Wang et al. (2021)	0.8889	7.5990	0.8141	7.5840	0.7836	5.9080	**0.8289**	7.0303
Valanarasu et al. (2021)	0.8760	8.9420	0.7392	9.8930	0.7321	6.3230	0.7824	8.3860
OURS	**0.8942**	**5.3490**	**0.8142**	10.8988	0.7710	5.8548	0.8265	7.3675

Bold and underlined values indicate the best scores and second best scores on each metric (each column in the tables) among all the methods.

TABLE 5 Objective evaluation results of different brain tumor segmentation methods on the BraTS 2020 validation sets.

References	WT		TC		ET		Average	
	Dice	Hausdorff	Dice	Hausdorff	Dice	Hausdorff	Dice	Hausdorff
Jun et al. (2020)	0.8780	6.3000	0.7790	11.0200	0.7520	30.6500	0.8030	15.9900
Liu et al. (2020a)	0.8823	6.4900	0.8012	**6.6800**	0.7637	21.3900	0.8157	11.5200
Messaoudi et al. (2020)	0.8413	-	0.6804	-	0.6537	-	0.7251	-
Sun et al. (2020)	0.8920	-	0.7880	-	0.7230	-	0.8010	-
Cirillo et al. (2020)	0.8926	6.3900	0.7919	14.0700	0.7504	36.0000	0.8116	18.8200
Pang et al. (2020)	0.8811	18.0901	0.7605	29.0570	0.7538	34.2391	0.7985	27.1287
Sundaresan et al. (2020)	0.8900	**4.4000**	0.7700	15.3000	0.7700	29.4000	0.8100	16.3667
Ballestar and Vilaplana (2020)	0.8300	12.3400	0.7700	13.1100	0.7200	37.4200	0.7733	20.9567
McHugh et al. (2020)	0.8810	6.7200	0.7890	10.2000	0.7120	40.6000	0.7940	19.1733
Ma et al. (2020b)	0.8794	-	0.7731	-	0.7040	-	0.7855	-
Cao et al. (2021)	0.8934	7.855	0.7760	14.5940	**0.7895**	11.0050	0.8196	11.1513
Wang et al. (2021)	0.8900	6.4690	0.8136	10.4680	0.7850	16.7160	**0.8295**	11.2177
Zhang et al. (2021b)	0.8800	6.9500	0.7400	30.1800	0.7000	38.6000	0.7733	25.2433
OURS	**0.8950**	5.3117	**0.8178**	9.4285	0.7745	**4.4715**	0.8291	**6.4039**

Bold and underlined values indicate the best scores and second best scores on each metric (each column in the tables) among all the methods.

that the performance of proposed method may be slightly inferior to some latest state-of-the-art methods. However, the main purpose of this work is to verify the effectiveness of the proposed pixel-level and feature-level image fusion approaches for brain tumor segmentation. The segmentation model and loss function adopted in this work are both plain while popular approaches (i.e., the original V-Net and the BCE-and-Dice-based loss) in 3D medical image segmentation. By introducing some advanced architectures and loss functions, we believe that the segmentation performance can be further improved.

5. Conclusion

In this paper, we mainly introduce pixel-level and feature-level image fusion techniques for MRI-based brain tumor segmentation, aiming to achieve more sufficient and finer utilization of multimodal information. Specifically, we present a CNN-based 3D pixel-level image fusion network named PIF-Net to enrich the input modalities of the segmentation model and design an attention-based feature fusion module named MSFF for multimodal feature refinement. A two-stage brain tumor segmentation framework is accordingly proposed based on the PIF-Net, the MSFF module and the V-Net. Experimental results on the BraTS 2019 and BraTS 2020 benchmarks show that the proposed components on both pixel-level and feature-level fusion can effectively improve the segmentation accuracy of all the three tumor sub-regions including whole tumor, tumor core and enhancing tumor. The pixel-level image fusion network in this work is trained independently to the segmentation model. Future work may concentrate on integrating image fusion and segmentation into a unified network for better feature learning to further improve the segmentation performance.

Data availability statement

The datasets for this study can be found in the BraTS 2019 dataset available at: https://www.med.upenn.edu/cbica/brats2019/data.html and in the BraTS 2020 dataset available at: https://www.med.upenn.edu/cbica/brats2020/data.html.

Author contributions

YL: conceptualization, methodology, and writing. FM: methodology, experiments, and Writing. YS: methodology and experiments. JC: methodology and review. CL: experiments and review. XC: methodology, review, and supervision. All authors contributed to the work and approved the submission.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grants 62176081, 61922075, 62171176, and 41901350, and in part by the Fundamental Research Funds for the Central Universities under Grants JZ2020HGPA0111 and JZ2021HGPA0061.

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

Abraham, N., and Khan, N. M. (2019). "Multimodal segmentation with MGF-Net and the focal tversky loss function," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 191–198. doi: 10.1007/978-3-030-46643-5_18

Ahmad, P., Qamar, S., Hashemi, S. R., and Shen, L. (2019). "Hybrid labels for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 158–166. doi: 10.1007/978-3-030-46643-5_15

Astaraki, M., Wang, C., Carrizo, G., Toma-Dasu, I. and Smedby, Ö. (2019). "Multimodal brain tumor segmentation with normal appearance autoencoder," in *International MICCAI Brainlesion Workshop* (Springer), 316–323. doi: 10.1007/978-3-030-46643-5_31

Baid, U., Shah, N. A., and Talbar, S. (2019). "Brain tumor segmentation with cascaded deep convolutional neural network," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 90–98. doi: 10.1007/978-3-030-46643-5_9

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629*.

Ballestar, L. M., and Vilaplana, V. (2020). "MRI brain tumor segmentation and uncertainty estimation using 3D-UNet architectures," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 376–390. doi: 10.1007/978-3-030-72084-1_34

Bauer, S., Nolte, L.-P., and Reyes, M. (2011). "Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Berlin, Heidelberg: Springer), 354–361. doi: 10.1007/978-3-642-23626-6_44

Bhalerao, M., and Thakur, S. (2019). "Brain tumor segmentation based on 3D residual U-Net," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 218–225. doi: 10.1007/978-3-030-46643-5_21

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., et al. (2021). Swin-unet: UNet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*. doi: 10.48550/arXiv.2105.05537

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). TransUNet: transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*. doi: 10.48550/arXiv.2102.04306

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-assisted Intervention* (Athens, Greece: Springer), 424–432. doi: 10.1007/978-3-319-46723-8_49

Cirillo, M. D., Abramian, D., and Eklund, A. (2020). "Vox2Vox: 3D-GAN for brain tumour segmentation," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 274–284. doi: 10.1007/978-3-030-72084-1_25

Du, J., Li, W., Lu, K., and Xiao, B. (2016). An overview of multi-modal medical image fusion. *Neurocomputing* 215, 3–20. doi: 10.1016/j.neucom.2015.07.160

Goetz, M., Weber, C., Bloecher, J., Stieltjes, B., Meinzer, H.-P., and Maier-Hein, K. (2014). "Extremely randomized trees based brain tumor segmentation," in *Proceeding of MICCAI BRATS Challenge* (Boston, MA, USA), 6–11.

González, S. R., Sekou, T. B., Hidane, M., and Tauber, C. (2019). "3D automatic brain tumor segmentation using a multiscale input U-Net network," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 113–123. doi: 10.1007/978-3-030-46643-5_11

Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., et al. (2022). Attention mechanisms in computer vision: a survey. *Comp. Visual Media* 8, 331–368. doi: 10.1007/s41095-022-0271-y

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., et al. (2017). Brain tumor segmentation with deep neural networks. *Med. Image Anal.* 35, 18–31. doi: 10.1016/j.media.2016.05.004

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, USA), 7132–7141. doi: 10.1109/CVPR.2018.00745

Iantsen, A., Jaouen, V., Visvikis, D., and Hatt, M. (2019). "Encoder-decoder network for brain tumor segmentation on multi-sequence MRI," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 296–302. doi: 10.1007/978-3-030-46643-5_29

Isensee, F., Jager, P. F., Full, P. M., Vollmuth, P., and Maier-Hein, K. H. (2020). "nnU-Net for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 118–132. doi: 10.1007/978-3-030-72087-2_11

Jun, W., Haoxiang, X., and Wang, Z. (2020). "Brain tumor segmentation using dual-path attention U-Net in 3D MRI images," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 183–193. doi: 10.1007/978-3-030-72084-1_17

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., et al. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* 36, 61–78. doi: 10.1016/j.media.2016.10.004

Kleesiek, J., Biller, A., Urban, G., Kothe, U., Bendszus, M., and Hamprecht, F. (2014). "Ilastik for multi-modal brain tumor segmentation," in *Proceeding of MICCAI BRATS Challenge* (Boston, MA, USA), 12–17.

Li, S., Kang, X., Fang, L., Hu, J., and Yin, H. (2017). Pixel-level image fusion: a survey of the state of the art. *Inform. Fusion* 33, 100–112. doi: 10.1016/j.inffus.2016.05.004

Li, X., Luo, G., and Wang, K. (2019a). "Multi-step cascaded networks for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 163–173. doi: 10.1007/978-3-030-46640-4_16

Li, X., Wang, W., Hu, X., and Yang, J. (2019b). "Selective kernel networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (California, USA), 510–519. doi: 10.1109/CVPR.2019.00060

Liang, X., Hu, P., Zhang, L., Sun, J., and Yin, G. (2019). MCFNet: multi-layer concatenation fusion network for medical images fusion. *IEEE Sensors J.* 19, 7107–7119. doi: 10.1109/JSEN.2019.2913281

Liu, C., Ding, W., Li, L., Zhang, Z., Pei, C., Huang, L., et al. (2020a). "Brain tumor segmentation network using attention-based fusion and spatial relationship constraint," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 219–229. doi: 10.1007/978-3-030-72084-1_20

Liu, J., Li, M., Wang, J., Wu, F., Liu, T., and Pan, Y. (2014). A survey of MRI-based brain tumor segmentation methods. *Tsinghua Sci. Technol.* 19, 578–595. doi: 10.1109/TST.2014.6961028

Liu, Y., Chen, X., Cheng, J., and Peng, H. (2017). "A medical image fusion method based on convolutional neural networks," in *2017 20th International Conference on Information Fusion (Fusion)* (Xian, China), 1070–1076. doi: 10.23919/ICIF.2017.8009769

Liu, Y., Chen, X., Liu, A., Ward, R. K., and Wang, Z. J. (2021). Recent advnaces in sparse representation based medical image fusion. *IEEE Instrument. Meas. Mag.* 24, 45–53. doi: 10.1109/MIM.2021.9400960

Liu, Y., Chen, X., Wang, Z., Wang, Z., Ward, R., and Wang, X. (2018). Deep learning for pixel-level image fusion: recent advances and future prospects. *Inform. Fusion* 42, 158–173. doi: 10.1016/j.inffus.2017.10.007

Liu, Y., Chen, X., Ward, R. K., and Wang, Z. J. (2016). Image fusion with convolutional sparse representation. *IEEE Signal Process. Lett.* 23, 1882–1886. doi: 10.1109/LSP.2016.2618776

Liu, Y., Chen, X., Ward, R. K., and Wang, Z. J. (2019). Medical image fusion via convolutional sparsity based morphological component analysis. *IEEE Signal Process. Lett.* 26, 485–489. doi: 10.1109/LSP.2019.2895749

Liu, Y., Liu, S., and Wang, Z. (2015). A general framework for image fusion based on multi-scale transform and sparse representation. *Inform. Fusion* 24, 147–164. doi: 10.1016/j.inffus.2014.09.004

Liu, Y., Shi, Y., Mu, F., Cheng, J., Li, C., and Chen, X. (2022). Multimodal mri volumetric data fusion with convolutional neural networks. *IEEE Trans, Instrument. Meas.* 71, 4006015. doi: 10.1109/TIM.2022.3184360

Liu, Y., Wang, L., Cheng, J., Li, C., and Chen, X. (2020b). Multi-focus image fusion: a survey of the state of the art. *Inform. Fusion* 64, 71–91. doi: 10.1016/j.inffus.2020.06.013

Lorenzo, P. R., Marcinkiewicz, M., and Nalepa, J. (2019). "Multi-modal U-Nets with boundary loss and pre-training for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 135–147. doi: 10.1007/978-3-030-46643-5_13

Lyu, C., and Shu, H. (2020). "A two-stage cascade model with variational autoencoders and attention gates for MRI brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 435–447. doi: 10.1007/978-3-030-72084-1_39

Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., and Ma, Y. (2022). SwinFusion: cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA. Automat. Sin.* 9, 1200–1217. doi: 10.1109/JAS.2022.105686

Ma, J., Xu, H., Jiang, J., Mei, X., and Zhang, X.-P. (2020a). DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans. Image Process.* 29, 4980–4995. doi: 10.1109/TIP.2020.2977573

Ma, J., and Yang, X. (2018). "Automatic brain tumor segmentation by exploring the multi-modality complementary information and cascaded 3D lightweight CNNs," in *International MICCAI Brainlesion Workshop* (Granada, Spain: Springer), 25–36. doi: 10.1007/978-3-030-11726-9_3

Ma, S., Zhang, Z., Ding, J., Li, X., Tang, J., and Guo, F. (2020b). "A deep supervision CNN network for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 158–167. doi: 10.1007/978-3-030-72087-2_14

McHugh, H., Talou, G. M., and Wang, A. (2020). "2d Dense-UNet: a clinically valid approach to automated glioma segmentation," in *International MICCAI Brainlesion Workshop* (Springer), 69–80. doi: 10.1007/978-3-030-72087-2_7

Meier, R., Bauer, S., Slotboom, J., Wiest, R., and Reyes, M. (2014). "Patient-specific semi-supervised learning for postoperative brain tumor segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Boston, MA, USA: Springer), 714–721. doi: 10.1007/978-3-319-10404-1_89

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Messaoudi, H., Belaid, A., Allaoui, M. L., Zetout, A., Allili, M. S., Tliba, S., et al. (2020). "Efficient embedding network for 3D brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 252–262. doi: 10.1007/978-3-030-72084-1_23

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-Net: fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)* (Stanford, CA, USA: IEEE), 565–571. doi: 10.1109/3DV.2016.79

Myronenko, A. (2018). "3D MRI brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop* (Granada, Spain: Springer), 311–320. doi: 10.1007/978-3-030-11726-9_28

Pang, E., Shi, W., Li, X., and Wu, Q. (2020). "Glioma segmentation using encoder-decoder network and survival prediction based on cox analysis," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 318–326. doi: 10.1007/978-3-030-72084-1_29

Pinto, A., Pereira, S., Correia, H., Oliveira, J., Rasteiro, D. M., and Silva, C. A. (2015). "Brain tumour segmentation based on extremely randomized forest with high-level features," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Milan, Italy: IEEE), 3037–3040. doi: 10.1109/EMBC.2015.7319032

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention (MICCAI)* (Munich, Germany: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Subbanna, N. K., Precup, D., Collins, D. L., and Arbel, T. (2013). "Hierarchical probabilistic Gabor and MRF segmentation of brain tumours in MRI volumes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Nagoya, Japan), 751–758. doi: 10.1007/978-3-642-40811-3_94

Sun, J., Peng, Y., Li, D., and Guo, Y. (2020). "Segmentation of the multimodal brain tumor images used Res-U-Net," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 263–273. doi: 10.1007/978-3-030-72084-1_24

Sundaresan, V., Griffanti, L., and Jenkinson, M. (2020). "Brain tumour segmentation using a triplanar ensemble of U-Nets on MR images," in *International MICCAI Brainlesion Workshop* (Lima, Peru: Springer), 340–353. doi: 10.1007/978-3-030-72084-1_31

Tang, W., Liu, Y., Cheng, J., Li, C., and Chen, X. (2021). Green fluorescent protein and phase contrast image fusion via detail preserving cross network. *IEEE Trans. Comput. Imaging* 7, 584–597. doi: 10.1109/TCI.2021.3083965

Valanarasu, J. M. J., Sindagi, V. A., Hacihaliloglu, I., and Patel, V. M. (2021). Kiu-Net: overcomplete convolutional architectures for biomedical image and volumetric segmentation. *IEEE Trans. Med. Imaging* 41, 965–976. doi: 10.1109/TMI.2021.3130469

Wang, G., Li, W., Ourselin, S., and Vercauteren, T. (2017). "Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks," in *International MICCAI Brainlesion Workshop* (Quebec City, QC, Canada: Springer), 178–190. doi: 10.1007/978-3-319-75238-9_16

Wang, L., Li, B., and Tian, L. (2014). Multimodal medical volumetric data fusion using 3-D discrete shearlet transform and global-to-local rule. *IEEE Trans. Biomed. Eng.* 61, 197–206. doi: 10.1109/TBME.2013.2279301

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., and Li, J. (2021). "TransBTS: Multimodal brain tumor segmentation using transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Strasbourg, France: Springer), 109–119. doi: 10.1007/978-3-030-87193-2_11

Xu, H., and Ma, J. (2021). EMFusion: an unsupervised enhanced medical image fusion network. *Inform. Fusion* 76, 177–186. doi: 10.1016/j.inffus.2021.06.001

Xu, H., Ma, J., Jiang, J., Guo, X., and Ling, H. (2022). U2Fusion: a unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 502–518. doi: 10.1109/TPAMI.2020.3012548

Xu, X., Zhao, W., and Zhao, J. (2019). "Brain tumor segmentation using attention-based network in 3D MRI images," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 3–13. doi: 10.1007/978-3-030-466 43-5_1

Yan, K., Sun, Q., Li, L., and Li, Z. (2019). "3D Deep residual encoder-decoder CNNS with squeeze-and-excitation for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Shenzhen, China: Springer), 234–243. doi: 10.1007/978-3-030-46643-5_23

Yang, Y., Que, Y., Huang, S., and Lin, P. (2016). Multimodal sensor medical image fusion based on type-2 fuzzy logic in nsct domain. *IEEE Sensors J.* 16, 3735–3745. doi: 10.1109/JSEN.2016.2533864

Yin, H. (2018). Tensor sparse representation for 3-D medical image fusion using weighted average rule. *IEEE Trans. Biomed. Eng.* 65, 2622–2633. doi: 10.1109/TBME.2018.2811243

Yin, M., Liu, X., Liu, Y., and Chen, X. (2019). Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain. *IEEE Trans. Instrument. Meas.* 68, 49–64. doi: 10.1109/TIM.2018.2838778

Zhang, D., Huang, G., Zhang, Q., Han, J., Han, J., Wang, Y., et al. (2020a). Exploring task structure for brain tumor segmentation from multi-modality MR images. *IEEE Trans. Image Process.* 29, 9032–9043. doi: 10.1109/TIP.2020.3023609

Zhang, H., Xu, H., Tian, X., Jiang, J., and Ma, J. (2021a). Image fusion meets deep learning: a survey and perspective. *Inform. Fusion* 76, 323–336. doi: 10.1016/j.inffus.2021.06.008

Zhang, Q., Liu, Y., Blum, R., Han, J., and Tao, D. (2018). Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: a review. *Inform. Fusion* 40, 57–75. doi: 10.1016/j.inffus.2017.05.006

Zhang, W., Yang, G., Huang, H., Yang, W., Xu, X., Liu, Y., et al. (2021b). ME-Net: multi-encoder net framework for brain tumor segmentation. *Int. J. Imaging Syst. Technol.* 31, 1834–1848. doi: 10.1002/ima.22571

Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., et al. (2022). mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. *arXiv preprint arXiv:2206.02425*. doi: 10.48550/arXiv.2206.02425

Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., and Zhang, L. (2020b). IFCNN: a general image fusion framework based on convolutional neural network. *Inform. Fusion* 54, 99–118. doi: 10.1016/j.inffus.2019.07.011

Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y., and Fan, Y. (2018). A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. *Med. Image Anal.* 43, 98–111. doi: 10.1016/j.media.2017.10.002

Zhou, C., Ding, C., Wang, X., Lu, Z., and Tao, D. (2020). One-pass multi-task networks with cross-task guided attention for brain tumor segmentation. *IEEE Trans. Image Process.* 29, 4516–4529. doi: 10.1109/TIP.2020.2973510

Zhu, Z., Yin, H., Chai, Y., Li, Y., and Qi, G. (2018). A novel multi-modality image fusion method based on image decomposition and sparse representation. *Inform. Sci.* 432, 516–529. doi: 10.1016/j.ins.2017.09.010

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 15 September 2022
DOI 10.3389/fnins.2022.1009581

Check for updates

OPEN ACCESS

EDITED BY
Yu Liu,
Hefei University of Technology, China

REVIEWED BY
Yu Zhang,
Beihang University, China
Xiaoqing Luo,
Jiangnan University, China
Lu Tang,
Xuzhou Medical University, China

*CORRESPONDENCE
Li Yin
yl1@cqu.edu.cn
Yuehui Yin
yinyh@hospital.cqmu.edu.cn
Hao Chen
Chenhao9364@163.com

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 02 August 2022
ACCEPTED 24 August 2022
PUBLISHED 15 September 2022

CITATION
Xu Y, He X, Xu G, Qi G, Yu K, Yin L, Yang P, Yin Y and Chen H (2022) A medical image segmentation method based on multi-dimensional statistical features. *Front. Neurosci.* 16:1009581. doi: 10.3389/fnins.2022.1009581

A medical image segmentation method based on multi-dimensional statistical features

Yang Xu[1], Xianyu He[1], Guofeng Xu[1], Guanqiu Qi[2], Kun Yu[1], Li Yin[3]*, Pan Yang[4,5], Yuehui Yin[6]* and Hao Chen[4]*

[1]College of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China, [2]Department of Computer Information Systems, Buffalo State College, Buffalo, NY, United States, [3]Chongqing Key Laboratory of Translational Research of Cancer Metastasis and Individualized Treatment, Chongqing University Cancer Hospital, Chongqing, China, [4]Department of Cardiovascular Surgery, Chongqing General Hospital, University of Chinese Academy of Sciences, Chongqing, China, [5]Department of Emergency, The Second Affiliated Hospital of Chongqing Medical University, Chongqing, China, [6]Department of Cardiology, The Second Affiliated Hospital of Chongqing Medical University, Chongqing, China

Medical image segmentation has important auxiliary significance for clinical diagnosis and treatment. Most of existing medical image segmentation solutions adopt convolutional neural networks (CNNs). Althought these existing solutions can achieve good image segmentation performance, CNNs focus on local information and ignore global image information. Since Transformer can encode the whole image, it has good global modeling ability and is effective for the extraction of global information. Therefore, this paper proposes a hybrid feature extraction network, into which CNNs and Transformer are integrated to utilize their advantages in feature extraction. To enhance low-dimensional texture features, this paper also proposes a multi-dimensional statistical feature extraction module to fully fuse the features extracted by CNNs and Transformer and enhance the segmentation performance of medical images. The experimental results confirm that the proposed method achieves better results in brain tumor segmentation and ventricle segmentation than state-of-the-art solutions.

KEYWORDS

medical image segmentation, deep learning, convolutional neural network, transformer, neural network

1. Background

Medical image segmentation is not only an important step in medical image analysis, but also an indispensable part of computer-aided diagnosis and pathology research. With the continuous development of computer vision in recent years, convolutional neural networks (CNNs), especially fully convolutional networks (FCNs), have made breakthroughs in the applications of medical image segmentation. For example, they have been applied to brain Magnetic Resonance Imaging (MRI) (Li et al., 2021), multi-organ segmentation, and cardiac ventricle (Moeskops et al., 2016; Hesamian et al., 2019).

FCNs enable end-to-end image semantic segmentation and have evolved many variants during development, U-Net (Ronneberger et al., 2015), V-Net (Milletari et al., 2016), 3D U-Net (Çiçek et al., 2016), Res-UNet (Xiao et al., 2018), density-unet (Li et al., 2018), Y-Net (Mehta et al., 2018), etc. have been specially proposed for image and volume segmentation according to various medical imaging modalities. Existing CNN-based methods have good image segmentation performance. Due to the limitation of convolution kernel size, each convolution kernel only focuses on local information. Therefore, it is difficult for these existing methods to generate any long-distance dependencies when performing image segmentation tasks. The ability to construct global contextual information is crucial for intensive prediction tasks during medical image segmentation.

To effectively address the issues on global contextual information, Transformer (Vaswani et al., 2017; Dosovitskiy et al., 2020) was proposed to handle the issues in sequence-to-sequence prediction. It uses a completely attention-based encoder-decoder architecture, which is completely different from CNN-based methods. A one-dimensional sequence is taken as input, so Transformer has a powerful modeling ability, not only in constructing global context information. The powerful capability, works well for downstream tasks in the case of large-scale pre-training.

Transformer has been widely used in medical image segmentation, but it only focuses on building global context information at all stages. Therefore, its ability to obtain local information is weakened, and the lack of detailed location information encoding reduces the distinguishability between background and target. Various CNN architectures such as U-Net provide a way to extract low-level visual information, which can well compensate for the spatial details of Transformer's local information.

Therefore, considering the above-mentioned advantages, some studies integrated CNNs and Transformer. For example, TransUNet (Chen et al., 2021), first used CNNs to extract local features, and then applied Transformer to global context modeling. This architecture not only establishes a self-attention mechanism, but also reduces the loss of local feature resolution brought by Transformer, making it have better image segmentation accuracy. However, TransUNet is only a simple integration of CNNs and Transformer, and there are some shortcomings in practical applications.

The low-dimensional image texture features mainly include structural features and statistical features. The image information contained in these features plays an important role in semantic segmentation. Chen et al. (2018) proposed the DeepLabv3+ model by adding an encoder to the DeepLabv3 (Chen et al., 2017) model to achieve the extraction and fusion of both shallow and deep image features. Li et al. (2020) proposed an edge preservation module to enhance low-dimensional edge features, effectively improving the performance of semantic segmentation. However, the above methods are all applied to shallow features or low-dimensional edge features. Although low-dimensional statistical features play an importance role in grasping global image features, only a small percent of existing solutions try to analyze them.

Therefore, this paper proposes a hybrid feature extraction network based on CNNs and Transformer. The proposed network can not only utilize the Transformer's ability to construct global contextual information, but can also use the CNN's ability to capture local information. Additionally, in order to use the statistical image features, this paper designs a multi-scale statistical feature extraction module to extract statistical image features to improve segmentation performance.

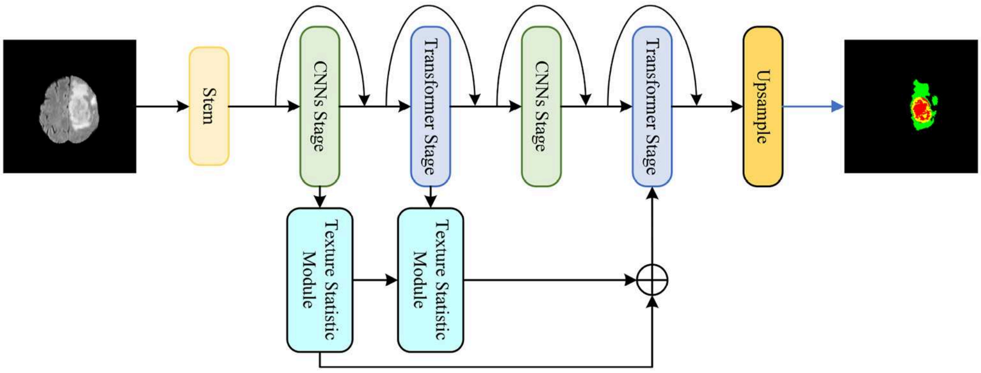


FIGURE 1
The proposed medical image segmentation method based on multi-dimensional statistical features.

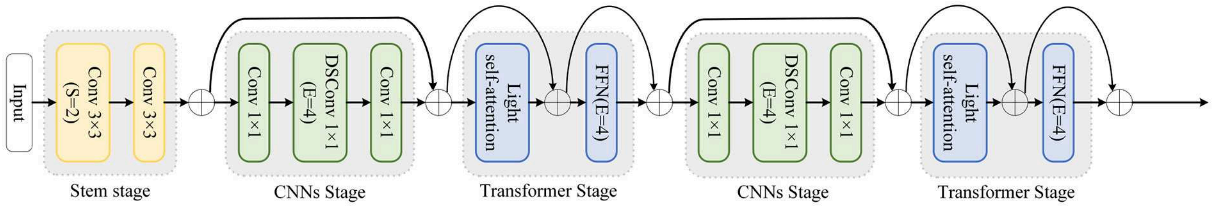


FIGURE 2
The proposed hybrid network consisting of CNNs stages and Transformer stage.

2. Related work

2.1. Semantic segmentation network

In the past few years, CNNs have been used as the main framework for various computer vision tasks, especially in semantic segmentation. The mainstream medical image segmentation methods use the encoder-decoder structured FCN and U-Net. U-Net++ (Zhou et al., 2018) designs more dense skip connections based on U-Net. Res-UNet (Xiao et al., 2018) introduces a residual module in ResNet (He et al., 2016), and designs a deeper network for feature extraction.

In the past 2 years, Vision Transformer (ViT) (Dosovitskiy et al., 2020) has demonstrated its powerful modeling capability in computer vision tasks. ViT splits the source image into patches and uses these patches to perform self-attention operations. The Swin Transformer (Liu et al., 2021) uses the shift idea to calculate the attention of different windows and layer the corresponding feature maps. MedT (Valanarasu et al., 2021) improves gated self-attention and applies Transformer to medical image segmentation.

Some recent solutions try to use the advantages of CNN and Transformer by integrating the two architectures as a new backbone network. The CMT (Guo et al., 2022) block consists of a depthwise convolution-based local perception unit and a light-weight transformer module. CoAtNet (Dai et al., 2021) fuses the two frameworks based on MBConv and relative self-attention. TransUNet (Chen et al., 2021) first fuses the U-shape structure of Transformer and U-Net and applies Transformer to medical image segmentation.

2.2. Statistical features

Statistical features as low-dimensional texture features play a key role in improving semantic segmentation performance. Many existing solutions exploit the texture information of statistical features. Simonyan et al. (2013) applied Fisher vector layers to enhance features using handcrafting. Wang et al. (2016) first proposed learnable histograms for semantic segmentation and object detection. Zhu et al. (2021) proposed a texture enhancement module and a pyramid texture extraction module to extract image texture features for the enhancement of semantic segmentation performance.

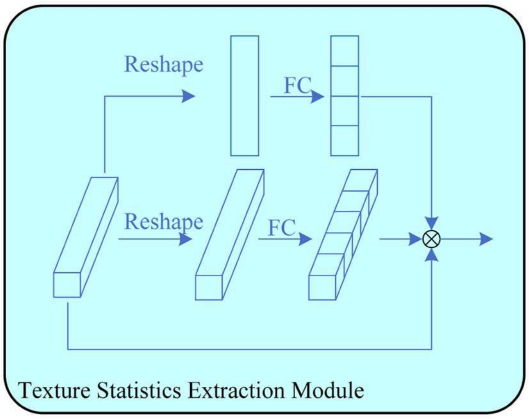


FIGURE 3
The proposed Texture Statistics Extraction Module. It is used to extract statistics at different stages.

3. Method

3.1. Semantic segmentation network

A medical image segmentation method is proposed based on multi-dimensional statistical features as shown in Figure 1. This method integrates CNNs and Transformer into the feature extraction network, and designs a texture statistics extraction module (TSEM) for the extraction and fusion of multi-dimensional statistical features.

3.2. Hybrid network

The proposed hybrid feature extraction network aims to utilize the advantages of CNNs and Transformer to achieve more

TABLE 1 Comparison of segmentation metrics on BraTS2018.

BraTS2018	WT		TC		ET		Average	
	Dice	HD	Dice	HD	Dice	HD	Dice	HD
Myronenko	90.40	4.483	85.90	8.278	81.40	3.805	85.90	5.500
U-Net++	88.96	5.327	84.65	8.535	79.49	4.285	84.36	6.049
CENet	89.53	5.271	84.31	8.493	79.95	4.379	84.60	6.193
D. Zhang	89.60	5.733	82.40	9.270	78.20	3.567	83.40	6.190
TransUNet	90.25	**4.390**	**87.19**	5.539	80.41	3.731	85.95	4.553
Proposed	**90.45**	4.923	86.96	**5.327**	**81.53**	**3.279**	**86.31**	**4.510**

Bold font represents the best result.

TABLE 2 Comparison of segmentation metrics on medical segmentation decathlon.

Cardiac Dataset	IoU	Dice	HD
U-Net	90.07	93.86	1.7414
U-Net++	90.55	94.38	1.7197
CENet	90.23	94.17	1.7682
TransUNet	90.67	94.54	1.7300
Proposed	**91.30**	**94.86**	**1.6772**

Bold font represents the best result.

accurate segmentation tasks. As shown in Figure 2, the proposed hybrid network is divided into five stages.

Stem is the first stage. CNNs and Transformer alternate in the remaining four stages. At the beginning of each stage, downsampling is applied to decrease feature map size and increase the number of channels. Additionally, the proposed network refers to the residual connection of ResNet and performs shortcuts at each stage.

Specifically, stem as the first stage contains two layers of simple 3×3 convolution. CNNs stage is the second stage, because the feature map is too large at this moment and not suitable for using Transformer in global feature extraction. The CNNs stage uses a Depthwise Separable Convolution block (DSConv) (Howard et al., 2017) to reduce the amount and size of model parameters. There is a 1×1 convolution layer before and after DSConv to change the feature map size and the number of channels. The third stage is the Transformer stage, which extracts global features after CNNs. The proposed network adopts a lightweight multi-head self-attention.

In the original self-attention module, the input $X \in \mathbb{R}^{C\times H\times W}$is linearized to query $Q \in \mathbb{R}^{n\times d_k}$, key $K \in \mathbb{R}^{n\times d_k}$, and value $V \in \mathbb{R}^{n\times d_v}$, where $n = H \times W$ is the number of patches, d, d_k, d_v represent input, key, and value's dimension. The self-attention output is obtained by the following formula.

$$Atten\left(Q,K,V\right) = \text{Soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

In order to reduce the overhead, the proposed network uses a $k\times k$ depthwise convolution with a stride of k to reduce the dimensions of K, V, ie $K' = DSVConv\left(K\right) \in \mathbb{R}^{\frac{n}{k^2}\times d_k}$ and $V' = DSVConv\left(V\right) \in \mathbb{R}^{\frac{n}{k^2}\times d_v}$, so the lightweight attention output is obtained by the following formula.

$$Atten\left(Q,K,V\right) = \text{Soft max}\left(\frac{QK'^T}{\sqrt{d_k}}\right)V' \qquad (2)$$

The CNNs and Transformer operations in the second and third stages are repeated in the subsequent fourth and fifth stages. Additionally, each stage is repeated L times. Stages 1 to 5 of the proposed network are were repeated 2, 2, 4, 2, and 8 times, respectively.

3.3. Texture statistics extraction module

The image texture information contains local structural features and global statistical properties. For poorly visualized images, the global statistical features are more suitable for segmentation. To effectively utilize statistical image features, a texture statistics extraction module (TSEM) is proposed. TSEM extracts statistical image features by encoding feature maps, as shown in Figure 3.

Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, the input is divided into three branches for multi-scale feature encoding. One branch is first processed by global average pooling to obtain channel average features, and then multiplied with the input feature map $X \in \mathbb{R}^{C\times 1\times 1}$ to obtain the final output feature map. Another branch first average pooling on one channel to obtain the feature map $X \in \mathbb{R}^{1\times H\times W}$, and then multiplies it with the input feature map $X \in \mathbb{R}^{C\times H\times W}$ to obtain the output feature map. The last two input feature maps are multiplied to obtain the output feature map of this module.

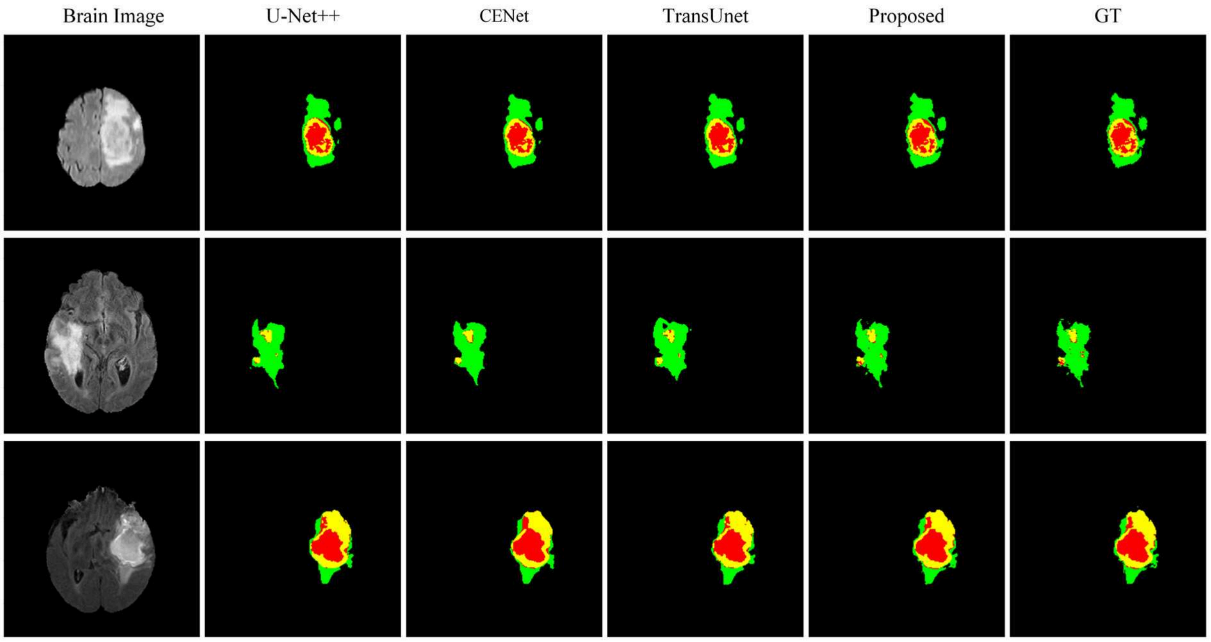


FIGURE 4
Comparison of the proposed method and other state-of-the-art methods on BraTS2018.

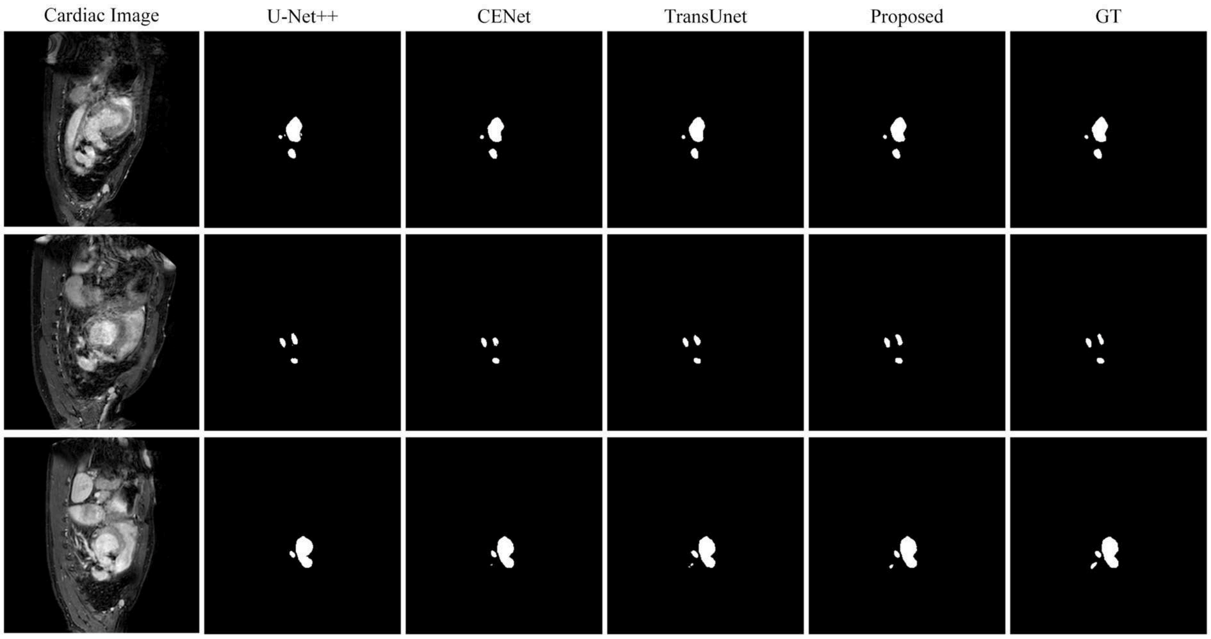


FIGURE 5
Comparison of the proposed method and other state-of-the-art methods on the Cardiac Dataset.

3.4. Loss function

To achieve the end-to-end training effect, a fusion loss function L_{fusion} is used to optimize the proposed method in the training process, training segmentation prediction and ground truth (GT). The loss function uses BCEDiceLoss, which is composed of binary cross entropy loss (BCELoss) and dice loss. The formula is given as follows:

$$L_{fusion} = \sum \left(0.5 * \left(-y\log\left(\hat{y}\right) - \left(1-y\right)\log\left(1-\hat{y}\right)\right) + \left(1 - \frac{2|y \cap \hat{y}|}{|y| + |\hat{y}|}\right)\right) \quad (3)$$

Where y represents GT and $\hat{y}$ represents the network prediction result.

4. Experiments

4.1. Datasets

To verify the effectiveness of the proposed method, BraTS2018 (Menze et al., 2014; Bakas et al., 2017, 2018) and the cardiac segmentation dataset in the medical segmentation (Antonelli et al., 2022) decathlon are used as training and testing datasets in the experiments. The BraTS2018 dataset has 285 annotated brain tumor magnetic resonance imaging (MRI) cases, and each case has four different modalities, namely Flair, T1, T1ce, and T2. This dataset needs to segment three different brain tumor regions, which are Whole Tumor (WT), Tumor Core (TC), Ehance Tumor (ET). The decathlon cardiac segmentation dataset contains 20 annotated mono-modal MRI cases, and this dataset requires the segmentation of the left atrium.

4.2. Experimental details

The model frameworks in this paper are all implemented based on Pytorch. The image size and batch size of the input BraTS2018 dataset are 240*240 and 8, respectively. The image size and batch size of the input cardiac dataset are 320*320 and 8, respectively. Four Tesla P100 GPUs were used in training. Adam (Kingma and Ba, 2014) is the optimizer used in this paper, and all parameters are set as default. The initial learning rate and weight decay for model training are 1e-3 and 1e-5, respectively.

4.3. Comparative experiments

To verify the efficiency of the proposed model framework, three most common metrics used in medical image segmentation, IoU score, Dice score and Hausdorff score (HD) are used. The corresponding formulas are given:

$$IoU = \frac{Y \cap \hat{Y}}{Y \cup \hat{Y}} \quad (4)$$

$$Dice = \frac{2\left|Y \cap \hat{Y}\right|}{|Y| + \left|\hat{Y}\right|} \quad (5)$$

Where Y represents GT and $\hat{Y}$ represents the network prediction result.

$$H(A,B) = \max\left(\max_{a\in A}\left\{\min_{b\in B}\|a-b\|\right\}, \max_{b\in B}\left\{\min_{a\in A}\|b-a\|\right\}\right) \quad (6)$$

Where $A = \left\{a_1, a_2, ..., a_p\right\}, B = \left\{b_1, b_2, ..., b_q\right\}$ represents the pixels of the prediction result and GT. $\|\cdot\|$ represents the norm between A and B.

This paper conducts comparative experiments with state-of-the-art image segmentation frameworks on the BraTS2018 and cardiac segmentation datasets. These frameworks include

TABLE 3 Comparison of the model size and flops cost.

Model	Input size	Parameter(M)	FLOPS(G)
U-Net	3, 224, 224	39.40	55.84
U-Net++	3, 224, 224	9.34	34.65
TransUNet	3, 224, 224	105.32	38.52
MedT	3, 224, 224	**1.60**	21.24
Proposed	3, 224, 224	37.25	**15.24**

Bold font represents the best result.

TABLE 4 Ablation experiment results on BraTS2018.

BraTS2018	WT		TC		ET		Average	
	Dice	HD	Dice	HD	Dice	HD	Dice	HD
C-C-C-C	88.31	5.322	86.32	5.531	80.68	4.293	85.10	5.049
T-T-T-T	88.15	5.514	86.19	6.681	80.64	4.450	84.99	5.548
C-T-C-T	89.04	5.357	86.91	5.554	80.91	3.315	85.32	4.742
Proposed	**90.45**	**4.923**	**86.96**	**5.327**	**81.53**	**3.279**	**86.31**	**4.510**

Bold font represents the best result.

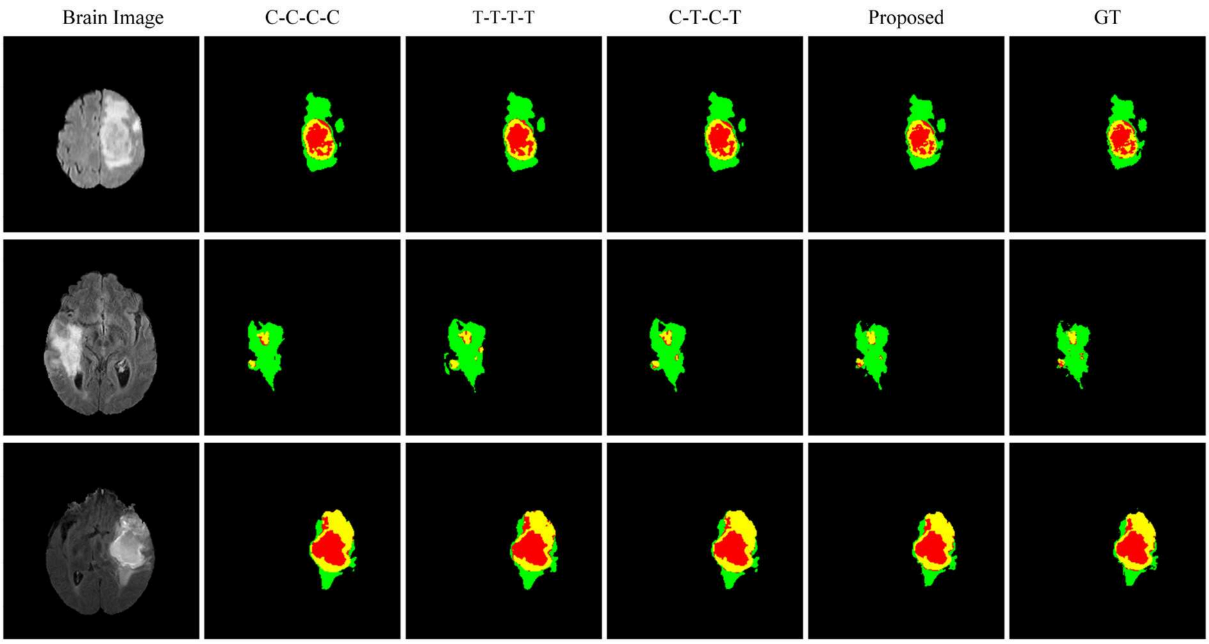


FIGURE 6
Visual results of ablation experiments on BraTS2018.

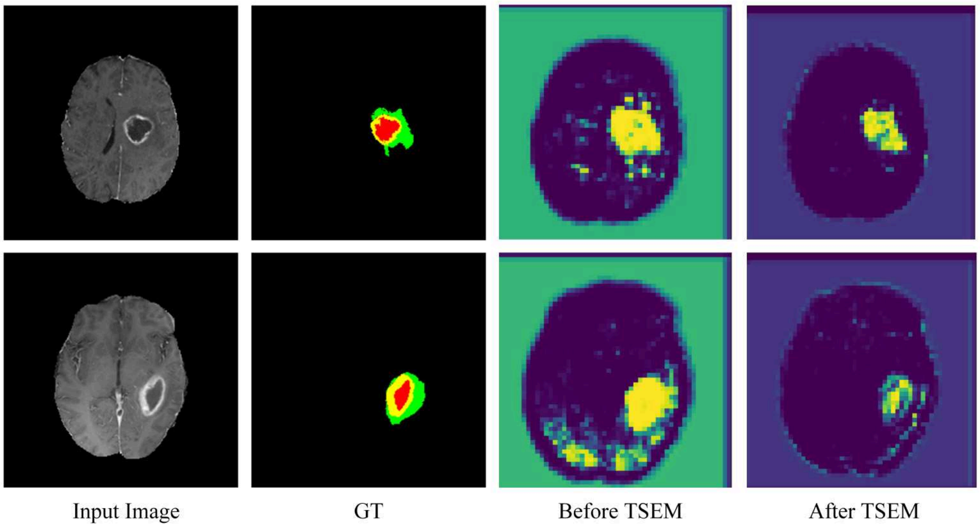


FIGURE 7
Performance comparison before and after adding TSEM.

2D CNN, 3D CNN segmentation frameworks (Ronneberger et al., 2015; Myronenko, 2018; Zhou et al., 2018; Gu et al., 2019; Zhang et al., 2020) and partial Transformer segmentation framework (Chen et al., 2021). The corresponding experimental results obtained by each method are shown in Tables 1, 2, and the visualized results are shown in Figures 4, 5. The number of parameters and computation cost are compared, as shown in Table 3.

According to the comparison results, the proposed segmentation framework obtains better scores and achieves a more significant performance improvement compared with state-of-the-art segmentation models. The proposed segmentation model achieves an average Dice of 86.31% on the BraTS2018 dataset and an average Dice of 94.86% on the medical segmentation decathlon, which are better than other state-of-the-art segmentation models.

According to the visualized results shown in Figure 4, the proposed method significantly improves the refinement of tumor and its texture features by using TSEM. Compared with other state-of-the-art, the model developed based on the integration of CNNs and Transformer has achieved better results in the context feature extraction and statistical feature fusion, and provides a reference for medical image segmentation of brain tumors and hearts. According to Table 3, the proposed method also has the lowest flops.

4.4. Ablation experiments

In order to further verify the importance and practical contribution of the backbone network used in this paper and the designed modules, the relevant ablation experiments are carried out. The index comparison of ablation experiments is shown in Table 4, and the experimental results are shown in Figures 6, 7.

This paper uses a fully convolutional layer as the Baseline for segmentation, and then replaces the backbone network blocks one by one for experiments. The experiments cover the full convolution network of C-C-C-C, the full transformer network of T-T-T-T, the hybrid network of C-T-C-T, and the TSEM is finally. The corresponding indicator values are shown in Table 4. The proposed module can improve the segmentation performance of baseline to a certain extent. After adding TSEM to the baseline, the corresponding improvement is the most obvious.

According to Table 4, the average Dice of the full Transformer is slightly lower than the result of the full CNN. The C-T-C-T result of the integration of CNNs and Transformer is significantly improved, confirming the effectiveness of the proposed hybrid network. After adding TSEM, the corresponding performance is further improved, the Dice of WT is increased by 1.41%, and the average Dice is increased by 0.99%.

Figure 6 shows the visualized brain tumor segmentation results obtained by each method in ablation experiments. After the backbone network becomes a hybrid network, the segmentation performance is further improved. After adding the texture statistics extraction module, the brain tumor edges after segmentation are significantly better, and the involved edges regions are closer to the actual situation compared with the segmentation result obtained by the hybrid network.

To further verify the role of TSEM, an intermediate experimental procedure is added. As shown in Figure 7, the area of interest in the feature map is concentrated and accurate after adding TSEM. Before adding TSEM, the feature map is mainly concentrated in the segmented area. Therefore, the proposed TSEM is conducive for the network to paying more attention to the segmented area and can effectively improve segmentation results.

5. Conclusion

This paper proposes a medical image segmentation method based on multi-dimensional statistical features. It consists of a hybrid feature extraction network and a multi-dimensional statistical feature extraction module. The hybrid feature extraction network is composed by CNNs and Transformer, and the lightweight processing is adopted to adapt to practical application scenarios. The multi-dimensional statistical feature extraction module is used to strengthen low-dimensional image texture features and enhance medical image segmentation performance. Experimental results show that the proposed medical image segmentation method achieves excellent results on brain tumor and heart segmentations.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://medicaldecathlon.com/; https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=37224922.

Author contributions

YX, XH, and GX: conceptualization, data curation, and visualization. XH and GQ: methodology and writing—original draft preparation. YX, XH, and GQ: software. KY, LY, and YY: validation. LY and HC: formal analysis. YX: investigation. YY and HC: resources. GQ, HC, LY, and YY: writing—review and editing. HC, LY, and YY: supervision. YX and HC: project administration. HC: funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding

This research was sponsored by the China Postdoctoral Science Foundation (2020M670111ZX), Chongqing medical scientific research project (Joint project of Chongqing Health Commission and Science and Technology Bureau, 2020GDRC019 and 2022MSXM184), Natural Science Foundation of Chongqing (cstc2020jcyj-bshX0068), and Special Fund for Young and Middle-aged Medical Top Talents of Chongqing (ZQNYXGDRCGZS2019005).

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

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., et al. (2022). The medical segmentation decathlon. *Nat. Commun.* 13, 1–13. doi: 10.1038/s41467-022-30695-9

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017). Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific Data* 4, 1–13. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*. doi: 10.48550/arXiv.1811.02629

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). Transunet: transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*. doi: 10.48550/arXiv.2102.04306

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*. doi: 10.48550/arXiv.1706.05587

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 801–818.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3D u-Net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 424–432.

Dai, Z., Liu, H., Le, Q. V., and Tan, M. (2021). Coatnet: marrying convolution and attention for all data sizes. *Adv. Neural Inf. Process. Syst.* 34, 3965–3977. doi: 10.48550/arXiv.2106.04803

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. doi: 10.48550/arXiv.2010.11929

Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., et al. (2019). Ce-net: context encoder network for 2d medical image segmentation. *IEEE Trans. Med. Imaging* 38, 2281–2292. doi: 10.1109/TMI.2019.2903562

Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., et al. (2022). "CMT: convolutional neural networks meet vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA), 12175–12185.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 770–778.

Hesamian, M. H., Jia, W., He, X., and Kennedy, P. (2019). Deep learning techniques for medical image segmentation: achievements and challenges. *J. Digit. Imaging* 32, 582–596. doi: 10.1007/s10278-019-00227-x

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). Mobilenets: efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*. doi: 10.48550/arXiv.1704.04861

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. doi: 10.48550/arXiv.1412.6980

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., and Heng, P.-A. (2018). H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Trans. Med. Imaging* 37, 2663–2674. doi: 10.1109/TMI.2018.2845918

Li, X., Li, X., Zhang, L., Cheng, G., Shi, J., Lin, Z., et al. (2020). "Improving semantic segmentation via decoupled body and edge supervision," in *European Conference on Computer Vision* (Springer), 435–452.

Li, Y., Wang, Z., Yin, L., Zhu, Z., Qi, G., and Liu, Y. (2021). X-net: a dual encoding-decoding method in medical image segmentation. *Vis. Comput.* 1–11. doi: 10.1007/s00371-021-02328-7

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal, QC: IEEE), 10012–10022.

Mehta, S., Mercan, E., Bartlett, J., Weaver, D., Elmore, J. G., and Shapiro, L. (2018). "Y-net: joint segmentation and classification for diagnosis of breast biopsy images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Granada: Springer), 893–901.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2014). The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)* (Stanford, CA: IEEE), 565–571.

Moeskops, P., Wolterink, J. M., van der Velden, B. H., Gilhuijs, K. G., Leiner, T., Viergever, M. A., et al. (2016). "Deep learning for multi-task medical image segmentation in multiple modalities," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Athens: Springer), 478–486.

Myronenko, A. (2018). "3D MRI brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop* (Granada: Springer), 311–320.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 234–241.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). "Deep fisher networks for large-scale image classification," in *Advances in Neural Information. Processing Systems* (Carson City, NV), 26.

Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., and Patel, V. M. (2021). "Medical transformer: gated axial-attention for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Strasbourg: Springer), 36–46.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information. Processing Systems* (Long Beach, CA), 30.

Wang, Z., Li, H., Ouyang, W., and Wang, X. (2016). "Learnable histogram: Statistical context features for deep neural networks," in *European Conference on Computer Vision* (Amsterdam: Springer), 246–262.

Xiao, X., Lian, S., Luo, Z., and Li, S. (2018). "Weighted res-unet for high-quality retina vessel segmentation," in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)* (Hangzhou: IEEE), 327–331.

Zhang, D., Huang, G., Zhang, Q., Han, J., Han, J., Wang, Y., et al. (2020). Exploring task structure for brain tumor segmentation from multi-modality mr images. *IEEE Trans. Image Process.* 29, 9032–9043. doi: 10.1109/TIP.2020.3 023609

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. (2018). "Unet++: a nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Springer), 3–11.

Zhu, L., Ji, D., Zhu, S., Gan, W., Wu, W., and Yan, J. (2021). "Learning statistical texture for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12537–12546.

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 28 October 2022
DOI 10.3389/fnins.2022.1055451

Check for updates

OPEN ACCESS

EDITED BY
Wei Wei,
The First Affiliated Hospital of University of Science and Technology of China Anhui Provincial Hospital, China

REVIEWED BY
Huafeng Li,
Kunming University of Science and Technology, China
Guanqiu Qi,
Buffalo State College, United States

*CORRESPONDENCE
Xiangzhi Bai
jackybxz@buaa.edu.cn
Li Zhang
chinazhangli@mail.tsinghua.edu.cn
Qing Zhang
jstzhangqing@vip.sina.com

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 27 September 2022
ACCEPTED 11 October 2022
PUBLISHED 28 October 2022

CITATION
Zhang Y, Xiang W, Zhang S, Shen J, Wei R, Bai X, Zhang L and Zhang Q (2022) Local extreme map guided multi-modal brain image fusion. *Front. Neurosci.* 16:1055451. doi: 10.3389/fnins.2022.1055451

Local extreme map guided multi-modal brain image fusion

Yu Zhang[1], Wenhao Xiang[2], Shunli Zhang[3], Jianjun Shen[2], Ran Wei[4], Xiangzhi Bai[1]*, Li Zhang[2]* and Qing Zhang[5]*

[1]School of Astronautics, Beihang University, Beijing, China, [2]Department of Electronic Engineering, Tsinghua University, Beijing, China, [3]School of Software Engineering, Beijing Jiaotong University, Beijing, China, [4]Department of Radiation Oncology, National Cancer Center, National Clinical Research Center for Cancer, Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, [5]Department of Orthopaedic Oncology, Beijing Jishuitan Hospital, Beijing, China

Multi-modal brain image fusion targets on integrating the salient and complementary features of different modalities of brain images into a comprehensive image. The well-fused brain image will make it convenient for doctors to precisely examine the brain diseases and can be input to intelligent systems to automatically detect the possible diseases. In order to achieve the above purpose, we have proposed a local extreme map guided multi-modal brain image fusion method. First, each source image is iteratively smoothed by the local extreme map guided image filter. Specifically, in each iteration, the guidance image is alternatively set to the local minimum map of the input image and local maximum map of previously filtered image. With the iteratively smoothed images, multiple scales of bright and dark feature maps of each source image can be gradually extracted from the difference image of every two continuously smoothed images. Then, the multiple scales of bright feature maps and base images (i.e., final-scale smoothed images) of the source images are fused by the elementwise-maximum fusion rule, respectively, and the multiple scales of dark feature maps of the source images are fused by the elementwise-minimum fusion rule. Finally, the fused bright feature map, dark feature map, and base image are integrated together to generate a single informative brain image. Extensive experiments verify that the proposed method outperforms eight state-of-the-art (SOTA) image fusion methods from both qualitative and quantitative aspects and demonstrates great application potential to clinical scenarios.

KEYWORDS

multi-modal brain images, image fusion, image guided filter, local extreme map, bright and dark feature map

1. Introduction

With the development of the medical imaging techniques, patients are often required to take multiple modalities of images, such as computed tomography (CT), magnetic resonance (MR) image, positron emission tomography (PET), and single-photon emission computed tomography (SPECT). Specifically, CT image mainly captures dense structures, such as bones and implants. MR image can capture soft-tissue information clearly, such as muscle and tumor. PET image can help reveal the metabolic or biochemical function of tissues and organs and SPECT image can visualize the conditions of organs, tissues, and bones through delivering a gamma-emitting radioisotope into the patient. Then, through observing all these captured medical images, the doctors can precisely diagnose the possible diseases. However, accurately locating the lesions and diagnosing the corresponding diseases from multiple modalities of images are still complex and time-consuming for the doctors. Therefore, the image fusion technique can be applied to merge the salient and complementary information of the multi-modal images into a single image for better perception of both doctors and intelligent systems (Yin et al., 2018; Liu et al., 2019, 2022a,b,c,d; Xu and Ma, 2021; Wang et al., 2022).

In recent years, many methods have been proposed for the task of multi-modal image fusion. Generally, these methods can be divided in two categories, i.e., spatial-domain image fusion methods and transform-domain methods (Liu et al., 2015, 2018; Zhu et al., 2018, 2019; Yin et al., 2019; Xu et al., 2020a; Zhang et al., 2020). Specifically, the spatial-domain image fusion methods first decompose the source images into multiple regions, and then integrate the salient regions together to generate their fusion image (Bai et al., 2015; Liu et al., 2017, 2018, 2020; Zhang et al., 2017). The fusion images of these methods often yield unsatisfactory effect due to their inaccurate segmentation results. The transform-domain methods are more popular in the field of image fusion. These methods first convert the source images into a specific domain, then fuse the salient features in this domain, and finally generates the fusion image by converting the fused features back to the image domain (Liu et al., 2015; Xu et al., 2020a; Zhang et al., 2020). The fusion images of these methods are usually more suitable for human to perceive, but might suffer from the blurring effect (Ma et al., 2019a). Moreover, with the fast development of deep-learning techniques, many deep-learning (mainly convolutional neural network, CNN) based image fusion methods have been proposed (Liu et al., 2017; Li and Wu, 2018; Ma et al., 2019b; Wang et al., 2020; Zhang et al., 2020). These methods adopt CNN to extract the deep convolutional features, then fuse the features of the source images by a feature fusion module, and finally reconstruct the fused features as their fusion images. Even though these deep-learning based methods have achieved great success in the field of image fusion, many of these methods would generate fusion images of low contrast or having other kinds of defects.

Amongst the transform-domain methods, the guided image filter (He et al., 2012) demonstrates to be a state-of-the-art (SOTA) edge-preserving image filter, and has been widely used in the field of image fusion (Li et al., 2013; Gan et al., 2015). But in these methods, the guided image filter is often used to refine the decision map or weight map rather than used to extract salient features due to its relatively weak ability in feature extraction. Therefore, in this study, we aim to improve the feature extraction ability of the guided image filter, and based on our improved guided filter to further develop a multi-modal brain image fusion method.

To be specific, we have developed a local extreme map guided image filter, which consists of a local minimum map guided image filter and a local maximum map guided image filter. The developed local extreme map guided image filter is able to more effectively smooth the input image as compared to the original image filter guided by the input image itself, then the features extracted from the difference image of the smoothed image and input image by our filter will be naturally more salient than those extracted by the original image filter guided by the input image itself. Through extending the local extreme map guided filter to multiple scales, we propose a local extreme map guided image filter based multi-modal brain image fusion method. Specifically, we first apply the local extreme map guided image filter iteratively on each source image to extract their multi-scale bright and dark feature maps. Then, the multi-scale bright feature maps, multi-scale dark feature maps, and the base images of the multi-modal brain images are fused, respectively. Finally, the fused bright feature map, dark feature map, and base image are integrated together to generate our fused brain image.

The contributions of this study can be concluded in three parts:

- We propose a new scheme to improve the feature extraction ability of the guided image filter, i.e., using two guided image filters with a local minimum map and a local maximum map, respectively, as their guidance images. This scheme can be incorporated with various guided filters or other similar filters in pursuit of improving their feature extraction ability.
- Based on the local extreme map guided image filter, we further propose an effective image fusion method for fusing multi-modal brain images. Moreover, the proposed method can be easily adapted to fuse other modalities of images while achieving superior fusion performance.
- Extensive experiments verify that our method performs comparably to or even better than eight SOTA image fusion methods (including three conventional methods and five deep learning based methods) in terms of both qualitative and quantitative evaluations.

The rest of this paper is organized as follows. In Section 2, the constructed local extreme map guided image filter and the proposed multi-modal brain image fusion method are elaborated, respectively. Then, the experimental results and discussions are

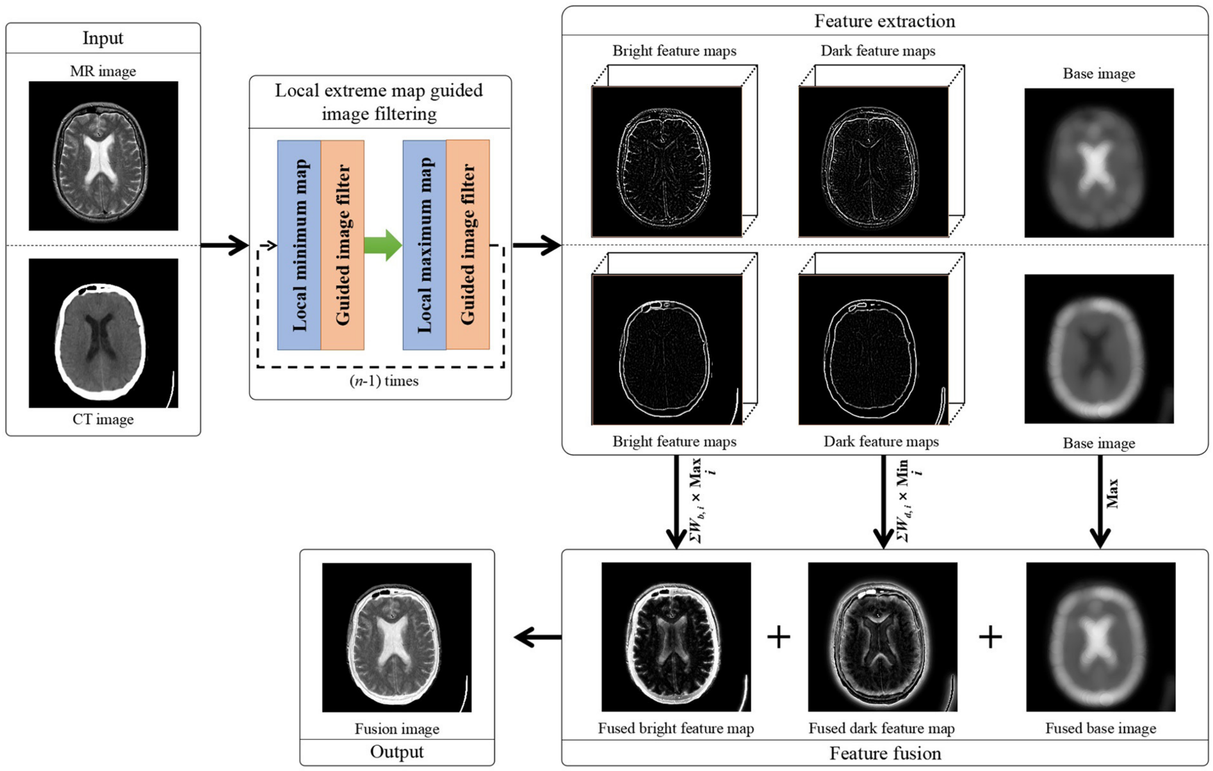


FIGURE 1
Flowchart of our proposed local extreme map guided multi-modal brain image fusion method. (Note that the dark feature maps in this figure have been illustrated as their absolute feature maps in order to properly visualize the dark features).

made in Section 3. Finally, this study is concluded in Section 4.

2. Proposed method

The overall structure of the proposed method is illustrated in Figure 1. The major procedures of the proposed method include: First, the two multi-modal brain image are iteratively smoothed by the local extreme map guided filter, respectively. Then, different scales of bright and dark feature maps are extracted, respectively, from the two multi-modal brain images, and the two smoothed brain images in the last iteration are taken as their base images, respectively. Afterwards, each scale of bright feature maps of the two brain images and each scale of dark feature maps of the two brain images are fused by selecting their elementwise maximum values and their elementwise minimum values, respectively. Further, the fused multi-scale bright feature maps and dark feature maps are integrated as a single bright feature map and a single dark feature map, respectively, and the two base image are fused as their elementwise maximum values as well. Finally, the fusion image is generated by integrating the fused bright feature map, dark feature map, and base image together. In the following subsections, the local extreme map guided image filter and our proposed image fusion are elaborated, respectively.

2.1. Local extreme map guided image filter

In the guided image filter based image fusion methods (Li et al., 2013; Gan et al., 2015), the guided image filter was often used to adjust the decision maps or weight maps for fusing the feature maps of input images rather than directly extracting the salient feature maps from the input images, due to its limited feature extraction ability. Therefore, in this study, we focus on improving the feature extraction ability of the guided image filter by designing appropriate guidance images.

In the official demonstration of guided image filter (He et al., 2012), the input image is smoothed under the guidance of the input image itself to approach the edge preserving effect. However, in this way, the feature map generated by subtracting the filtered image from the input image is usually not salient enough for the task of image fusion. In order to enhance the feature map, we have modified the guidance image from the input image to its local extreme maps, so that the salient features of the input image can be largely suppressed and accordingly these salient features can be effectively extracted from the difference image of the input image and filtered image. The detailed construction method of our local extreme map guided image filter is described as follows.

First, the input image is filtered under the guidance of the local minimum map of the input image as:

$$I_f' = guidedfilter\left(I, I_{\min}, r\right), \tag{1}$$

where *guidedfilter* denotes the guided filter (He et al., 2012). I and $I_{\min}$ are the input image and guidance image, respectively. r denotes the size of the local window for constructing the linear model between input image and guidance image. Moreover, $I_{\min}$ denotes the local minimum image of I. Under the guidance of the local minimum map, the salient bright features could be sufficiently removed from the input image. Specifically, $I_{\min}$ can be solved by the morphological erosion operation as:

$$I_{\min} = imerode\left(I, se\right),$$

where $imerode\left(\cdot\right)$ denotes the morphological erosion operator. se denotes the structuring element of flat-disk shape, radius of which is denoted by k.

Then, I_f' is further filtered under the guidance of its local maximum map as:

$$I_f = guidedfilter\left(I_f', I_{\max}, r\right), \tag{2}$$

where I_f' and $I_{\max}$ are the input image and guidance image, respectively, and $I_{\max}$ denotes the local maximum image of I_f'. Under the guidance of the local maximum map, the salient dark features could be further removed from the finally filtered image. Similar to the solution of $I_{\min}$, $I_{\max}$ can be efficiently solved by the morphological dilation operation as:

$$I_{\max} = imdilate\left(I_f', se\right),$$

where $imdilate\left(\cdot\right)$ denotes the morphological dilation operator.

In order to conveniently introduce the following image fusion method, we denote by $leguidedfilter\left(\cdot\right)$ the function of our constructed local extreme map guided image filter [composed by Equations (1) and (2)], then smoothing an image with the local extreme map guided image filter can be expressed as:

$$I_f = leguidedfilter\left(I, se, r\right), \tag{3}$$

where r and se correspond to the parameters in Equations (1) and (2).

As is known, there exist both bright features and dark features in an image, such as the bright person and the dark roof in Figure 2A. Through sequentially smoothing the input image guided by the local minimum map and local maximum map, respectively, both the salient bright and dark features will be removed from the input image and a well-smoothed image will be obtained. Then, the salient features of the input image can be obtained by subtracting the filtered image I_f from the input image I according to Equation (4), and the positive part of $(I - I_f)$ corresponds to the bright features, and the negative part corresponds to the dark features.

$$\begin{cases} F_b = \max\left(I - I_f, 0\right) \\ F_d = \min\left(I - I_f, 0\right) \end{cases}, \tag{4}$$

where F_b and F_d denote the bright feature map and dark feature map of I, respectively.

A demonstration example of the proposed local extreme map guided image filter performed on an infrared image is illustrated in Figure 2. In this figure, we have compared the smoothed images (see Figures 2B–E), respectively, by the original guided image filter, single local minimum map guided filter, single local maximum map guided filter, and our complete extreme map guided filter, and also compared the feature maps extracted from their difference images with respect to the original infrared image in Figure 2A. It can be seen from Figures 2B–E that the smoothed image by our extreme map guided filter has suppressed more salient features (textural details) compared to those of the original guided filter, single local minimum map guided filter, and single local maximum map guided filter. Accordingly, the salient features (see Figures 2I,M) extracted by our extreme map guided filter are far more than those extracted by the original guided filter, single local minimum map guided filter, and single local maximum map guided filter. Moreover, intensities of our extracted feature maps are much higher than those of feature maps extracted by the other three filters. Overall, the results in this figure suggest that our constructed local extreme map guided filter is able to extract the input image's bright and dark features well and significantly outperforms the original guided filter, single local minimum map guided filter, and single local maximum map guided filter.

Naturally, the local extreme map guided image filter can be extended to multiple scales by iteratively applying the image filter guided with local minimum map and that guided with local maximum map on the input image I according to Equation (5).

$$I_f^i = leguidedfilter\left(I_f^{(i-1)}, se_i, r_i\right), \tag{5}$$

where i denotes the current scale of the guided filter, and i is increased from 1 to n one by one. I_f^i denotes the ith-scale filtered image and especially I_f^0 is the original input image I. se_i and r_i denote the structuring element and size of the local window at the ith scale, respectively.

Accordingly, different scales of bright and dark features can be simultaneously extracted from the difference image of every

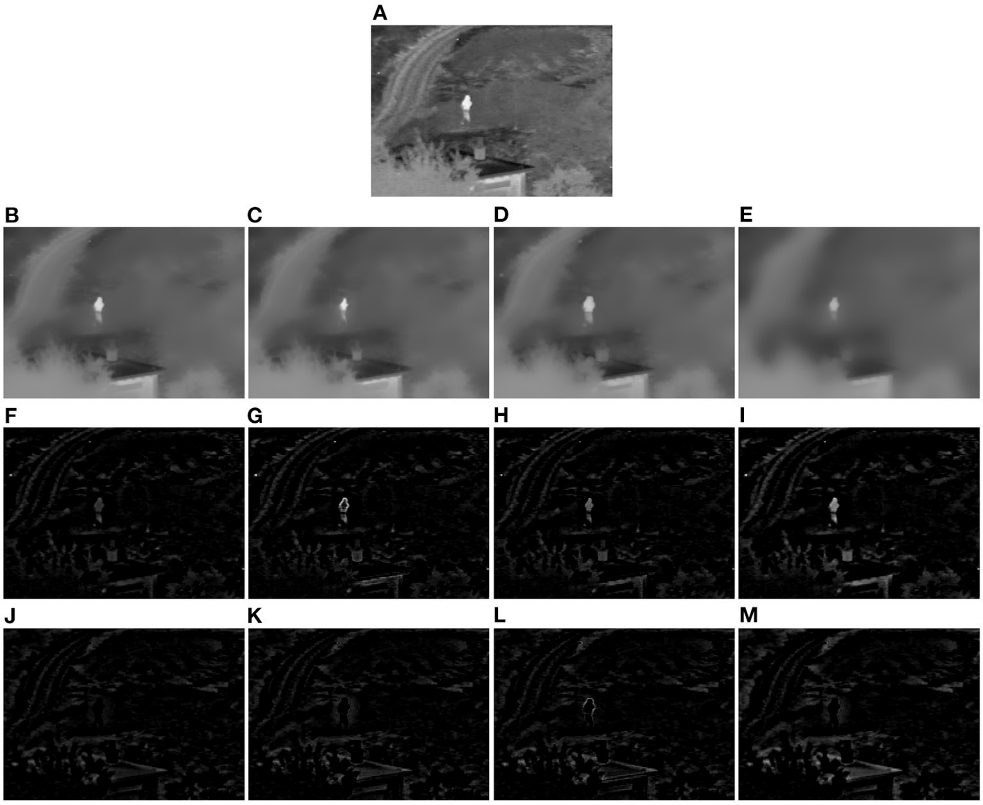


FIGURE 2
Demonstration example of the local extreme map guided image filter (Toet, 2017). **(A)** Original infrared image. **(B)** Image smoothed by the image filter guided by the input image itself. **(C)** Image smoothed by single local minimum map guided image filter. **(D)** Image smoothed by single local maximum map guided image filter. **(E)** Image smoothed by our local extreme map guided image filter. **(F–I)** Bright feature maps extracted from the difference images of **(B–E)** and **(A)**, respectively. **(J–M)** Dark feature maps extracted from the difference images of **(B–E)** and **(A)**, respectively (Note that the dark feature maps in this figure have been illustrated as their absolute feature maps in order to properly visualize the dark features).

two continuously filtered images according to Equation (6).

$$\begin{cases} F_{b,i} = \max\left(I_f^{i-1} - I_f^i, 0\right) \\ F_{d,i} = \min\left(I_f^{i-1} - I_f^i, 0\right) \end{cases}. \tag{6}$$

Finally, the last scale of filtered image is taken as the base image of I, as expressed in Equation (7).

$$I_{base} = I_f^i. \tag{7}$$

2.2. Local extreme map guided image fusion

In this study, we aim to fuse two multi-modal brain images (denoted by I^1 and I^2). According to the feature extraction method introduced in the previous subsection, we can well extract the multi-scale bright feature maps (denoted by $F_{b,i}^j$) and dark feature maps (denoted by $F_{d,i}^j$) of each input image I^j, and simultaneously obtain their base images (denoted by I_{base}^j). j denotes index of the input image, and is ranged from 1 to 2. Then, the detailed procedures for fusing two multi-modal brain images are introduced as follows.

As the high-frequency features of high intensities are usually corresponding to the salient sharp features in the image, thus we fuse each scale of bright feature maps of the two multi-modal brain images by selecting their elementwise-maximum values and fuse each scale of dark feature maps of the two multi-modal images as their elementwise-minimum values as:

$$\begin{cases} F_{b,i}^{fuse} = \max\left(F_{b,i}^1, F_{b,i}^2\right) \\ F_{d,i}^{fuse} = \min\left(F_{d,i}^1, F_{d,i}^2\right) \end{cases}, \tag{8}$$

Like other feature extractors, the proposed local extreme map guided image filter cannot extract the entire bright and dark features from the source images either, thus we have enhanced the fused bright and dark features by multiplying each scale of fused bright feature map and dark feature map by an information-amount related weight. Further, the enhanced bright feature maps and dark feature maps are integrated, respectively. The above two procedures can be mathematically expressed as:

$$\begin{cases} F_b^{fuse} = \sum_{i=1}^{n} w_{b,i} \cdot F_{b,i}^{fuse} \\ F_d^{fuse} = \sum_{i=1}^{n} w_{d,i} \cdot F_{d,i}^{fuse} \end{cases}, \tag{9}$$

where $w_{b,i}$ denotes the weight of the ith scale of bright feature map and $w_{d,i}$ denotes the weight of the ith scale of dark feature map. Generally, the feature map with more information should be assigned to a large weight, thus $w_{b,i}$ and $w_{d,i}$ are set according to the entropy of $F_{b,i}^{fuse}$ and $F_{d,i}^{fuse}$, respectively, as:

$$\begin{cases} w_{b,i} = \frac{e_{b,i}}{\min\limits_{j}(e_{b,j})} \\ w_{d,i} = \frac{e_{d,i}}{\min\limits_{j}(e_{d,j})} \end{cases}, \tag{10}$$

where $e_{b,i}$ denotes the entropy of $F_{b,i}^{fuse}$ and $e_{d,i}$ denotes the entropy of $\left(-F_{d,i}^{fuse}\right)$. In this way, the minimum weight, i.e., weight of feature map with the lowest entropy, will be 1, and weights of other scales of feature maps will all be higher than 1. Accordingly, most scales of bright and dark feature maps will be enhanced to some degree according to their information amount.

As for the low-frequency base images, we directly fuse them by computing their elementwise-maximum values according to Equation (11). In this manner, most basic information of the multi-modal medical images will preserved into the final fusion image.

$$F_{base}^{fuse} = \max\left(I_{base}^{1}, I_{base}^{2}\right). \tag{11}$$

Finally, the fusion image can be generated by combining the fused bright feature map, dark feature map, and base image together as expressed in Equation (12). In this way, our fused image can not only preserve as much as basic information of the multi-modal source images, but also well enhance the salient sharp features of the multi-modal source images.

$$I^{fuse} = F_b^{fuse} + F_d^{fuse} + F_{base}^{fuse}. \tag{12}$$

2.3. Parameter settings

In our method, there are mainly three parameters, including the scale number n, the size of the local window r_i in the guided image filter, and the radius of the structuring element k_i in the morphological erosion and dilation operations. In order to balance the time cost and fusion effect of the multi-modal brain images, n is set to five in this study, i.e., $n = 5$. As for r_i and k_i, we keep them same with each other, i.e., $k_i = r_i$, in in each iteration i of local extreme map guided image filtering. Moreover, in order to effectively extract the salient image features, we set $r_i = 2 \times i + 1$ where i is gradually increased from 1 to n in this study. The extensive experimental results verify the above settings are effective for fusing the multi-modal brain images.

3. Experimental results and discussions

In order to verify the effectiveness of the proposed image fusion method, we have compared it with eight representative image fusion methods on three commonly used multi-modal brain image datasets (Xu and Ma, 2021). The detailed experimental settings, implementation details, results, and discussions are introduced in the following five subsections.

3.1. Experimental settings

At first, we take 30 pairs of commonly used multi-modal brain images from http://www.med.harvard.edu/aanlib as our testing sets, including 10 pairs of CT and MR brain images, 10 pairs of PET and MR images, and 10 pairs of SPECT and MR images. The three used datasets have been shown in Figures 3–5, respectively. In particular, the spatial resolution of the images in the three datasets are all 256 × 256.

Second, we have compared our method with eight SOTA image fusion methods, including the discrete wavelet transform based method (DWT) (Li et al., 1995), the guided-filter based method (GFF) (Li et al., 2013), the Laplacian pyramid and sparse representation base method (LPSR) (Liu et al., 2015), the unified image fusion network (U2Fusion) (Xu et al., 2020a), the GAN based method (DDcGAN) (Ma et al., 2020), the general CNN based image fusion network (IFCNN) (Zhang et al., 2020), the enhanced medical image fusion network (EMFusion) (Xu and Ma, 2021), and the disentangled representation based brain image fusion network (DRBIF) (Wang et al., 2022). Moreover, in order to verify the efficacy of the guidance of local extreme maps, we have also added our method without the guidance of local extreme maps (denoted by LEGFF$_0$) for comparison.

At last, qualitative evaluation heavily depends on the subjective observation which is inaccurate and laborious, thus 11 commonly-used quantitative metrics are further used to objectively compare the 10 methods' performance. The 11 quantitative metrics are spatial frequency (SF) (Li and Yang, 2008), average absolute gradient (AbG), perceptual saliency (PS) (Zhou et al., 2016), standard deviation (STD), entropy (E), Chen-Blum Metric (QCB) (Chen and Blum, 2009), visual information fidelity (VIFF) (Han et al., 2013), edge preservation metric (Qabf) (Xydeas and Petrovic, 2000), gradient similarity metric (QGS) (Liu et al., 2011), weighted structural similarity metric (WSSIM) (Piella and Heijmans, 2003), and multi-scale structural similarity (NSSIM) (Ma et al., 2015). Among

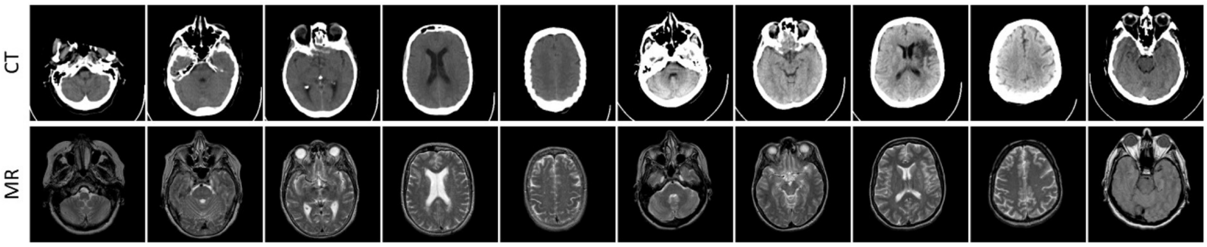


FIGURE 3
Ten pairs of images in the CT-MR image dataset.

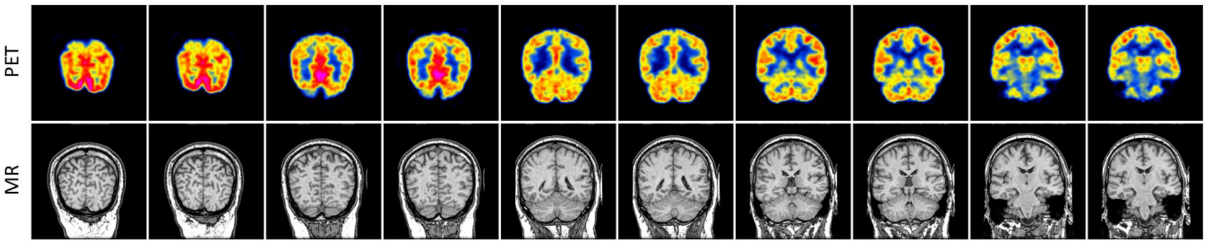


FIGURE 4
Ten pairs of images in the PET-MR image dataset.

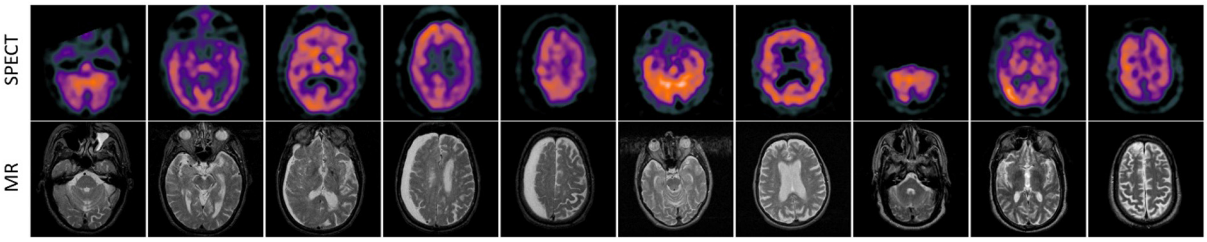


FIGURE 5
Ten pairs of images in the SPECT-MR image dataset.

these metrics, SF, AG, PS, and STD quantify the amount of details reserved in the fusion image, E measures the intensity distribution of the fusion image, QCB measures the amount of the preserved contrast information of the fusion image compared to the source images, VIFF measures the information fidelity of the fusion image with respect to the source images, Qabf measures the amount of the preserved edge information of the fusion image compared to the source images, QGS measures the gradient similarity of the fusion image and the corresponding source images, and WSSIM and NSSIM both measure the structural information of the fusion image preserved from the source images. Overall, the 11 selected metrics can quantitatively evaluate the fusion images of different image fusion methods from various aspects, and the larger values of all the 11 metrics indicate the better performance of the corresponding image fusion method.

3.2. Implementation details

Among the 10 comparison methods, IFCNN and DRBIF can be directly used to fuse color images, and the other eight fusion methods can only fuse gray-scale images directly. Thus, DWT, GFF, LPSR, U2Fusion, DDcGAN, EMFusion, LEGFF$_0$, and our method can be directly applied to fuse the pair of gray-scale CT and MR images in the CT-MR image dataset. As for fusing images in the PET-MR and SPECT-MR image datasets, the color image (PET or SPECT image) is first transformed from the RGB color space to the YCbCr color space. Then, these

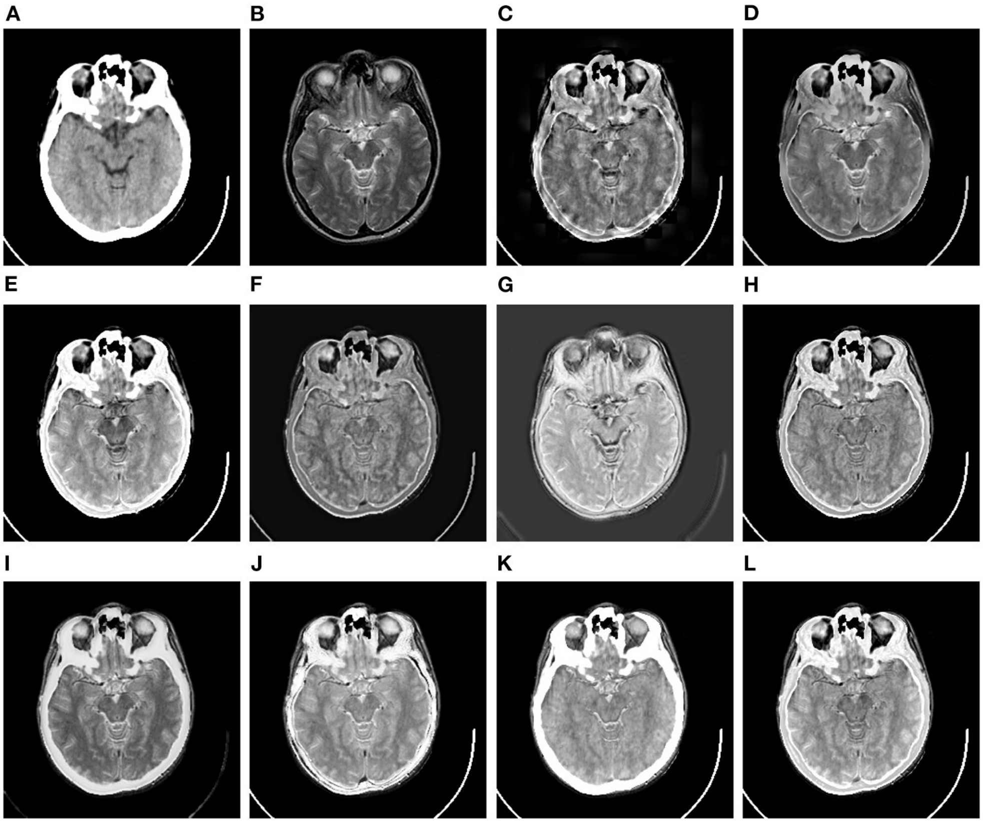


FIGURE 6
Comparison example on the CT-MR image dataset. **(A,B)** are the original CT image and MR image, respectively. **(C–L)** are the fusion results of DWT, GFF, LPSR, U2Fusion, DDcGAN, IFCNN, EMFusion, DRBIF, LEGFF$_0$, and our method, respectively.

eight methods fuse the Y channel of the color image and the gray-scale MR image together. Finally, the fused color image is generated by concatenating the fused gray-scale image and Cb and Cr channels of the original color image, and transforming the fused image in the YCbCr color space back to the RGB color space. Moreover, most quantitative metrics are designed to quantify the quality of gray-scale fusion images. Thus, during computing the quantitative metric values on the PET-MR and SPECT-MR image datasets, we covert the color source image and the corresponding color fusion image to the YCbCr color space and take their Y channels to compute the metric value of this color fusion image. Finally, code of our proposed method will be released on https://github.com/uzeful/LEGFF.

3.3. Qualitative evaluation results

In this subsection, the 10 image fusion methods are evaluated by the qualitative method, i.e., comparing their fusion results through visual observation. Specifically, we have shown three comparison examples of the 10 image fusion methods in Figures 6–8, respectively.

Figure 6 shows a set of fusion results of the 10 image fusion methods on the CT-MR image dataset. It can be seen from Figure 6C that the fusion image of DWT demonstrates severe blocking effect around the head. Figures 6D,F reflect that the fusion images of GFF and U2Fusion are of relatively low contrast. Figure 6G shows that the background of the DDcGAN's fusion image becomes gray and leads to low-contrast effect. It can be seen from Figures 6I,K that EMFusion and LEGFF$_0$ fail to integrate the textures of soft tissues in the skull region of the original MR image into their fusion images. Figure 6J shows that DRBIF fails to integrate several parts of skull region of the original CT image into its fusion image. Finally, the fusion images of LPSR, IFCNN, and our method in Figures 6E,H,J achieve the best visual effect among all the fusion images, i.e., having better contrast and integrating the salient textures of the original MR image and CT image into their fusion images.

Figure 7 shows a set of fusion results of the 10 image fusion methods on the PET-MR image dataset. It can be seen from

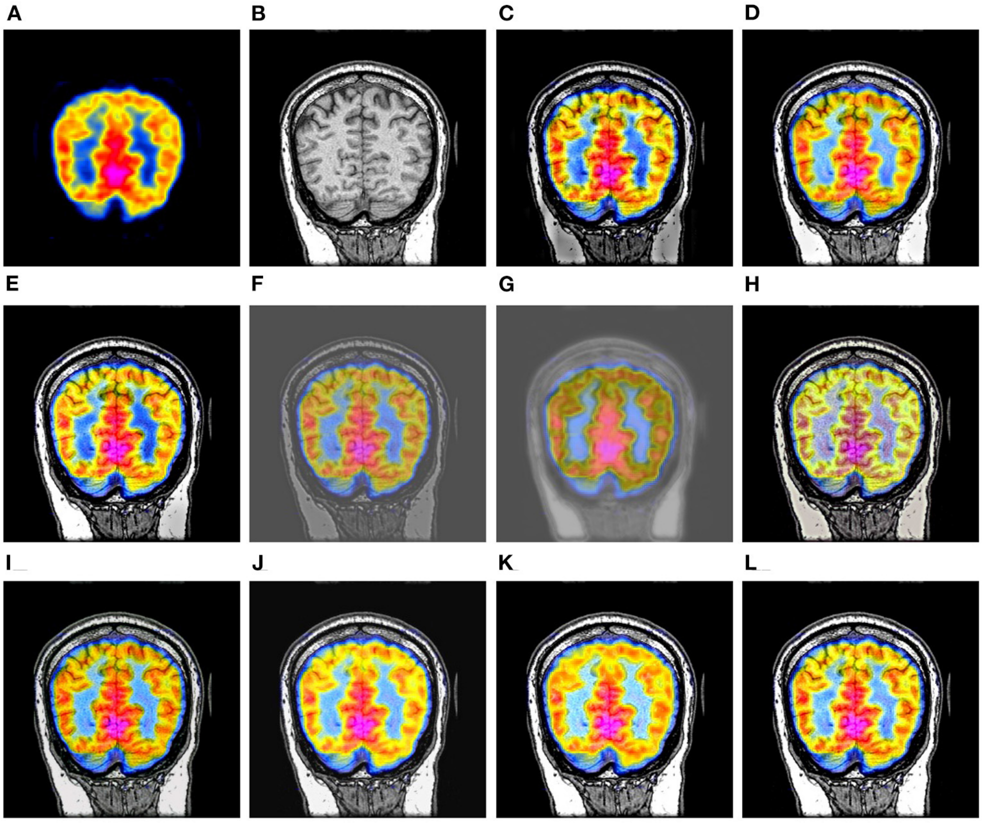


FIGURE 7
Comparison example on the PET-MR image dataset. **(A,B)** are the original PET image and MR image, respectively. **(C–L)** are the fusion results of DWT, GFF, LPSR, U2Fusion, DDcGAN, IFCNN, EMFusion, DRBIF, LEGFF$_0$, and our method, respectively.

Figure 7C that the intensities of the bottom part of DWT's fusion image are significantly lower than that of the original MR image in Figure 7B. Figure 7E shows that the intensities of the bottom right of LPSR's fusion image are slightly lower than that of the original MR image in Figure 7B. The fusion results of U2Fusion and DDcGAN in Figures 7F,G have much lower contrast than those of other methods. Figure 7H shows that the color style of IFCNN's fusion image is significantly changed as compared to that of the original PET image in Figure 7A. Figures 7J,K show that DRBIF and LEGFF$_0$ fail to integrate some dark features of the MR image into their fusion images. Overall, the fusion images of GFF, EMFusion, and our method in Figures 7D,I,L integrate most salient features of the original PET and MR images into their fusion images, but contrast of EMFusion's fusion image is a little lower than that of GFF's fusion image and ours.

Figure 8 shows a set of fusion results of the 10 image fusion methods on the SPECT-MR image dataset. We can see from Figure 8C that the fusion image of DWT loses a few textures around the center regions of the two eyes. It can be seen from Figures 8D,E that GFF and LPSR only integrate a few details of the bottom skull region of the original MR image into their fusion images. The fusion image of U2Fusion and DDcGAN in Figures 8F,G still have the defect of lower contrast and gray background. The fusion image of IFCNN in Figure 8H is of low contrast compared to the original SPECT and MR images in Figures 8A,B. Figure 8J shows that the color style of DRBIF's fusion image is significantly different from that of the original SPECT image in Figure 8A and DRBIF fails to integrate a few bright features of the original MR image into its fusion image due to its relatively high intensity. Figure 8K shows that LEGFF$_0$ fails to integrate many bright features of the original MR image into its fusion image. Overall, the fusion images of DWT, EMFusion, and our method in Figures 8C,I,L exhibit the best visual effects among all the fusion images, but the salient features integrated in our fusion image are more complete than those integrated in the fusion images of DWT and EMFusion.

The three comparison examples could verify that the proposed method can effectively fuse the salient bright and dark features of the multi-modal brain images into a comprehensive fusion image, and outperforms the eight SOTA image fusion methods according to the visual comparison results. Moreover,

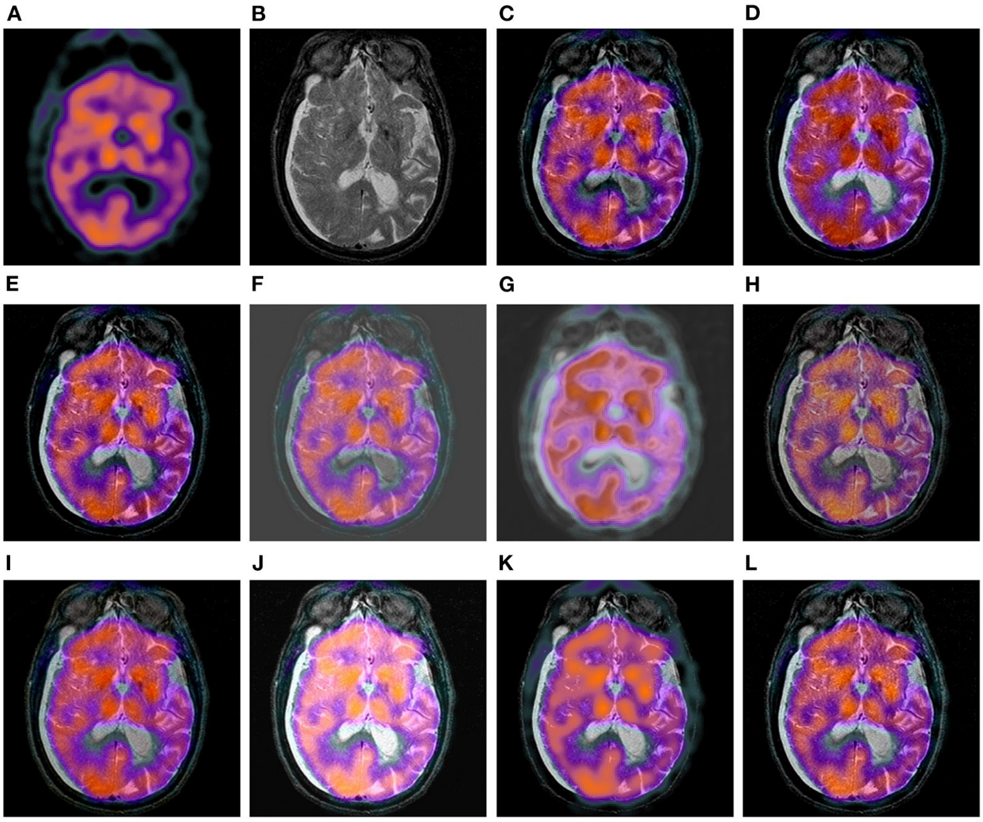


FIGURE 8
Comparison example on the SPECT-MR image dataset. **(A,B)** are the original SPECT image and MR image, respectively. **(C–L)** are the fusion results of DWT, GFF, LPSR, U2Fusion, DDcGAN, IFCNN, EMFusion, DRBIF, LEGFF$_0$, and our method, respectively.

through visually comparing the fusion results of LEGFF$_0$ and our method, it could be verified that the incorporation of the local extreme map guidance is critical for improving the feature extraction ability and feature fusion ability of the guided image filter.

3.4. Quantitative evaluation results

The quantitative metric values of the eight image fusion methods are first calculated according to their fusion results on each dataset, then the average metric values of the eight methods on each dataset are listed in Tables 1–3, respectively. In each table, the values in the **bold**, underline, and *italic* fonts indicate the best, second-best and third-best results, respectively.

It can be seen from Table 1 that the proposed method has achieved the best performance on two metrics (i.e., VIFF and NSSIM), obtained second-best performance on three metrics (i.e., PS, QGS, and WSSIM), ranked the third place on the STD metric on the CT-MR image dataset. To be specific, the largest VIFF and NSSIM values and second-largest QGS and WSSIM values of our method suggest that our fusion images have preserved relatively more edge and structural information from the original CT and MR images than the fusion images of other methods. The second-largest PS value and the third-largest STD value of our method indicate that the fusion images of our method have slightly more textural details than those generated by the other eight comparison methods. Since in our method the base images of the source images are fused as their elementwise-maximum values, thus intensity distribution of our fusion images might be not that uniform along the gray-scale space leading to relatively lower E and QCB values. Besides our method, LPSR has achieved the best performance on three metrics (i.e., SF, PS, and QCB) and the second performance on two metrics (i.e., STD and MSSIM) and IFCNN has achieved the best performance on three metrics (i.e., AbG, QGS, and WSSIM) and the second performance on three metrics (i.e., SF, QCB, and Qabf). Overall, consistent to the qualitative comparison results, the quantitative evaluation results in Table 1 shows LPSR, IFCNN, and our method perform slightly better than the other seven methods on fusing the CT and MR images.

TABLE 1 Quantitative evaluation results on the CT-MRI dataset.

Metrics	Methods									
	DWT	**GFF**	**LPSR**	**U2Fusion**	**DDcGAN**	**IFCNN**	**EMFusion**	**DRBIF**	**LEGFF0**	**Ours**
SF	*33.6948*	28.4859	**35.7214**	22.6498	21.2063	34.0565	21.8423	29.4565	31.3377	32.8535
AbG	16.6486	12.8718	14.7253	12.7615	13.8349	**16.7553**	11.7355	14.5474	12.8055	*15.1338*
STD	69.4181	62.5530	86.7440	55.0136	63.6475	76.6713	76.1188	83.2861	**89.8708**	*84.8222*
QPS	41.6377	35.8390	**47.8699**	30.9477	29.9360	43.2325	35.6807	42.0657	*44.8173*	44.9017
E	**5.0467**	4.3135	3.7990	4.6787	4.9987	4.2717	*4.5214*	4.3620	3.1752	4.2705
QCB	0.5436	0.6598	**0.7074**	0.2906	0.1683	0.6843	*0.6699*	0.6562	0.6603	0.6342
VIFF	0.3539	0.2733	0.4280	0.3185	0.2002	0.4256	0.3992	*0.4318*	0.4539	**0.4800**
Qabf	0.5538	0.7319	0.7394	0.6503	0.5934	0.7598	0.7247	*0.7461*	**0.7821**	0.7170
QGS	0.8796	0.8521	*0.9000*	0.8088	0.7776	**0.9135**	0.8053	0.8580	0.8766	0.9054
WSSIM	0.7113	0.8245	0.8178	0.3525	0.1931	**0.8456**	0.8088	*0.8266*	0.8139	0.8415
MSSIM	0.8731	0.8460	0.9395	0.8736	0.6584	*0.9384*	0.8819	0.9095	0.8926	**0.9505**

TABLE 2 Quantitative evaluation results on the PET-MRI dataset.

Metrics	Methods									
	DWT	**GFF**	**LPSR**	**U2Fusion**	**DDcGAN**	**IFCNN**	**EMFusion**	**DRBIF**	**LEGFF0**	**Ours**
SF	33.5860	35.0668	34.5411	10.0922	6.2573	33.4545	29.6977	28.2939	*34.7476*	**38.0866**
AbG	22.2784	*22.6448*	22.4732	7.2796	4.5950	22.9097	20.2215	18.6718	21.5265	**25.0111**
STD	68.8455	*74.7923*	73.3129	27.8267	24.9354	72.8165	68.6928	72.6768	**80.3887**	79.2849
QPS	36.7423	*40.6313*	39.2959	13.9280	10.4126	38.0456	34.6725	35.4353	41.2365	**42.5936**
E	4.8620	4.9216	4.8230	4.6586	4.9882	*5.1515*	5.3884	**5.4553**	4.5049	4.8948
QCB	0.5747	0.5968	0.6075	0.3435	0.2272	*0.5997*	0.5989	0.3475	0.5683	**0.6145**
VIFF	0.4784	0.4248	*0.4926*	0.1779	0.0446	0.4969	0.4018	0.4702	0.4403	**0.5175**
Qabf	0.6443	*0.7340*	0.7070	0.4017	0.3277	0.7118	0.7171	0.6761	0.7346	**0.7392**
QGS	0.8905	0.9256	0.9123	0.7595	0.7164	*0.9142*	0.9015	0.8954	0.9128	**0.9344**
WSSIM	0.6754	0.7170	0.7027	0.3623	0.1611	0.7098	0.7052	0.6569	0.6708	**0.7201**
MSSIM	0.9238	0.9003	**0.9463**	0.6364	0.3814	*0.9387*	0.8983	0.9216	0.8817	0.9436

Table 2 shows that our method has achieved the best performance on eight metrics (i.e., SF, AbG, PS, QCB, VIFF, Qabf, QGS, and WSSIM) and the second-best performance on two metrics (i.e., STD and MSSIM). As addressed previously, the E metric value of our method is relatively lower than those of other methods, due to our usage of the elementwise-maximum strategy for fusing the base images. Overall, the quantitative evaluation results on the PET-MR image dataset suggest our method significantly outperforms the other nine methods by a large margin in particular on fusing the PET and MR images. This conclusion is also consistent to the visual comparison results from Figure 7.

Finally, it can be seen from Table 3 that our method has ranked the first place on six metrics (i.e., SF, AbG, QCB, Qabf, WSSIM, and MSSIM), ranked the second place on the QPS, VIFF, and QGS metrics, and ranked the third place on the STD metric. Besides, DRBIF have obtained the best performance on five metrics (i.e., STD, QPS, E, VIFF, and QGS) and the second place on three metrics (i.e., SF, AbG, and MSSIM). These results suggest the fusion images of our method and DRBIF have more textural details and persevered more structural information from the original SPECT and MR images compared to those of the other eight methods. Moreover, the quantitative results in Tables 1–3 indicate that our method with the local extreme map guidance significantly outperforms that without the local extreme map guidance. Thus, the incorporation of the local extreme map guidance is effective for fusing the multi-modal medical images.

Besides, in order to test the efficiency of our proposed method, we have compared the average time cost of each method on the SPECT-MRI image dataset. All methods were evaluated on the same computation platform with Intel Core i7-11700K CPU and NVIDIA GeForce RTX 3090 GPU. The evaluation results have been listed in Table 4. It can be seen from Table 4

TABLE 3 Quantitative evaluation results on the SPECT-MRI dataset.

Metrics	Methods									
	DWT	GFF	LPSR	U2Fusion	DDcGAN	IFCNN	EMFusion	DRBIF	LEGFF0	Ours
SF	16.4700	16.4841	*16.6859*	8.0194	6.4764	15.6082	13.2987	17.1988	14.7571	**18.5201**
AbG	11.4319	11.0090	11.3386	5.7527	4.9230	*11.7558*	9.6794	12.7670	9.0886	**13.4389**
STD	40.8780	46.4581	47.2084	24.2069	39.5366	42.6516	42.9472	**64.6827**	49.2280	*48.9651*
QPS	21.4775	23.2906	*23.4398*	12.5464	12.7030	20.3133	18.7972	**26.1537**	21.7281	24.3273
E	4.4663	4.2638	4.4474	4.5896	5.6676	5.1628	*5.2005*	**5.7996**	4.3966	4.9760
QCB	0.5633	0.5926	*0.5858*	0.3456	0.2249	0.5838	0.5807	0.3706	0.5380	**0.6000**
VIFF	0.4749	0.4230	*0.4873*	0.2327	0.2141	0.4814	0.4411	**0.7241**	0.4531	0.5646
Qabf	0.5656	0.6269	0.6137	0.3996	0.2824	0.6860	0.6461	*0.6501*	0.5976	**0.6866**
QGS	0.9194	0.9332	*0.9368*	0.8646	0.8368	0.9353	0.9184	**0.9461**	0.8984	0.9411
WSSIM	0.6371	0.6509	0.6546	0.4302	0.2113	*0.6512*	0.6425	0.5305	0.5936	**0.6548**
MSSIM	0.9299	0.8973	0.9411	0.7797	0.5437	*0.9475*	0.9372	0.9496	0.8949	**0.9556**

TABLE 4 Time cost comparison.

Methods	DWT	GFF	LPSR	U2Fusion	DDcGAN	IFCNN	EMFusion	DRBIF	LEGFF0	Ours
Time costs	0.0077	0.1035	**0.0021**	0.3019	0.7935	*0.0391*	0.1323	0.0997	0.0401	0.1230

that LPSR and DWT run much faster than the other methods. As for our method, it costs about 0.1230 s to fuse a pair of multi-modal brain images, and it is slightly faster than three deep learning based methods including U2Fusion, DDcGAN, and EMFusion. Therefore, in term of time cost evaluation, the proposed method is relatively time-efficient as compared to the other nine comparison methods. Moreover, in order to verify the generalization ability of our method, we have apply it to fuse other modalities of images, including the multi-focus images, infrared and visual images, multi-exposure images, and green-fluorescent and phase-contrast protein images. Figure 9 shows that our method can well integrate the salient features of each pair of source images into the corresponding fusion images. Thus, the good fusion results in Figure 9 can verify the good generalization ability of our method for fusing other modalities of images. Overall, both qualitative and quantitative evaluation results indicate that our method performs comparably to or even better than eight SOTA image fusion methods and owns good generalization ability.

3.5. Limitations and future prospects

Even though the experimental results validate the advantages of our image fusion method, there still exist several limitations in our method. At first, our local extreme map guided image filter is constructed on the basis of the guided image filter, thus the feature extraction ability of our filter will be inevitably impacted by that of the original guided image filter. Second, compared to LEGFF$_0$ (which uses the original guided filter solely for feature extraction and image fusion), the time cost of our image fusion method increases by a large margin due to iterative calculation of local extreme maps. In future, with the development of guided image filter, performance of our image fusion method can be further boosted by incorporating more advanced guided image filter. Moreover, integrating the local extreme map guidance and the deep-learning frameworks is another way to simultaneously improve the performance and efficiency of the local extreme map guided image fusion methods. Finally, the proposed image fusion method does not contain the image denoising and registration procedures, thus before applying our method in the clinical scenarios the pair of multi-modal source images should be denoised and aligned first.

4. Conclusion

In this study, we propose an effective multi-modal brain image fusion method based on a local extreme map guided image filter. The local extreme map guided image filter can well smooth the image, thus it can further be used to extract the salient bright and dark features of the image. By iteratively applying this local extreme map guided image filter, our method is able to extract multiple scales of bright and dark features from the multi-modal brain images, and integrate these salient features into one informative fusion image. Extensive experimental results suggest that the proposed method outperforms eight SOTA image fusion methods from

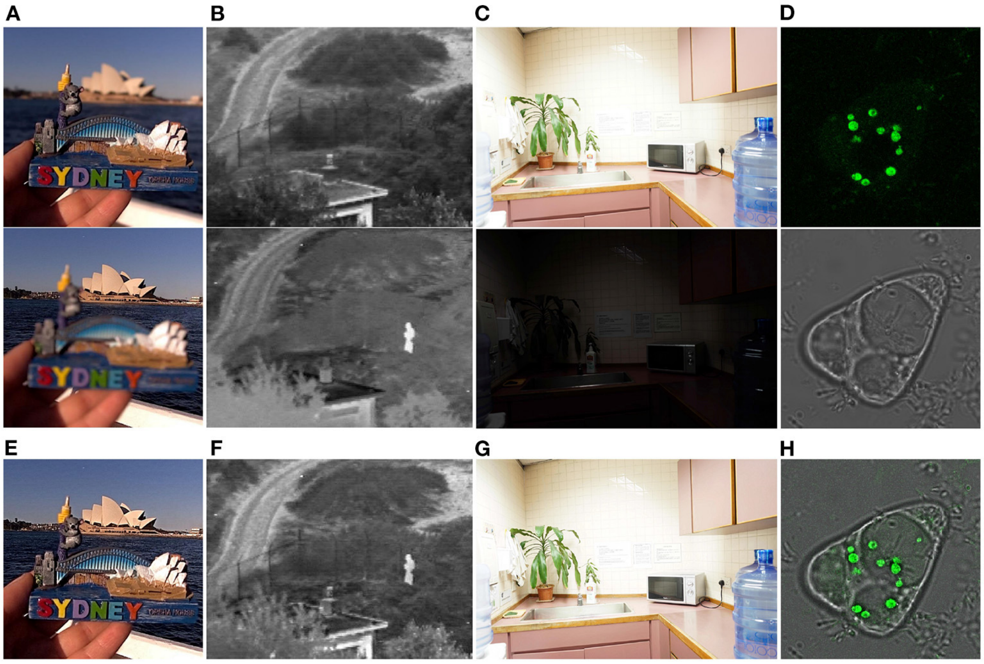


FIGURE 9
Our fusion results on other modalities of images. **(A)** Shows a pair of multi-focus images (Nejati et al., 2015). **(B)** Shows a pair of visual and infrared images (Toet, 2017). **(C)** Shows a pair of over- and under-exposed images (Xu et al., 2020b). **(D)** Shows a pair of green-fluorescent and phase-contrast protein images (Tang et al., 2021). **(E–H)** are the fusion results of **(A–D)**, respectively.

both qualitative and quantitative aspects and it demonstrates very good generalization ability to fuse other modalities of images. Therefore, the proposed method exhibits great possibility to apply in the real clinical scenarios.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://www.med.harvard.edu/aanlib. Code and used images are available at https://github.com/uzeful/LEGFF.

Author contributions

YZ and QZ designed the study. YZ, WX, SZ, and JS performed data analysis. YZ wrote the manuscript. QZ, LZ, XB, and RW revised the manuscript. All authors contributed to the article and approved the final submitted version.

Funding

This study was supported in part by the National Key Research and Development Program of China under Grant No. 2019YFB1311301, in part by the National Natural Science Foundation of China under Grant Nos. 62171017, 61871248, 62132002, 62173005, and 61976017, in part by the Beijing Natural Science Foundation under Grant No. 4202056, and in part by China Postdoctoral Science Foundation under Grant No. 2021M690297.

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

Bai, X., Zhang, Y., Zhou, F., and Xue, B. (2015). Quadtree-based multi-focus image fusion using a weighted focus-measure. *Inform. Fus.* 22, 105–118. doi: 10.1016/j.inffus.2014.05.003

Chen, Y., and Blum, R. S. (2009). A new automated quality assessment algorithm for image fusion. *Image Vis. Comput.* 27, 1421–1432. doi: 10.1016/j.imavis.2007.12.002

Gan, W., Wu, X., Wu, W., Yang, X., Ren, C., He, X., et al. (2015). Infrared and visible image fusion with the use of multi-scale edge-preserving decomposition and guided image filter. *Infrared Phys. Technol.* 72, 37–51. doi: 10.1016/j.infrared.2015.07.003

Han, Y., Cai, Y., Cao, Y., and Xu, X. (2013). A new image fusion performance metric based on visual information fidelity. *Inform. Fus.* 14, 127–135. doi: 10.1016/j.inffus.2011.08.002

He, K., Sun, J., and Tang, X. (2012). Guided image filtering. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1397–1409. doi: 10.1109/TPAMI.2012.213

Li, H., Manjunath, B., and Mitra, S. K. (1995). Multisensor image fusion using the wavelet transform. *Graph. Models Image Process.* 57, 235–245. doi: 10.1006/gmip.1995.1022

Li, H., and Wu, X.-J. (2018). DenseFuse: a fusion approach to infrared and visible images. *IEEE Trans. Image Process.* 28, 2614–2623. doi: 10.1109/TIP.2018.2887342

Li, S., Kang, X., and Hu, J. (2013). Image fusion with guided filtering. *IEEE Trans. Image Process.* 22, 2864–2875. doi: 10.1109/TIP.2013.2244222

Li, S., and Yang, B. (2008). Multifocus image fusion using region segmentation and spatial frequency. *Image Vis. Comput.* 26, 971–979. doi: 10.1016/j.imavis.2007.10.012

Liu, Y., Chen, X., Peng, H., and Wang, Z. (2017). Multi-focus image fusion with a deep convolutional neural network. *Inform. Fus.* 36, 191–207. doi: 10.1007/978-3-319-42999-1

Liu, Y., Chen, X., Wang, Z., Wang, Z. J., Ward, R. K., and Wang, X. (2018). Deep learning for pixel-level image fusion: recent advances and future prospects. *Inform. Fus.* 42, 158–173. doi: 10.1016/j.inffus.2017.10.007

Liu, Y., Chen, X., Ward, R. K., and Wang, Z. J. (2019). Medical image fusion via convolutional sparsity based morphological component analysis. *IEEE Signal Process. Lett.* 26, 485–489. doi: 10.1109/LSP.2019.2895749

Liu, Y., Liu, S., and Wang, Z. (2015). A general framework for image fusion based on multi-scale transform and sparse representation. *Inform. Fus.* 24, 147–164. doi: 10.1016/j.inffus.2014.09.004

Liu, Y., Mu, F., Shi, Y., and Chen, X. (2022a). SF-Net: a multi-task model for brain tumor segmentation in multimodal MRI via image fusion. *IEEE Signal Process. Lett.* 29, 1799–1803. doi: 10.1109/LSP.2022.3198594

Liu, Y., Mu, F., Shi, Y., Cheng, J., Li, C., and Chen, X. (2022b). Brain tumor segmentation in multimodal MRI via pixel-level and feature-level image fusion. *Front. Neurosci.* 16:1000587. doi: 10.3389/fnins.2022.1000587

Liu, Y., Shi, Y., Mu, F., Cheng, J., and Chen, X. (2022c). Glioma segmentation-oriented multi-modal mr image fusion with adversarial learning. *IEEE/CAA J. Autom. Sin.* 9, 1528–1531. doi: 10.1109/JAS.2022.105770

Liu, Y., Shi, Y., Mu, F., Cheng, J., Li, C., and Chen, X. (2022d). Multimodal MRI volumetric data fusion with convolutional neural networks. *IEEE Trans. Instrum. Meas.* 71, 1–15. doi: 10.1109/TIM.2022.3184360

Liu, Y., Wang, L., Cheng, J., Li, C., and Chen, X. (2020). Multi-focus image fusion: a survey of the state of the art. *Inform. Fus.* 64, 71–91. doi: 10.1016/j.inffus.2020.06.013

Liu, Z., Blasch, E., Xue, Z., Zhao, J., Laganiere, R., and Wu, W. (2011). Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 94–109. doi: 10.1109/TPAMI.2011.109

Ma, J., Ma, Y., and Li, C. (2019a). Infrared and visible image fusion methods and applications: a survey. *Inform. Fus.* 45, 153–178. doi: 10.1016/j.inffus.2018.02.004

Ma, J., Xu, H., Jiang, J., Mei, X., and Zhang, X.-P. (2020). DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans. Image Process.* 29, 4980–4995. doi: 10.1109/TIP.2020.2977573

Ma, J., Yu, W., Liang, P., Li, C., and Jiang, J. (2019b). FusionGAN: a generative adversarial network for infrared and visible image fusion. *Inform. Fus.* 48, 11–26. doi: 10.1016/j.inffus.2018.09.004

Ma, K., Zeng, K., and Wang, Z. (2015). Perceptual quality assessment for multi-exposure image fusion. *IEEE Trans. Image Process.* 24, 3345–3356. doi: 10.1109/TIP.2015.2442920

Nejati, M., Samavi, S., and Shirani, S. (2015). Multi-focus image fusion using dictionary-based sparse representation. *Infm. Fusion.* 25, 72–84. doi: 10.1016/j.inffus.2014.10.004

Piella, G., and Heijmans, H. (2003). "A new quality metric for image fusion," in *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, Vol. 3. IEEE (Barcelona), 111–173. doi: 10.1109/ICIP.2003.1247209

Tang, W., Liu, Y., Cheng, J., Li, C., and Chen, X. (2021). Green fluorescent protein and phase contrast image fusion via detail preserving cross network. *IEEE Trans. Comput. Imag.* 7, 584–597. doi: 10.1109/TCI.2021.3083965

Toet, A. (2017). The TNO multiband image data collection. *Data Breif.* 15, 249–251. doi: 10.1016/j.dib.2017.09.038

Wang, A., Luo, X., Zhang, Z., and Wu, X.-J. (2022). A disentangled representation based brain image fusion via group lasso penalty. *Front. Neurosci.* 16:937861. doi: 10.3389/fnins.2022.937861

Wang, K., Zheng, M., Wei, H., Qi, G., and Li, Y. (2020). Multi-modality medical image fusion using convolutional neural network and contrast pyramid. *Sensors* 20, 2169. doi: 10.3390/s20082169

Xu, H., and Ma, J. (2021). EMFusion: an unsupervised enhanced medical image fusion network. *Inform. Fus.* 76, 177–186. doi: 10.1016/j.inffus.2021.06.001

Xu, H., Ma, J., Jiang, J., Guo, X., and Ling, H. (2020a). U2Fusion: a unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 502–518. doi: 10.1109/TPAMI.2020.3012548

Xu, H., Ma, J., and Zhang, X. -P. (2020b). MEF-GAN: Multi-exposure image fusion via generative adversarial networks. *IEEE Trans. Imag. Process.* 29, 7203–7316. doi: 10.1109/TIP.2020.2999855

Xydeas, C. S., and Petrovic, V. S. (2000). "Objective pixel-level image fusion performance measure," in *Proceedings SPIE 4051, Sensor Fusion: Architectures, Algorithms, and Applications IV*, (Orlando, FL), 89–98.

Yin, L., Zheng, M., Qi, G., Zhu, Z., Jin, F., and Sim, J. (2019). A novel image fusion framework based on sparse representation and pulse coupled neural network. *IEEE Access* 7, 98290–98305. doi: 10.1109/ACCESS.2019.2929303

Yin, M., Liu, X., Liu, Y., and Chen, X. (2018). Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain. *IEEE Trans. Instrum. Meas.* 68, 49–64. doi: 10.1109/TIM.2018.2838778

Zhang, Y., Bai, X., and Wang, T. (2017). Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure. *Inform. Fus.* 35, 81–101. doi: 10.1016/j.inffus.2016.09.006

Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., and Zhang, L. (2020). IFCNN: a general image fusion framework based on convolutional neural network. *Inform. Fus.* 54, 99–118. doi: 10.1016/j.inffus.2019.07.011

Zhou, Z., Dong, M., Xie, X., and Gao, Z. (2016). Fusion of infrared and visible images for night-vision context enhancement. *Appl. Opt.* 55, 6480–6490. doi: 10.1364/AO.55.006480

Zhu, Z., Yin, H., Chai, Y., Li, Y., and Qi, G. (2018). A novel multi-modality image fusion method based on image decomposition and sparse representation. *Inform. Sci.* 432, 516–529. doi: 10.1016/j.ins.2017.09.010

Zhu, Z., Zheng, M., Qi, G., Wang, D., and Xiang, Y. (2019). A phase congruency and local laplacian energy based multi-modality medical image fusion method in nsct domain. *IEEE Access* 7, 20811–20824. doi: 10.1109/ACCESS.2019.2898111

TYPE Original Research
PUBLISHED 03 November 2022
DOI 10.3389/fnins.2022.1019572

OPEN ACCESS

EDITED BY
Wei Wei,
The First Affiliated Hospital of University of Science and Technology of China Anhui Provincial Hospital, China

REVIEWED BY
Juan Cheng,
Hefei University of Technology, China
Aiping Liu,
University of Science and Technology of China, China

*CORRESPONDENCE
Tommi Mononen
tommi.mononen@helsinki.fi

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 15 August 2022
ACCEPTED 06 October 2022
PUBLISHED 03 November 2022

CITATION
Mononen T, Kujala J, Liljeström M, Leppäaho E, Kaski S and Salmelin R (2022) The relationship between electrophysiological and hemodynamic measures of neural activity varies across picture naming tasks: A multimodal magnetoencephalography-functional magnetic resonance imaging study. *Front. Neurosci.* 16:1019572. doi: 10.3389/fnins.2022.1019572

The relationship between electrophysiological and hemodynamic measures of neural activity varies across picture naming tasks: A multimodal magnetoencephalography-functional magnetic resonance imaging study

Tommi Mononen[1,2,3,4]*, Jan Kujala[1,5], Mia Liljeström[1,2,6], Eemeli Leppäaho[3], Samuel Kaski[3] and Riitta Salmelin[1,2]

[1]Department of Neuroscience and Biomedical Engineering, Aalto University School of Science, Espoo, Finland, [2]Aalto NeuroImaging, Aalto University, Espoo, Finland, [3]Department of Computer Science, Aalto University, Espoo, Finland, [4]Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland, [5]Department of Psychology, University of Jyväskylä, Jyväskylä, Finland, [6]BioMag Laboratory, Helsinki University Hospital, Helsinki, Finland

Different neuroimaging methods can yield different views of task-dependent neural engagement. Studies examining the relationship between electromagnetic and hemodynamic measures have revealed correlated patterns across brain regions but the role of the applied stimulation or experimental tasks in these correlation patterns is still poorly understood. Here, we evaluated the across-tasks variability of MEG-fMRI relationship using data recorded during three distinct naming tasks (naming objects and actions from action images, and objects from object images), from the same set of participants. Our results demonstrate that the MEG-fMRI correlation pattern varies according to the performed task, and that this variability shows distinct spectral profiles across brain regions. Notably, analysis of the MEG data alone did not reveal modulations across the examined tasks in the time-frequency windows emerging from the MEG-fMRI correlation analysis. Our results suggest that the electromagnetic-hemodynamic correlation could serve as a more sensitive proxy for task-dependent neural engagement in cognitive tasks than isolated within-modality measures.

KEYWORDS

multimodal data, data fusion, fMRI, MEG, picture naming, clustering, correlation patterns

Introduction

Functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG) are widely used non-invasive neuroimaging methods that both have their own strengths. FMRI measures hemodynamic modulations resulting from multiple neural and vascular phenomena, and yields an accurate three-dimensional map of brain activity with a good spatial resolution. However, its temporal precision is modest due to the sluggishness of the hemodynamic response. MEG measures the magnetic field elicited by electric activity, and possesses a temporal resolution in the millisecond scale. The distribution of cortical activity yielded by MEG is relatively smooth.

Thus, depending on the importance of the spatial and temporal aspects of brain activity to the research question at hand, either method may be the preferred option. It has also been proposed that combining MEG and fMRI would allow one to obtain accurate spatiotemporal maps of brain activity (Dale et al., 2000; Henson et al., 2010). Such approaches have, for example, utilized fMRI to spatially constrain the MEG source-level estimates, either explicitly (Matchin et al., 2019) or in a probabilistic manner (Cottereau et al., 2015; Wang and Holland, 2022). Another principle that has been applied to combine electrophysiological and hemodynamic signals is to use the MEG or electroencephalography (EEG) based individual measures of neural activity to model the fMRI response instead of common regressors across subjects, leading to increased statistical power (Renvall et al., 2012a; Iannaccone et al., 2015). In this effort, various kinds of computational analyses have been applied to obtain more comprehensive spatiotemporal accounts than can be afforded by MEG or fMRI data alone. In some instances, machine-learning based classification analyses have been conducted separately for MEG and fMRI data to obtain maximally accurate temporal and spatial accounts of neural phenomena (Brandman and Peelen, 2017). Computational models have been applied to identify electrophysiological correlates of behavioral processes which, in turn, have been used to model the trial-level variability within fMRI signals (Pisauro et al., 2017). Some studies have further utilized representational similarity analyses across MEG and fMRI data to accomplish spatially and temporally detailed characterization of neuronal activity (Cichy et al., 2014, 2016; Leonardelli and Fairhall, 2022). Generative models have also been utilized to capture the state transitions in both fMRI and MEG resting-state networks (Jiang et al., 2022). In a clinical setting, fusion of distinct neuroimaging measures, such as MEG and fMRI, is increasingly being used to improve the ability to distinguish between patient groups and control participants (Calhoun and Sui, 2016). Together, these reports demonstrate the versatile ways that MEG and fMRI can be merged to obtain spatiotemporally detailed accounts of neural-level processing.

In order to use electrophysiological and hemodynamic measures together in a principled manner, it is important to understand the relationship between the different types of measures. Numerous neuroimaging studies have therefore attempted to deepen this understanding. Initially, the emphasis was on how electrophysiological and hemodynamic techniques would allow the identification and localization of neural responses to the same types of stimuli (Sanders et al., 1996; Stippich et al., 1998). Subsequently, the focus has been more on identifying electrophysiological phenomena that correlate with the blood-oxygen-level dependent (BOLD) fMRI signal. In general, such studies have revealed robust and spectrally systematic correlation patterns between neural and hemodynamic activity (Logothetis et al., 2001; Mukamel et al., 2005; Scheeringa et al., 2011). However, it has also been shown that the correlation between electrophysiological measures and the BOLD signal varies across brain regions (Conner et al., 2011; Kujala et al., 2014). Moreover, comparisons between large-scale networks derived from MEG and fMRI have indicated a complex frequency-specific relationship between fMRI and the electrophysiological connectivity (Hipp and Siegel, 2015; Liljeström et al., 2015b). Furthermore, studies examining the MEG and fMRI signals in identical experimental settings from the same subjects have revealed systematic functional differences between the electrophysiological and BOLD responses (Liljeström et al., 2009; Vartiainen et al., 2011). Accordingly, it is commonly accepted that when integrating the temporally and spatially accurate views of neural processing from MEG and fMRI, it is crucial to consider the complex nature of the origins of hemodynamic fluctuations (Logothetis, 2008; Ekstrom, 2010; Lauritzen et al., 2012; Whitman et al., 2013).

One aspect that has received less attention in combining MEG and fMRI measures both for obtaining detailed spatiotemporal accounts as well as investigating neurovascular coupling has been the role of the applied stimulation or experimental tasks themselves. Naturally, a broad range of stimuli from different sensory modalities as well as various kinds of cognitive experiments have been applied. However, the main goal in those explorations has been to induce detectable signals in different neural systems across the cortex and to develop approaches that maximize the association between the two signal types (Lankinen et al., 2018), not to examine how the different stimuli and tasks might influence the joint modulation of electrophysiological and hemodynamic signals. Yet, it has been shown that the local neural and hemodynamic signals can be partially decoupled (O'Herron et al., 2016), and that the relationship between electrophysiological and hemodynamic signals depends on the correlation between the local inputs (Butler et al., 2017), effects that could cause variability in the MEG-fMRI correlations across stimuli and tasks. In the present study, we sought to explicitly utilize the inherently complex relationship between BOLD fluctuations and modulations of electrophysiological brain activity as well as the possible task-induced variability in this relationship to track and dissociate the neural engagement of different brain regions across distinct cognitive tasks. We asked whether any differences in neural processing related to distinct picture naming tasks could be highlighted through MEG-fMRI fusion as compared to isolated within-modality measures. Specifically, we investigated the variability of MEG-fMRI correlation patterns across three naming tasks (naming objects and actions from action images, and objects from object images) using a dataset where MEG and fMRI data were recorded from the same participants in identical experiments (Liljeström et al., 2009). The correlation patterns were obtained by first computing the MEG-fMRI correlation separately for each brain region, time window and frequency band, and by then applying variance minimizing hierarchical clustering to find clusters of similarly correlated brain areas. The approach allows the grouping of both task-invariant and task-dependent correlation patterns across brain regions regardless of their spatial adjacency. We predicted that our approach would reveal both types of correlation patterns and, critically, facilitate identification of neural engagement that could not be detected using one imaging modality alone.

Materials and methods

Subjects

Magnetoencephalography and fMRI data were collected from 10 healthy (nine right-handed, one ambidextrous), native Finnish-speaking subjects (four females, six males; ages 20–33 years). Informed consent was obtained from all subjects, in agreement with a prior approval of the Local Ethics Committee (Hospital District of Helsinki and Uusimaa). The subjects did not report any neurological disorders, and all had normal or corrected-to-normal vision. All methods were conducted in accordance with the guidelines of the Finnish National Board on Research Integrity.

Experimental design

The task was to silently name pictures of objects or actions presented as simple black line art on a gray background. There were two categories of drawings. In the first category, an action performed with an object was depicted, whereas in the second category, a single object was shown. To achieve the same visual complexity as in action images the object images were constructed from the action images by dissolving the action figures into non-meaningful lines in the background. The experiment consisted of three different cognitive tasks: Object naming from object images (100 trials), action naming from action images (100 trials), and object naming from action images (100 trials). The experiment had a blocked design with 10 stimuli of the same task presented within each 30-s block. Each image was shown for 300 ms at 1.8–4.2 s intervals. Each block started with an instruction indicating the task for the block. The task blocks were separated by 21-s rest blocks. The experiment was divided into two runs, with different sets of stimuli in the two runs (150 images per run, 50 per task). The experimental design was identical in MEG and fMRI leading to two matched runs per subject. The order of the three naming conditions was randomized in both runs and silent naming was used to avoid muscular artifacts. A complete description of the experiment can be found in Liljeström et al. (2009). The design permits identification of effects that are related to the naming task (comparing action naming to both object naming conditions) and to the picture type (comparing object-only images to both action image conditions). Behaviorally, (overt) object naming from action images leads to longer reaction times than for naming objects from object images or actions from action images (Liljeström et al., 2015a), indicating that increased effort or additional processing is required when naming objects from action images compared to the two other tasks. It is therefore of interest also to compare the object naming from action image condition to the other tasks.

Functional magnetic resonance imaging data collection

The MRI data were collected at the Advanced Magnetic Imaging Centre (Aalto University) with a Signa VH/i 3.0 T MRI scanner (GE Healthcare, Chalfont St Giles, UK). Anatomical MRIs were acquired using a T1-weighted 3D spoiled gradient-echo sequence. Functional MRI data were collected using a single-shot gradient-echo planar imaging sequence (TR 3 s, TE = 32 ms, FA = 90, slice thickness 3 mm, in-plane resolution either 3 mm × 3 mm, or 3.4 mm × 3.4 mm). The first five functional volumes were discarded from the analysis.

Magnetoencephalography data collection

Magnetoencephalography data recordings were conducted using a 306-channel whole-head device (Elekta Oy, Helsinki, Finland) in a magnetically shielded room. The data were bandpass filtered to 0.03–200 Hz and sampled at 600 Hz. The temporal extension of the Signal Space Separation method (Taulu and Simola, 2006) was applied in order to suppress contributions from external artifacts. Eye movements were monitored with electro-oculogram (EOG).

Functional magnetic resonance imaging and magnetoencephalography data analysis

The overview of the analysis pipeline including key formulae for the conducted computations is presented in **Figure 1**. First, to facilitate the across-subjects evaluation of MEG-fMRI correlation, the data of each subject were transformed to an average brain *via* a surface-based transformation (Fischl et al., 1999) using Freesurfer 5.3 (Fischl, 2012). Before the transformation, the individual fMRI data were realigned to the first volume and susceptibility artifacts caused by movements were corrected for using SPM8 (Wellcome Department of Cognitive Neurology, London, UK). The mean image of the functional series was used for co-registering the fMRI data with the individual anatomical images. For each vertex in the average brain, the fMRI values of a spatially matching voxels were then taken to represent the fMRI activity at the cortical surface level.

The vertex-level data were averaged within 188 parcels covering the entire cortical surface (see e.g., **Figure 3**). This parcellation was based on the automatic anatomic parcellation of the human cortical gyri and sulci consisting of 144 parcels (Destrieux et al., 2010) that was subsequently computationally modified to form a parcellation that would be more suitable for the analysis of MEG data. Specifically, the parcellation scheme was obtained by applying a PCA algorithm implemented in MNE-python (Gramfort et al., 2013) for splitting parcels defined by the automatic anatomical labeling scheme of the cortical surface (Destrieux et al., 2010). The splitting yields parcels that are relatively symmetrical and small enough to be relatively homogeneous with respect to local activations. Notably, the splitting was based solely on the anatomical information without utilizing functional data, leading to a parcellation that is less optimized (Thirion et al., 2014) but more generalizable to multiple datasets. While the exclusively anatomical parcellation obtained *via* splitting the original parcels using spatial PCA does not ensure an exact alignment between MEG and fMRI responses and the parcels, it allows better separation of responses that are spatially distinct than when using the original Destrieux atlas. From the entire parcellation consisting of 188 parcels, regions that are prone to artifacts or signal loss in either of the imaging methods (anterior parts of the frontal lobe, deepest parts of the medial surface, and inferior parts of the temporal lobe) were omitted. The set of parcels used in the final analysis consisted of 70 parcels per hemisphere.

In the fMRI analysis the goal was to determine, for each fMRI run and experimental condition, the BOLD signal change with respect to rest within each parcel. This was accomplished by first high-pass filtering the parcel-level representation of the fMRI data in SPM8 with a cut-off frequency of 1/510 Hz. Baseline effects were removed using a rest block (6 volumes) that preceded each stimulus block (11 volumes), thus removing slow drifts taking place during the scanning runs. For each fMRI block and parcel, the data were averaged across the collected 11 volumes. The data were then normalized by subtracting the mean activity across all blocks and tasks from the block and task specific values, and by dividing these values by the standard deviation of the whole run's data. The normalization was done separately for each subject to remove inter-subject differences in signal scales and means. Subsequently, within each run, blocks of each task were averaged per participant. For the correlation analysis, we thus obtained a total of 20 fMRI values per task (10 participants, 2 runs).

Magnetoencephalography data estimates were obtained for the same parcels in six different frequency bands: Theta (4–7 Hz), alpha (8–13 Hz), low beta (15–21 Hz), high beta (23–29 Hz), low gamma (36–46 Hz), and high gamma (54–90 Hz), from 100 to 800 ms with respect to stimulus presentation. The gamma-band analyses were conducted in two separate bands to avoid including 50 Hz line noise into the estimates. The estimation was done using event-related Dynamic Imaging of Coherent Sources (Laaksonen et al., 2008), a beamforming technique in the time-frequency domain. Here, only data from the 204 gradiometers were used. In the estimation, a surface-based grid consisting of 5,122 points was first created in the average brain with MNE (Gramfort et al., 2014) and transformed to each individual's anatomy using Freesurfer 5.3 (Fischl, 2012). Brain activity estimates for each task and block were then computed for each grid point in the six different frequency bands, in 22 partially overlapping 200-ms time-windows (33-ms time difference between two successive time-windows). The 200-ms window length was chosen as it was the shortest length that allowed the accurate estimation of data covariance and thus brain activity given the signal-to-noise ratio (SNR) and number of trials across experimental tasks within the present dataset (see, e.g., Brookes et al., 2008). This window length is likely to be sufficiently short for exploring the sustained neural phenomena in higher-order cortical regions but could be sub-optimal for determining the temporally intricate early processes within the visual hierarchy. A baseline value was computed from the prestimulus interval −200 to 0 ms, separately for each block and grid point. Trials in which the amplitude of either the vertical or the horizontal EOG exceeded 150 μV were rejected. The parcel-level values were obtained by calculating, in each grid point, the difference between the post-stimulus values and the corresponding baseline values (divided by the same baseline values) and computing the average across all these baseline relative changes within each parcel. The parcel values were normalized separately for each subject, run and frequency band. This was done similarly as for the fMRI data, by subtracting the average activity across all blocks and tasks from the block and task specific values, and by dividing these values by the standard deviation of the data from the entire run. The run-level data were then obtained by averaging the block-specific data within each run. Similarly to the fMRI data, we thus obtained a total

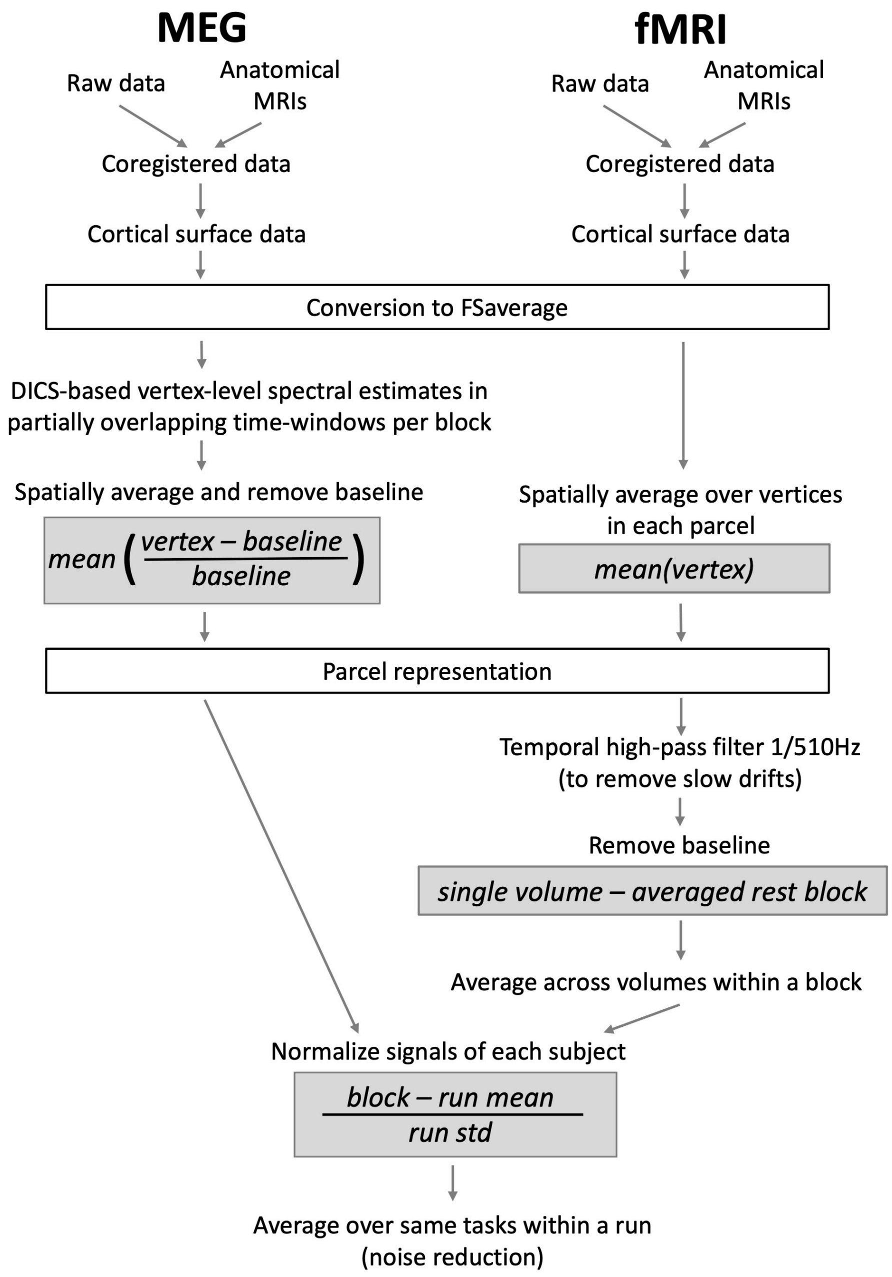

MEG
Raw data
Anatomical MRIs
Coregistered data
Cortical surface data
fMRI
Raw data
Anatomical MRIs
Coregistered data
Cortical surface data
Conversion to FSaverage
DICS-based vertex-level spectral estimates in partially overlapping time-windows per block
Spatially average and remove baseline
$mean\left(\frac{vertex - baseline}{baseline}\right)$
Spatially average over vertices in each parcel
mean(vertex)
Parcel representation
Temporal high-pass filter 1/510Hz (to remove slow drifts)
Remove baseline
single volume – averaged rest block
Average across volumes within a block
Normalize signals of each subject
$\frac{block - run\ mean}{run\ std}$
Average over same tasks within a run (noise reduction)

FIGURE 1
An outline of MEG and fMRI analysis pipelines, displaying the most important steps and their order. Gray boxes show essential normalizations that aim to equalize the measures obtained with the two modalities.

of 20 MEG values per task (10 participants, 2 runs) for the correlation analysis.

Correlation analysis and clustering

We computed a vector of MEG-fMRI correlation estimates for each parcel using Spearman's rank correlation (see **Figure 2**). Within a parcel, separate correlations were computed for all tasks, time intervals and frequency bands (3 tasks, 22 time-windows, 6 frequency bands: in total 396 MEG-fMRI correlation estimates per parcel). Each of these estimates was computed based on 20 MEG and 20 fMRI observations (10 subjects and 2 runs). We applied an agglomerative (merging) hierarchical clustering algorithm on our Spearman's rho value vectors to find clusters of similarly correlated regions. For this, we used the Ward minimum variance method (Ward, 1963) that aims to minimize the within-cluster variance, leading to a clustering where correlation patterns inside a cluster are as similar as possible [function linkage(. . .,"ward") in Matlab]. Information about hemispheres was not passed to the clustering algorithm. Ward's method measures Euclidean distances between cluster centroids during its merging steps. The clustering algorithm produces a hierarchical cluster tree structure that describes the merging process. The leaves of the tree can be reordered without changing the structure itself. The optimal leaf order is such that the similarities of adjacent leaves are maximized [function optimalleaforder() in Matlab]. The final clustering allows for visual comparison of the correlation patterns across the three tasks and different frequency bands.

To evaluate the possible differences in correlations across the tasks, we estimated the 99% confidence limits for each task across the identified clusters, separately for left and right-hemisphere parcels, using bootstrapping (Efron, 1979). The bootstrapping was conducted by re-sampling the data 10,000 times, by computing the new MEG-fMRI correlation values for each sample, and by estimating the 99% confidence limits for each task from the obtained distribution. In the re-sampling, 80% of the data were randomly selected at each round. In this evaluation, we only considered those clusters and frequency bands in which at least one of the three tasks showed significant MEG-fMRI correlation ($p < 0.05$, Bonferroni-corrected over time points).

To compare a joint analysis approach and a more conventional approach utilizing a single brain imaging method alone, we also evaluated the differences in the MEG activity patterns between the tasks with paired t-tests ($p < 0.05$, Bonferroni-corrected over time points) for the identified clusters. This analysis was performed with the same temporal and spectral resolution as the MEG-fMRI correlation analysis and was, thus, only applicable to MEG; fMRI lacks the temporal resolution that would be needed for comparison of fMRI activity and MEG-fMRI correlation modulations. The comparison was therefore restricted to MEG activity and MEG-fMRI correlation patterns. Potential differences in the temporal-spectral aspects of the findings between the two approaches would reveal

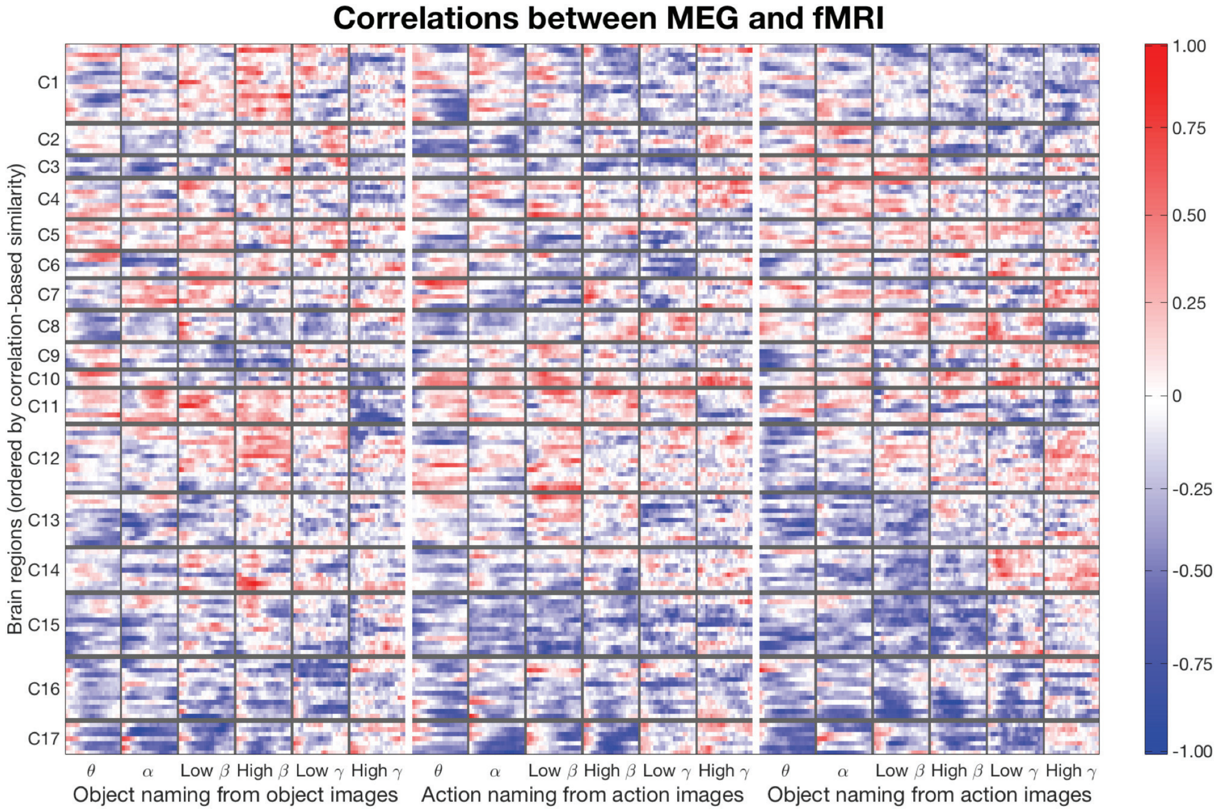


FIGURE 2
Matrix of correlations between MEG and fMRI for the three experimental conditions (separated with thick white vertical lines). Each condition-related submatrix is divided into six frequency bands: Theta, alpha, low beta, high beta, low gamma, and high gamma, from left to right (columns separated by thin vertical gray lines). Each frequency band consists of a sequence of 22 time points (sub-columns). All 140 brain regions (70 per hemisphere) are displayed on the y-axis, ordered with respect to the optimal leaf order of a cluster tree. This leads to a solution where distances between similarly behaving brain regions are minimized. The brain regions (rows) are divided into 17 clusters (C1–C17, separated by horizontal thin gray lines; see Figure 3 for visualization of the areas on MRI). The clustering is the same for all three conditions. The color indicates the MEG-fMRI correlation strength (−1. . .+1), see scale on the right.

unique results that can be achieved only with one of the approaches, but not both.

Results

Clustering of correlation patterns

For clustering purposes, a matrix was constructed (see Figure 2), where each row lists the MEG-fMRI correlation values across the different frequency bands, time points and tasks. The clustering algorithm enables identification of clusters in which all three tasks behave similarly, but also clusters in which the tasks behave differently. In Figure 2, the rows are reordered according to a full cluster tree so that similar rows are close to each other. The ordering reveals salient MEG-fMRI correlation patterns, with consistent negative and positive correlation patterns across brain regions. The selected clustering consists of 17 clusters (Figure 3), chosen based on an appropriate level of spatial separation across parcels. With a smaller number of clusters, functionally distinct brain regions remain in larger shared clusters, whereas with a larger number of clusters single parcels start to form clusters by themselves. Accordingly, with a smaller set of clusters, regions with functionally distinct activity profiles would be merged together, whereas with a larger set of clusters individual parcels with very similar activity profiles would be segregated into distinct clusters. In general, the clusters were spatially concentrated, indicating that close-by regions show more similar MEG-fMRI correlation patterns than regions that are further apart. The clustering (Figure 3) agreed well with the known functional division of cortical processing related to picture naming, revealing, e.g., components representing both lower (C8) and higher-order (C16) visual, speech related motor/premotor (C5 and C7), and perisylvian language related processing (C9). In particular, lower-order regions involved in the basic visual processing formed clusters (C8, C11, C16, and C17) that did not include any higher-order cortical areas, whereas the clusters containing higher-order regions generally represented distinct neural functions associated with different cortical lobes and also with more fine-grained differences (e.g., separation of inferior vs. superior frontal cortices and lateral vs. medial cortical structures). Many of the identified clusters (C1, C4, C8, C9, C11, C12, C15, and C17) showed marked symmetry across the hemispheres, but temporal, central and inferior frontal cortical areas (e.g., C2, C5–C7, and C13–C14) critically involved

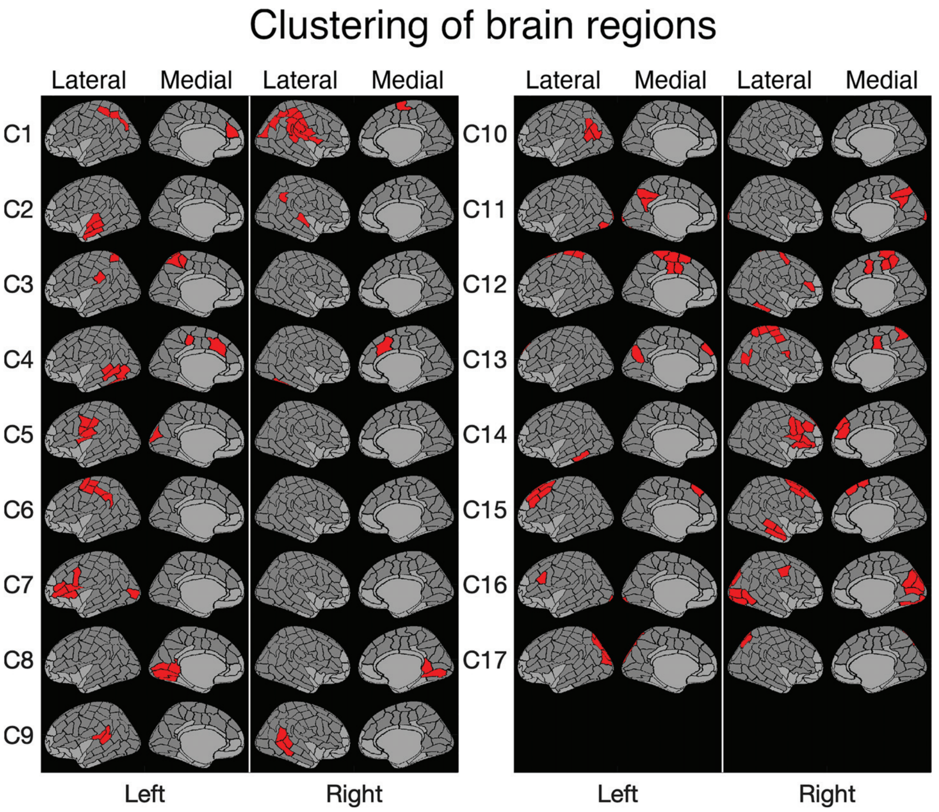


FIGURE 3
Clustering of brain regions. Level 17 of a clustering tree (used in the analyses). The deep medial and anterior frontal areas plotted in light gray were omitted from the analysis. Clusters are ordered according to the optimal leaf order and marked with labels C1–C17 (cf. Figure 2).

in picture naming tended form clusters exclusively within individual hemispheres.

Magnetoencephalography-functional magnetic resonance imaging correlation differences between tasks

For the clusters, we determined significant differences in MEG-fMRI correlation spectra between experimental conditions, across multiple frequency bands and time-windows (see Table 1 and Figure 4). We focused on identifying effects where one of the conditions differed from the other two conditions: (i) naming actions differed from both object naming conditions (different tasks; Figure 4, rectangles with solid orange line), (ii) naming objects from object pictures differed from naming objects or actions from action pictures (different images; Figure 4, rectangles with dotted black line), and (iii) naming objects from action pictures differed from both naming objects from object pictures and naming actions from action pictures (different reaction times; Figure 4, rectangles with solid gray line). Correlations were examined separately for parcels within each hemisphere.

Modulations of MEG-fMRI correlation across-tasks were detected predominantly in the left hemisphere. Different picture types elicited distinct correlation patterns in the occipital and parietal cortex, within the alpha and gamma frequency bands (left-hemisphere clusters C1, C3, and C13; see Figure 5A and Table 1). Between different naming tasks, correlations differed along the central sulcus and the posterior temporo-parietal cortex, mainly in the left hemisphere (left-hemisphere clusters C5, C6, and C10 and right-hemisphere cluster C2), particularly in the gamma-range. Distinct correlation patterns for the condition in which the participants named objects from action images as compared to the other two categories were observed exclusively in the left hemisphere and included brain regions within the posterior temporo-parietal cortex (cluster C10) as well as within the occipital cortex (clusters C11 and C17), with contributions from the theta band as well as low and high gamma-bands.

Task-invariant magnetoencephalography-functional magnetic resonance imaging correlation patterns

Figures 5B Shows the correlation patterns for two clusters within the occipital cortex (clusters C16 and C17). Parcels in the left-hemisphere cluster C17, covering the middle occipital cortex, and those in the right-hemisphere cluster C16, covering the medial and lateral parts of the occipital cortex, showed a significant negative MEG-fMRI correlation at low frequencies, but not in the gamma-range.

Magnetoencephalography activation vs. magnetoencephalography-functional magnetic resonance imaging correlation

Across the 17 identified clusters, the time-frequency windows in which MEG-fMRI correlation showed task-dependent modulation were highly distinct from the time-frequency windows in which MEG activity was modulated (Figure 6). Modulation of correlation was observed mainly in early time-windows (<500 ms), whereas modulations of activity (with band-limited power as measure) were exclusively detected more than 500 ms after stimulus onset. In the frequency domain, the MEG activity modulations were concentrated to the theta, alpha and (low and high) beta bands, whereas the MEG-fMRI correlation effects also showed a prominent contribution of gamma-band neural activity. No significant MEG-fMRI correlation effects were detected in the high beta band. Significant differences in activation were detected between object naming from object vs. action images, as well as for object naming from object images vs. action naming from action images; however, no differences were observed between object vs. action naming from action images (Figure 7). These effects were particularly prominent within the left hemisphere, predominantly in clusters with parcels in the parietal lobe. No significant effects of MEG signal changes were detected between object and action naming from identical images, in contrast to the MEG-fMRI correlation analysis which identified several left-hemisphere clusters in which action naming differed from the other two conditions (C2, C5, C10, and C13, Figure 5).

Discussion

We have shown that correlation between MEG and fMRI contains information that distinguishes between the three naming tasks. This finding aligns with observations that have demonstrated trial and stimulus dependent variability in the relationship between electrophysiological and hemodynamic activity within the visual cortex (O'Herron et al., 2016; Butler et al., 2017). Furthermore, our results demonstrate that the time-frequency windows in which the MEG-fMRI correlation patterns differ between the tasks are distinct from the windows showing task effects in a separate MEG-based analysis of modulation of neural activity. Interestingly, the differences in the correlation patterns between tasks were typically observed in markedly transient time-windows, highlighting the dynamic nature of the neural phenomena dissociating the different

TABLE 1 Significant effects detected with the given clusters, frequency bands, and significant time intervals.

Cluster	Hemisphere and primary location	Frequency band	Time (ms)	Confidence interval
Correlation patterns specific to object images				
C1	Left, superior parietal cortex	High gamma	300–600	99.9%
C3	Left, inferior parietal cortex and precuneus	Alpha	200–430	99.9%
C13	Left, cuneus and anterior frontal cortex	Alpha	100–370	99.9%
Correlation patterns specific to action naming				
C2	Right, superior temporal and inferior parietal cortex	Low gamma	100–300	99%
C5	Left, inferior precentral gyrus	Low gamma	230–430	99.9%
C6	Left, pre- and postcentral gyrus	Low gamma	300–500	99.9%
C10	Left, posterior temporo-parietal cortex	Low beta	200–470	99.9%
Correlation patterns specific to naming objects from action images				
C10	Left, posterior temporo-parietal cortex	Low gamma	230–430	99%
C11	Left, precuneus and occipital pole	High gamma	200–400	99%
		High gamma	300–500	99.9%
C17	Left, middle occipital cortex	Theta	330–570	99%

As significances are computed over 200-ms time-windows, it determines the lower bound for the size of significant time window. The correlation significance is Bonferroni corrected (p = 0.05) over 22 time-windows.

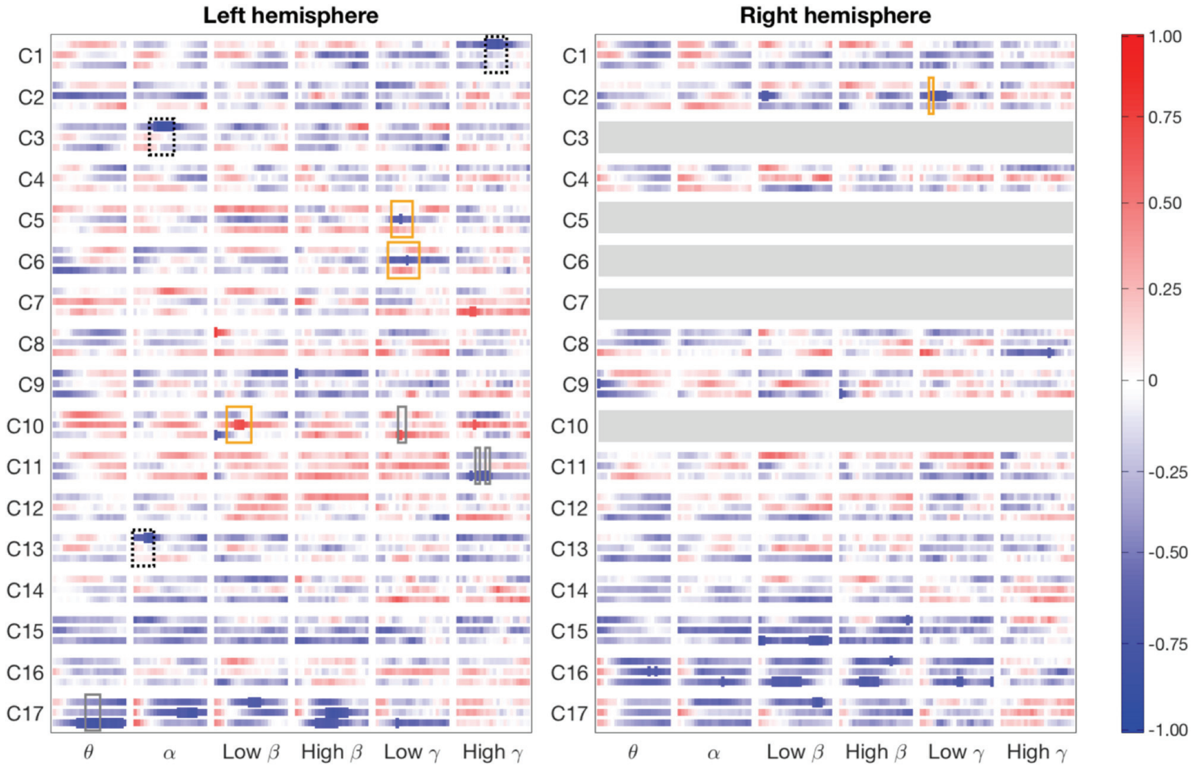


FIGURE 4
Magnetoencephalography-Functional magnetic resonance imaging correlation patterns divided into clusters (row labels) and hemispheres (left and right panels). The three rows in each cluster show correlation between fMRI and MEG for the three experimental conditions: from top to bottom, object naming from object images, action naming from action images and object naming from action images, over time in the different frequency bands (column labels). Significant correlations (p = 0.05, Bonferroni-corrected over the 22 time points) are marked as thicker parts of stripes. Rectangles indicate areas where the 99% confidence intervals of one condition do not overlap those of the other two conditions. A salient difference between naming tasks (naming actions vs. objects) is denoted by an orange rectangle, a difference between two picture types (action vs. object stimulus) is indicated by a dotted black rectangle, and a difference specific to naming objects from action images vs. the other two tasks with a gray rectangle. A rectangle is shown only when there is also a significant MEG–fMRI correlation inside the rectangle. Clusters C3, C5-C7, and C10 have parcels only in the left hemisphere (blank gray bars in the right-hemisphere).

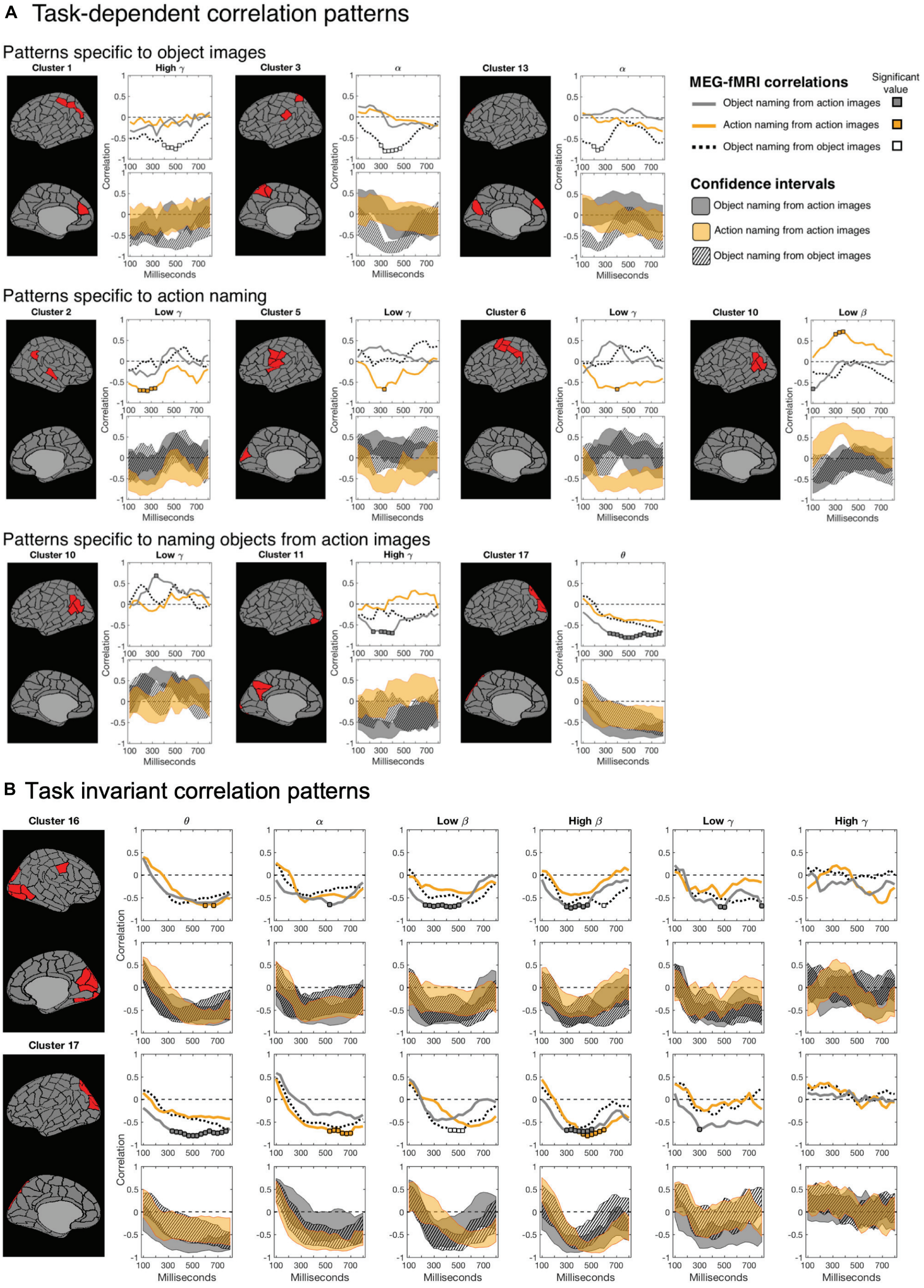


FIGURE 5

Magnetoencephalography-functional magnetic resonance imaging correlation as a function of time. For each cluster, the top row shows the correlation spectra for all tasks (naming object from action pictures in gray; naming actions in orange; and naming objects from object pictures in dotted black), and the bottom row the 99% confidence intervals for the three tasks (correspondingly gray, orange and a striped black pattern). In the correlation spectra, the colored squares indicate time instances at which the correlation is significant (p = 0.05, Bonferroni-corrected over time). **(A)** *Task-dependent instances*: One task shows significant correlation and differs from the other two tasks (non-overlapping confidence bounds) at a given time. White areas between the confidence intervals of experimental conditions indicate time instances of significantly different MEG-fMRI correlation between two or more conditions (p = 0.01, uncorrected). **(B)** *Task-invariant instances*: Clusters 16 and 17 suggest consistent negative correlation between MEG and fMRI at lower frequencies, among all experimental conditions, in the occipital cortex.

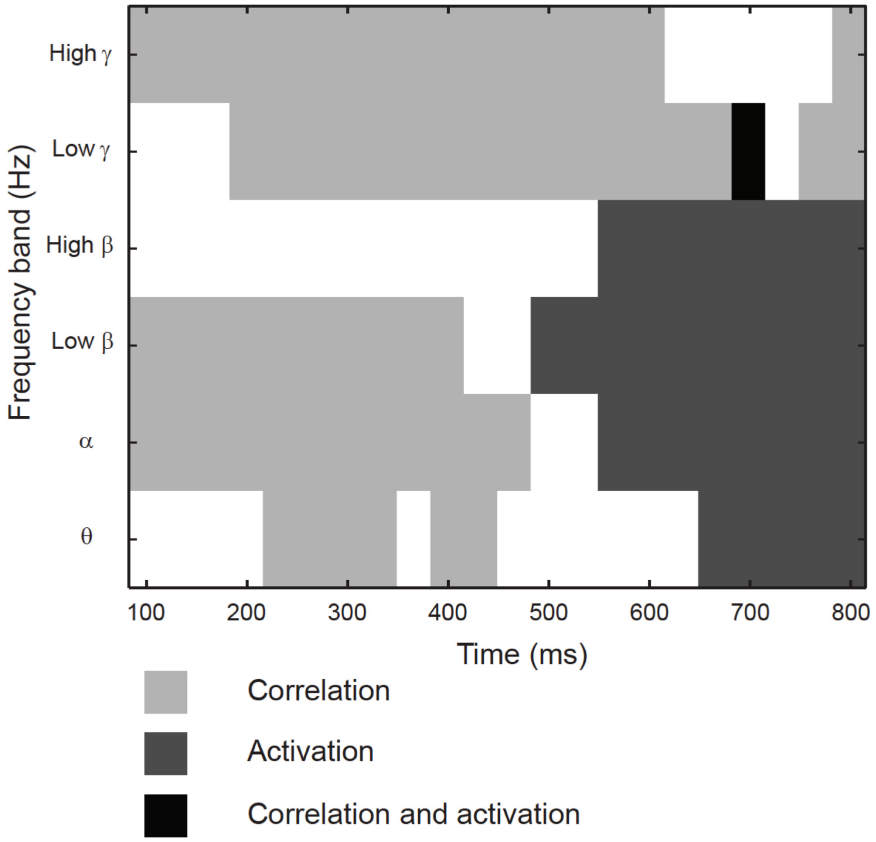


FIGURE 6
Temporo-spectral uniqueness and overlap in modulation of rhythmic activity and MEG-fMRI correlation. Timing with respect to picture presentation is plotted on the *x*-axis, and the different frequency bands on the y-axis. Time-frequency windows that showed differences between the conditions only for MEG band-limited power (light gray), only for MEG-fMRI correlation (dark gray) or both (black). Values averaged across all contrasts.

picture naming conditions also at the level of MEG-fMRI correlations. Notably, such correlation differences were not specific to any frequency bands but extended to a wide range of distinct oscillations (theta, alpha, beta, and gamma). On the other hand, task-invariant correlations especially in the theta- and alpha-bands tended to be more sustained, attesting to the distinct nature of task-dependent vs. task-invariant correlation patterns. From amongst the 17 identified clusters, nine showed significant differences between the three experimental conditions whereas no differences were observed in the other clusters covering, in particular, more anterior lateral frontal areas and primary visual cortices. Significant differences were observed for all contrasts in the parietal cortex, with more superior effects for different images and more inferior effects for different tasks and conditions with different reaction times. Differences in the MEG-fMRI correlation patterns were also observed for different images in the anterior medial frontal cortex and for different tasks in the post- and precentral gyri. Notably, the involvement of the parietal cortex was detected also in the analyses focusing on the MEG and fMRI activity, whereas the role of the anterior medial frontal cortex and the post- and precentral gyri in dissociating the different naming conditions was not observed in these studies (Liljeström et al., 2008, 2009). Our results thus illustrate that the multimodal correlations yield novel information about the task-dependent neural engagement that cannot be detected using one imaging method alone.

Detection of neural engagement using multiple neuroimaging methods

Task-dependent processing in neural circuits is a complex phenomenon that is supported by a wide range of mechanisms involving, e.g., electric, metabolic, and neurotransmitter activity (Singh, 2012). Measuring any of these processes yields one particular view of the full activity of the circuit. As it is not feasible to simultaneously record all possible processes related to the engagement of a circuit, its full activity remains a variable that may be estimated using specific proxies. As individual

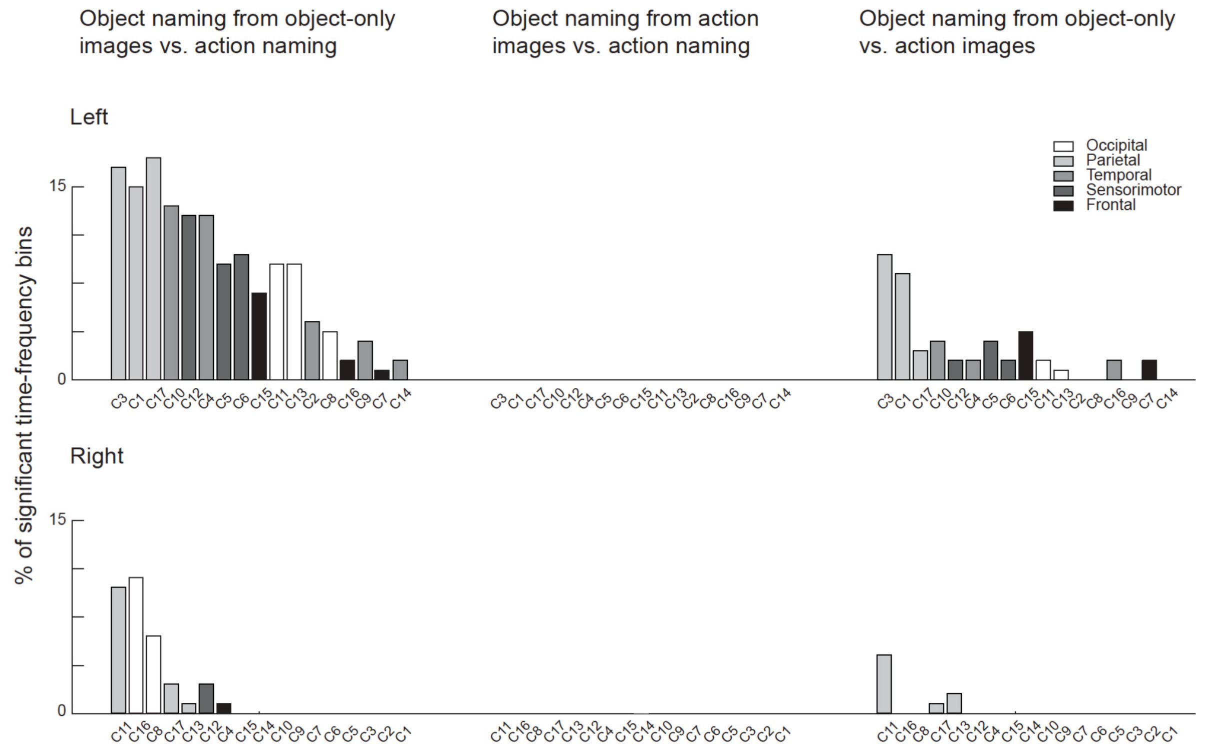


FIGURE 7
Significant results in the MEG activation analysis for each cluster. For each cluster the number of significant time-frequency bins are indicated as the percentage of all possible time-frequency bins (in total 132 bins from 6 frequency bands and 22 time-windows). The bars are color-coded according to the lobe to which the majority of the parcels belong to. The clusters are ordered according to the total percentage of significant time-frequency bins in all tasks (left and right- hemispheres separately). Note that there were no significant differences between neural activity during Object naming from action images and action naming, whereas for the other two contrasts where the stimulus contents were different multiple clusters showed significant differences.

proxies are noisy and give incomplete information, it may not be possible to accurately estimate the full brain activity in a region. Thus, the observed activation patterns determined by an individual proxy may not reveal any observable brain activity even if the neural circuit, in reality, participates in task-dependent processing. The same holds when the goal is to determine differences between levels of neural engagement between different experimental conditions.

It has been proposed that the complexity of the human brain coupled with the incomplete measurements make multimodal data fusion critical for identifying detailed, individual-level properties of brain anatomy and function (Calhoun and Sui, 2016). Multimodal data-fusion based approaches have proven particularly useful for combining genetic mapping with other measures in the study of brain disorders (Purcell et al., 2009; Pearlson et al., 2015) as well as for evaluating the variability of brain anatomy and function in healthy subjects (Hardoon et al., 2009; Le Floch et al., 2012; Renvall et al., 2012b; Salmela et al., 2016) and predicting the subjects' age (Engemann et al., 2020). So far, fusion of different neuroimaging data-types has been applied for identifying (in individual brain regions), e.g., the neural underpinnings of the BOLD response (Scheeringa et al., 2011; Kujala et al., 2014), also at the laminar level (Scheeringa et al., 2016; Warbrick, 2022), the effects of anatomical properties on functional data (Sepulcre et al., 2009; Schwarzkopf et al., 2012), or the effects of GABAergic inhibition on fMRI and MEG responses (Muthukumaraswamy et al., 2009; Kujala et al., 2015). While it has been proposed that by combining the temporally/spectrally and spatially sensitive measures of neural engagement provided by MEG and fMRI one could obtain a spatiotemporally accurate picture of brain activity (Dale et al., 2000), such data fusion has rarely been applied. Moreover, this type of combination has typically been used only in the primary sensory and motor neural systems (Schulz et al., 2004; Whittingstall et al., 2007; Stevenson et al., 2012; Renvall et al., 2012a; Cichy et al., 2014). Recently, similarity-based fusion methods combining MEG and fMRI have proven useful in learning relationships between visual objects and how they are represented within the visual system (Cichy et al., 2016) as well as within the semantic system (Leonardelli and Fairhall, 2022). In cognitive tasks, the improvement of SNR through group-level analysis (increased amount of data) may be limited by notable inter-subject variability, leading to a failure to detect the true engagement of neural circuits, even when multiple proxies are combined.

In the present study, we aimed to develop and apply a data-fusion based approach that would explicitly utilize the inter-subject and inter-block variability in combining different measurements (MEG and fMRI) to build a more sensitive and accurate picture of the neural engagement. Specifically, we used the correlation between MEG estimates of induced activity in different time-frequency windows and BOLD-fMRI estimates of hemodynamic activity to determine the neural circuits that are engaged in a distinct manner in three picture naming tasks. The MEG and fMRI proxies of neural activity can occasionally show salient negative or positive correlation when the brain activity is strong enough to be detected. In areas where one imaging method yields only noise and the other a good signal, task-wise correlations cannot be significant. To detect activity in a neural circuit, our approach requires that there is a causal connection between the engagement of the circuit and the two proxies (MEG and fMRI). Notably, unlike in typical neuroimaging studies, the sensitivity of the approach to detect neural engagement is in fact increased if the subjects or the blocks show considerable variability, given that the assumption of causality is met. In general, our approach as well as other approaches that profit from such variability are likely to be beneficial in cases where the SNR is low and where there is large individual variance in elicited neural processes. Hence, this type of approaches should prove useful in detecting neural engagement particularly in cognitive tasks.

Multimodal correlation as a spatially, temporally and spectrally unique view on neural engagement during picture naming

In the present study, we applied the developed MEG-fMRI correlation based method to a picture naming data set that had been previously analyzed separately using traditional MEG (evoked responses) and fMRI group-level statistical approaches for identification of neural activity related to different naming tasks (action vs. object naming) (Liljeström et al., 2009) as well as identification of task-relevant functional networks (Liljeström et al., 2015a). Several studies have shown a negative correlation between MEG and fMRI at lower alpha and beta frequencies, and a positive correlation within the gamma frequency range, especially in low-level sensory cortices (Logothetis et al., 2001; Mukamel et al., 2005; Scheeringa et al., 2011). In higher-level cortical regions and in cognitive tasks this relationship is more variable (Conner et al., 2011; Kujala et al., 2014). Moreover, analysis of functional networks has indicated a complex frequency-dependent relationship between MEG- and fMRI-derived networks that varies across-tasks (Liljeström et al., 2015b). In the present study, we observed task-invariant negative correlations between MEG and fMRI within the alpha and beta frequency bands in occipital and parietal regions, in line with previous studies (Logothetis et al., 2001; Scheeringa et al., 2011).

Our main goal was, however, to utilize the variability in the relationship between MEG and fMRI and identify clusters that manifested a task-varying relationship in MEG and fMRI correlation. This correlation-based approach revealed significant differences between the conditions in which the activation based analysis had not done so. Within the left parieto-temporal junction, along the central sulcus, and the inferior frontal cortex, the correlation pattern was different between the action naming condition and the two object naming conditions. In contrast, MEG activation analysis either with induced responses in the present study, or previously with evoked responses (Liljeström et al., 2009), did not reveal significant differences between action and object naming from identical images. These effects demonstrate that the correlation-based analysis can reveal neural engagement in functionally relevant circuits that are not detected in conventional activation-based analyses.

The most notable new insights revealed by the present approach were the spectral and temporal patterns of electrophysiological activity. For example, the correlation patterns differed in the parieto-occipital cortex for the conditions where the stimulus content was different. In the present analysis of the modulation of induced activity, effects were detected in late time-windows (>500 ms), whereas the correlation patterns revealed differences primarily in notably earlier intervals (200–400 ms). These findings suggest that the modulation of alpha/beta activity is distinct for different stimulus contents, a finding that could not be inferred from traditional analysis of MEG activation; the results also demonstrate that these early differences are linked with the BOLD activity that is measured in those cortical regions. Secondly, the correlation-based analysis revealed, in contrast to analysis of MEG induced activity, prominent effects in the gamma-band. This suggest that the present multimodal analysis may help reveal the role of high-frequency neural activity in cognitive processing that is often difficult to detect with non-invasive techniques.

Detection of cortical activity using clustering of magnetoencephalography-functional magnetic resonance imaging correlation patterns

In the present study, we computed the correlation between individual-level, run-wise MEG and fMRI recordings of the same experimental conditions from the same subjects. The goal was to develop an approach that would utilize the correlation between the two distinct proxies (MEG and fMRI) of brain activity to enhance the sensitivity of detecting the engagement of neural circuits in cognitive processing. The approach thus aims to capture effects related to the stimulus- and state-dependent input correlations and differences in the propagation of vascular dilation between neural columns (O'Herron et al., 2016; Butler et al., 2017) that would manifest as differences in the MEG-fMRI correlation patterns across experimental tasks. It should, however, be noted that our approach does not directly tell whether the circuit is more or less engaged during a task; the correlation-based measure can only reveal that the relationship between the applied proxies has changed. For example, our two proxies (MEG and fMRI) can be negatively or positively correlated, without indicating whether the amount of activity in the circuit has increased or decreased compared to the other conditions. In areas where one imaging method reveals only noise and the other detectable cortical activation, the task-wise correlations should not be significant. Our clustering approach corresponds to a conditioning which enforces the method to consider only those correlations that are related to the performed cognitive tasks. In the optimal situation, both proxies would have similar temporal granularity but, due to the highly integrative nature of the fMRI signal, precise temporal information was present only in the MEG signals. Nonetheless, we can track and utilize the temporal information in the MEG signals to dissociate even subtle effects in the integrative fMRI signals and, thereby, discover also small differences between cognitive tasks.

Spatially, our clustering-based analysis was designed to identify robust, large-scale effects in the correlation patterns that were specific to the given three naming tasks. Thus, the clustering results may not necessarily obey conventional knowledge about the locations of task-relevant functional brain regions. The reason is that the clustering is constructed using a very limited set of tasks. If these tasks do not distinguish between certain brain regions, then those regions will fall into the same cluster. Moreover, if the spatial extent of a cluster is too large, even a relatively strong signal may be masked by other contradicting signals or noise originating from the same cluster. If a cluster is too small, weak but significant signals may disappear as the region of activation has been split into parts. Some of the clusters are necessarily non-informative because none of the brain regions—including inactive regions—are left out in a clustering process.

In our study, we let the method cluster both hemispheres together. Thus, it is also possible that in some cases weaker, interesting signals might have been masked by stronger signals from the other hemisphere. Such a scenario could be avoided by conducting separate clustering for each hemisphere; however, this might hide some of the inter-hemispheric effects that were detected with the present approach.

Conclusion

We introduced a correlation-based data-fusion analysis pipeline that utilizes two proxies of brain activity to enhance sensitivity for detecting the engagement of neural circuits in cognitive processing. Our results demonstrate that the approach discovers spatially, spectrally, and temporally unique task-specific information on cortical processing during picture naming. Multimodal data fusion based on correlations between electromagnetic and hemodynamic activity can thus reveal task-dependent neural engagement that may not be detected using the proxies of brain activity offered by one imaging method alone.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The MEG and fMRI data cannot be made openly available, according to the ethical permission and national privacy regulations at the time of the study, but are available from the corresponding author on reasonable request and with permission of the Ethics Committee of the Hospital district of Helsinki and Uusimaa. Requests to access these datasets should be directed to TM, tommi.mononen@helsinki.fi.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of the Hospital district of Helsinki and Uusimaa. The patients/participants provided their written informed consent to participate in this study.

Author contributions

TM, JK, ML, EL, SK, and RS: conceptualization and writing—review and editing. TM, JK, ML, EL, and SK: methodology. TM, JK, and ML: validation and formal analysis. ML: investigation. TM, JK, ML, and RS: writing—original draft. SK and RS: supervision and funding acquisition. All authors contributions is based on the CRediT taxonomy.

Funding

This work was financially supported by the Academy of Finland (Finnish Center of Excellence in Computational Inference Research COIN and grants #292334, #294238 to SK;

#255349, #315553 to RS; #257576 to JK; #286405 funding for TM), the Sigrid Jusélius Foundation (grant to RS), the Finnish Cultural Foundation (grant to ML), the Swedish Cultural Foundation in Finland (grant to ML), the Maud Kuistila Memorial Foundation (grant to ML), and Aalto Brain Center.

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

Brandman, T., and Peelen, M. V. (2017). Interaction between scene and object processing revealed by human fMRI and MEG decoding. *J. Neurosci.* 37, 7700–7710. doi: 10.1523/JNEUROSCI.0582-17.2017

Brookes, M. J., Vrba, J., Robinson, S. E., Stevenson, C. M., Peters, A. M., Barnes, G. R., et al. (2008). Optimising experimental design for MEG beamformer imaging. *Neuroimage* 39, 1788–1802. doi: 10.1016/j.neuroimage.2007.09.050

Butler, R., Bernier, P. M., Lefebvre, J., Gilbert, G., and Whittingstall, K. (2017). Decorrelated input dissociates narrow band gamma power and BOLD in human visual cortex. *J. Neurosci.* 37, 5408–5418. doi: 10.1523/JNEUROSCI.3938-16.2017

Calhoun, V. D., and Sui, J. (2016). Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 1, 230–244. doi: 10.1016/j.bpsc.2015.12.005

Cichy, R. M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. *Nat. Neurosci.* 17, 455–462. doi: 10.1038/nn.3635

Cichy, R. M., Pantazis, D., and Oliva, A. (2016). Similarity-based fusion of MEG and fMRI reveals spatio-temporal dynamics in human cortex during visual object recognition. *Cereb. Cortex* 26, 3563–3579. doi: 10.1093/cercor/bhw135

Conner, C. R., Ellmore, T. M., Pieters, T. A., Disano, M. A., and Tandon, N. (2011). Variability of the relationship between electrophysiology and BOLD-fMRI across cortical regions in humans. *J. Neurosci.* 31, 12855–12865. doi: 10.1523/JNEUROSCI.1457-11.2011

Cottereau, B. R., Ales, J. M., and Norcia, A. M. (2015). How to use fMRI functional localizers to improve EEG/MEG source estimation. *J. Neurosci. Methods* 250, 64–73. doi: 10.1016/j.jneumeth.2014.07.015

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., et al. (2000). Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron* 26, 55–67. doi: 10.1016/S0896-6273(00)81138-1

Destrieux, C., Fischl, B., Dale, A., and Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* 53, 1–15. doi: 10.1016/j.neuroimage.2010.06.010

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Stat.* 7, 1–26. doi: 10.1214/aos/1176344552

Ekstrom, A. (2010). How and when the fMRI BOLD signal relates to underlying neural activity: The danger in dissociation. *Brain Res. Rev.* 62, 233–244. doi: 10.1016/j.brainresrev.2009.12.004

Engemann, D. A., Kozynets, O., Sabbagh, D., Lemaitre, G., Varoquaux, G., Liem, F., et al. (2020). Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers. *Elife* 9, 1–33. doi: 10.7554/eLife.54055

Fischl, B. (2012). FreeSurfer. *Neuroimage* 62, 774–781. doi: 10.1016/j.neuroimage.2012.01.021

Fischl, B., Sereno, M. I., and Dale, A. M. (1999). Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage* 9, 195–207. doi: 10.1006/nimg.1998.0396

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2013). MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* 7:267. doi: 10.3389/fnins.2013.00267

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2014). MNE software for processing MEG and EEG data. *Neuroimage* 86, 446–460. doi: 10.1016/j.neuroimage.2013.10.027

Hardoon, D. R., Ettinger, U., Mourao-Miranda, J., Antonova, E., Collier, D., Kumari, V., et al. (2009). Correlation-based multivariate analysis of genetic influence on brain volume. *Neurosci. Lett.* 450, 281–286. doi: 10.1016/j.neulet.2008.11.035

Henson, R. N., Flandin, G., Friston, K. J., and Mattout, J. (2010). A parametric empirical Bayesian framework for fMRI-constrained MEG/EEG source reconstruction. *Hum. Brain Mapp.* 31, 1512–1531. doi: 10.1002/hbm.20956

Hipp, J. F., and Siegel, M. (2015). BOLD fMRI correlation reflects frequency-specific neuronal correlation. *Curr. Biol.* 25, 1368–1374. doi: 10.1016/j.cub.2015.03.049

Iannaccone, R., Hauser, T. U., Staempfli, P., Walitza, S., Brandeis, D., and Brem, S. (2015). Conflict monitoring and error processing: New insights from simultaneous EEG–fMRI. *Neuroimage* 105, 395–407. doi: 10.1016/j.neuroimage.2014.10.028

Jiang, F., Jin, H., Gao, Y., Xie, X., Cummings, J., Raj, A., et al. (2022). Time-varying dynamic network model for dynamic resting state functional connectivity in fMRI and MEG imaging. *Neuroimage* 254:119131. doi: 10.1016/j.neuroimage.2022.119131

Kujala, J., Jung, J., Bouvard, S., Lecaignard, F., Lothe, A., Bouet, R., et al. (2015). Gamma oscillations in V1 are correlated with GABAA receptor density: A multi-modal MEG and Flumazenil-PET study. *Sci. Rep.* 5:16347. doi: 10.1038/srep16347

Kujala, J., Sudre, G., Vartiainen, J., Liljestrom, M., Mitchell, T., and Salmelin, R. (2014). Multivariate analysis of correlation between electrophysiological and hemodynamic responses during cognitive processing. *Neuroimage* 92, 207–216. doi: 10.1016/j.neuroimage.2014.01.057

Laaksonen, H., Kujala, J., and Salmelin, R. (2008). A method for spatiotemporal mapping of event-related modulation of cortical rhythmic activity. *Neuroimage* 42, 207–217. doi: 10.1016/j.neuroimage.2008.04.175

Lankinen, K., Saari, J., Hlushchuk, Y., Tikka, P., Parkkonen, L., Hari, R., et al. (2018). Consistency and similarity of MEG- and fMRI-signal time courses during movie viewing. *Neuroimage* 173, 361–369. doi: 10.1016/j.neuroimage.2018.02.045

Lauritzen, M., Mathiesen, C., Schaefer, K., and Thomsen, K. J. (2012). Neuronal inhibition and excitation, and the dichotomic control of brain hemodynamic and oxygen responses. *Neuroimage* 62, 1040–1050. doi: 10.1016/j.neuroimage.2012.01.040

Le Floch, E., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., et al. (2012). Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. *Neuroimage* 63, 11–24. doi: 10.1016/j.neuroimage.2012.06.061

Leonardelli, E., and Fairhall, S. L. (2022). Similarity-based fMRI-MEG fusion reveals hierarchical organisation within the brain's semantic system. *Neuroimage* 259:119405. doi: 10.1016/j.neuroimage.2022.119405

Liljeström, M., Tarkiainen, A., Parviainen, T., Kujala, J., Numminen, J., Hiltunen, J., et al. (2008). Perceiving and naming actions and objects. *NeuroImage* 41, 1132–1141. doi: 10.1016/j.neuroimage.2008.03.016

Liljeström, M., Hultén, A., Parkkonen, L., and Salmelin, R. (2009). Comparing MEG and fMRI views to naming actions and objects. *Hum. Brain Mapp.* 30, 1845–1856. doi: 10.1002/hbm.20785

Liljeström, M., Stevenson, C., Kujala, J., and Salmelin, R. (2015b). Task- and stimulus-related cortical networks in language production: Exploring similarity of MEG- and fMRI-derived functional connectivity. *Neuroimage* 120, 75–87. doi: 10.1016/j.neuroimage.2015.07.017

Liljeström, M., Kujala, J., Stevenson, C., and Salmelin, R. (2015a). Dynamic reconfiguration of the language network preceding onset of speech in picture naming. *Hum. Brain Mapp.* 36, 1202–1216. doi: 10.1002/hbm.22697

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature* 453, 869–878. doi: 10.1038/nature06976

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature* 412, 150–157. doi: 10.1038/35084005

Matchin, W., Brodbeck, C., Hammerly, C., and Lau, E. (2019). The temporal dynamics of structure and content in sentence comprehension: Evidence from fMRI-constrained MEG. *Hum. Brain Mapp.* 40, 663–678. doi: 10.1002/hbm.24403

Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., and Malach, R. (2005). Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. *Science* 309, 951–954. doi: 10.1126/science.1110913

Muthukumaraswamy, S. D., Edden, R. A., Jones, D. K., Swettenham, J. B., and Singh, K. D. (2009). Resting GABA concentration predicts peak gamma frequency and fMRI amplitude in response to visual stimulation in humans. *Proc. Natl. Acad. Sci. U.S.A.* 106, 8356–8361. doi: 10.1073/pnas.0900728106

O'Herron, P., Chhatbar, P. Y., Levy, M., Shen, Z., Schramm, A. E., Lu, Z., et al. (2016). Neural correlates of single-vessel haemodynamic responses *in vivo*. *Nature* 534, 378–382. doi: 10.1038/nature17965

Pearlson, G. D., Liu, J., and Calhoun, V. D. (2015). An introductory review of parallel independent component analysis (p-ICA) and a guide to applying p-ICA to genetic data and imaging phenotypes to identify disease-associated biological pathways and systems in common complex disorders. *Front. Genet.* 6:276. doi: 10.3389/fgene.2015.00276

Pisauro, M. A., Fouragnan, E., Retzler, C., and Philiastides, M. G. (2017). Neural correlates of evidence accumulation during value-based decisions revealed *via* simultaneous EEG-fMRI. *Nat. Commun.* 8:15808. doi: 10.1038/ncomms15808

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'donovan, M. C., Sullivan, P. F., et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460, 748–752. doi: 10.1038/nature08185

Renvall, H., Formisano, E., Parviainen, T., Bonte, M., Vihla, M., and Salmelin, R. (2012a). Parametric merging of MEG and fMRI reveals spatiotemporal differences in cortical processing of spoken words and environmental sounds in background noise. *Cereb. Cortex* 22, 132–143. doi: 10.1093/cercor/bhr095

Renvall, H., Salmela, E., Vihla, M., Illman, M., Leinonen, E., Kere, J., et al. (2012b). Genome-wide linkage analysis of human auditory cortical activation suggests distinct loci on chromosomes 2, 3, and 8. *J. Neurosci.* 32, 14511–14518. doi: 10.1523/JNEUROSCI.1483-12.2012

Salmela, E., Renvall, H., Kujala, J., Hakosalo, O., Illman, M., Vihla, M., et al. (2016). Evidence for genetic regulation of the human parieto-occipital 10-Hz rhythmic activity. *Eur. J. Neurosci.* 44, 1963–1971. doi: 10.1111/ejn.13300

Sanders, J. A., Lewine, J. D., and Orrison, W. W. Jr. (1996). Comparison of primary motor cortex localization using functional magnetic resonance imaging and magnetoencephalography. *Hum. Brain Mapp.* 4, 47–57. doi: 10.1002/(SICI)1097-0193(1996)4:1<47::AID-HBM3>3.0.CO;2-P

Scheeringa, R., Fries, P., Petersson, K. M., Oostenveld, R., Grothe, I., Norris, D. G., et al. (2011). Neuronal dynamics underlying high- and low-frequency EEG oscillations contribute independently to the human BOLD signal. *Neuron* 69, 572–583. doi: 10.1016/j.neuron.2010.11.044

Scheeringa, R., Koopmans, P. J., van Mourik, T., Jensen, O., and Norris, D. G. (2016). The relationship between oscillatory EEG activity and the laminar-specific BOLD signal. *Proc. Natl. Acad. Sci. U.S.A.* 113, 6761–6766. doi: 10.1073/pnas.1522577113

Schulz, M., Chau, W., Graham, S. J., Mcintosh, A. R., Ross, B., Ishii, R., et al. (2004). An integrative MEG-fMRI study of the primary somatosensory cortex using cross-modal correspondence analysis. *Neuroimage* 22, 120–133. doi: 10.1016/j.neuroimage.2003.10.049

Schwarzkopf, D. S., Robertson, D. J., Song, C., Barnes, G. R., and Rees, G. (2012). The frequency of visually induced gamma-band oscillations depends on the size of early human visual cortex. *J. Neurosci.* 32, 1507–1512. doi: 10.1523/JNEUROSCI.4771-11.2012

Sepulcre, J., Masdeu, J. C., Pastor, M. A., Goni, J., Barbosa, C., Bejarano, B., et al. (2009). Brain pathways of verbal working memory: A lesion-function correlation study. *Neuroimage* 47, 773–778. doi: 10.1016/j.neuroimage.2009.04.054

Singh, K. D. (2012). Which "neural activity" do you mean? fMRI, MEG, oscillations and neurotransmitters. *Neuroimage* 62, 1121–1130. doi: 10.1016/j.neuroimage.2012.01.028

Stevenson, C. M., Wang, F., Brookes, M. J., Zumer, J. M., Francis, S. T., and Morris, P. G. (2012). Paired pulse depression in the somatosensory cortex: Associations between MEG and BOLD fMRI. *Neuroimage* 59, 2722–2732. doi: 10.1016/j.neuroimage.2011.10.037

Stippich, C., Freitag, P., Kassubek, J., Soros, P., Kamada, K., Kober, H., et al. (1998). Motor, somatosensory and auditory cortex localization by fMRI and MEG. *Neuroreport* 9, 1953–1957. doi: 10.1097/00001756-199806220-00007

Taulu, S., and Simola, J. (2006). Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Phys. Med. Biol.* 51, 1759–1768. doi: 10.1088/0031-9155/51/7/008

Thirion, B., Varoquaux, G., Dohmatob, E., and Poline, J. B. (2014). Which fMRI clustering gives good brain parcellations? *Front. Neurosci.* 8:167. doi: 10.3389/fnins.2014.00167

Vartiainen, J., Liljeström, M., Koskinen, M., Renvall, H., and Salmelin, R. (2011). Functional magnetic resonance imaging blood oxygenation level-dependent signal and magnetoencephalography evoked responses yield different neural functionality in reading. *J. Neurosci.* 31, 1048–1058. doi: 10.1523/JNEUROSCI.3113-10.2011

Wang, Y., and Holland, S. K. (2022). Bayesian MEG time courses with fMRI priors. *Brain Imaging Behav.* 16, 781–791. doi: 10.1007/s11682-021-00550-4

Warbrick, T. (2022). Simultaneous EEG-fMRI: What have we learned and what does the future hold? *Sensors* 22:2262. doi: 10.3390/s22062262

Ward, J. H. Jr. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58, 236–244. doi: 10.1080/01621459.1963.10500845

Whitman, J. C., Ward, L. M., and Woodward, T. S. (2013). Patterns of cortical oscillations organize neural activity into whole-brain functional networks evident in the fMRI BOLD signal. *Front. Hum. Neurosci.* 7:80. doi: 10.3389/fnhum.2013.00080

Whittingstall, K., Stroink, G., and Schmidt, M. (2007). Evaluating the spatial relationship of event-related potential and functional MRI sources in the primary visual cortex. *Hum. Brain Mapp.* 28, 134–142. doi: 10.1002/hbm.20265

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 23 December 2022
DOI 10.3389/fnins.2022.951321

Check for updates

OPEN ACCESS

EDITED BY
Yu Liu,
Hefei University of Technology, China

REVIEWED BY
Ahmad Mayeli,
University of Pittsburgh, United States
Marco Marino,
KU Leuven, Belgium

*CORRESPONDENCE
Erick Pasaye-Alcaraz
✉ pasayerick@inb.unam.mx

SPECIALTY SECTION
This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroscience

RECEIVED 23 May 2022
ACCEPTED 08 December 2022
PUBLISHED 23 December 2022

CITATION
Gallego-Rudolf J, Corsi-Cabrera M, Concha L, Ricardo-Garcell J and Pasaye-Alcaraz E (2022) Preservation of EEG spectral power features during simultaneous EEG-fMRI. *Front. Neurosci.* 16:951321. doi: 10.3389/fnins.2022.951321

Preservation of EEG spectral power features during simultaneous EEG-fMRI

Jonathan Gallego-Rudolf[1], María Corsi-Cabrera[2,3], Luis Concha[4], Josefina Ricardo-Garcell[3] and Erick Pasaye-Alcaraz[1]*

[1]Unidad de Resonancia Magnética, Instituto de Neurobiología, Universidad Nacional Autónoma de México, Querétaro, Mexico, [2]Laboratorio de Sueño, Facultad de Psicología, Universidad Nacional Autónoma de México, Mexico City, Mexico, [3]Unidad de Neurodesarrollo, Instituto de Neurobiología, Universidad Nacional Autónoma de México, Santiago de Querétaro, Mexico, [4]Laboratorio de Conectividad Cerebral, Instituto de Neurobiología, Universidad Nacional Autónoma de México, Querétaro, Mexico

Introduction: Electroencephalographic (EEG) data quality is severely compromised when recorded inside the magnetic resonance (MR) environment. Here we characterized the impact of the ballistocardiographic (BCG) artifact on resting-state EEG spectral properties and compared the effectiveness of seven common BCG correction methods to preserve EEG spectral features. We also assessed if these methods retained posterior alpha power reactivity to an eyes closure-opening (EC-EO) task and compared the results from EEG-informed fMRI analysis using different BCG correction approaches.

Method: Electroencephalographic data from 20 healthy young adults were recorded outside the MR environment and during simultaneous fMRI acquisition. The gradient artifact was effectively removed from EEG-fMRI acquisitions using Average Artifact Subtraction (AAS). The BCG artifact was corrected with seven methods: AAS, Optimal Basis Set (OBS), Independent Component Analysis (ICA), OBS followed by ICA, AAS followed by ICA, PROJIC-AAS and PROJIC-OBS. EEG signal preservation was assessed by comparing the spectral power of traditional frequency bands from the corrected rs-EEG-fMRI data with the data recorded outside the scanner. We then assessed the preservation of posterior alpha functional reactivity by computing the ratio between the EC and EO conditions during the EC-EO task. EEG-informed fMRI analysis of the EC-EO task was performed using alpha power-derived BOLD signal predictors obtained from the EEG signals corrected with different methods.

Results: The BCG artifact caused significant distortions (increased absolute power, altered relative power) across all frequency bands. Artifact residuals/signal losses were present after applying all correction methods. The EEG reactivity to the EC-EO task was better preserved with ICA-based correction approaches, particularly when using ICA feature extraction to isolate alpha power fluctuations, which allowed to accurately predict hemodynamic signal fluctuations during the EEG-informed fMRI analysis.

Discussion: Current software solutions for the BCG artifact problem offer limited efficiency to preserve the EEG spectral power properties using this particular EEG setup. The state-of-the-art approaches tested here can be further refined and should be combined with hardware implementations to better preserve EEG signal properties during simultaneous EEG-fMRI. Existing and novel BCG artifact correction methods should be validated by evaluating signal preservation of both ERPs and spontaneous EEG spectral power.

KEYWORDS

simultaneous EEG-fMRI, ballistocardiographic artifact, EEG signal preservation, EEG-informed fMRI, EEG spectral analysis

1 Introduction

Simultaneous Electroencephalography and functional Magnetic Resonance Imaging (EEG-fMRI) records the electrophysiological and hemodynamic correlates of human brain activity non-invasively, aiming to combine the strengths of both modalities (Huster et al., 2012; Laufs, 2012; Scrivener, 2021; Ebrahimzadeh et al., 2022). EEG measures the sum of extracellular currents generated by the synchronous activity of large populations of neurons, using electrodes attached to the subject's scalp (Schomer and da Silva, 2011), while fMRI quantifies changes in cerebral blood oxygenation, blood flow and blood volume that result from neurovascular coupling responses mediated by astrocytes, blood vessels, and neurons and therefore, represents an indirect correlate of neuronal activity (Huettel et al., 2004; Figley and Stroman, 2011). Simultaneous EEG-fMRI recording aims to better understand the complex dynamics underlying brain function by combining EEG's temporal resolution and fMRI's spatial resolution (Mulert and Lemieux, 2010; Scrivener, 2021; Ebrahimzadeh et al., 2022). Simultaneous EEG-fMRI also opens the possibility of directly studying interactions between electrophysiological and hemodynamic responses (Jorge et al., 2014) which cannot be achieved when signals are recorded independently (Mulert and Lemieux, 2010; Jorge et al., 2014; Ebrahimzadeh et al., 2022).

The major challenge of simultaneous EEG-fMRI recording is the presence of artifacts that compromise the data quality of both modalities (Ives et al., 1993; Mulert and Lemieux, 2010). EEG hardware may produce distortions and signal loss in MRI due to electromagnetic noise generated by the EEG amplifier (Krakow et al., 2000), B0 field inhomogeneities produced by magnetic susceptibility of EEG electrodes (Krakow et al., 2000; Mullinger et al., 2008), and B1 field attenuation produced by EEG leads (Luo and Glover, 2011; Klein et al., 2015). Image distortions can be avoided by using adequate electrode materials and placing the EEG amplifier inside a radiofrequency containment system, efficiently preserving MRI data quality at 3T (Krakow et al., 2000; Mullinger et al., 2008; Laufs, 2012). On the other hand, the magnetic resonance (MR) environment severely compromises EEG data quality (Ives et al., 1993; Lemieux et al., 1999). Two main artifacts contaminate the EEG data during simultaneous EEG-fMRI recordings: The gradient artifact (GA) and the ballistocardiographic (BCG) artifact. The GA results from induced currents over the EEG electrodes and leads that are produced by magnetic flux variations due to the gradients switching during image acquisition (Allen et al., 2000). Since the properties of the GA depend entirely on the MR sequence, these are highly stable over time and across individuals. Therefore, Average Artifact Subtraction (AAS) approaches (Allen et al., 2000) combined with hardware synchronization between EEG and fMRI equipment (Mandelkow et al., 2006) have proven to be effective in completely removing the GA.

The BCG artifact is a large-amplitude artifact that results from the induced currents caused by cardiac related movement of the EEG sensors when the subject is inside a strong magnetic field (Allen et al., 1998; Yan et al., 2010). The largest BCG artifact peak is typically observed ∼200 milliseconds after the QRS-wave recorded on the electrocardiogram (Allen et al., 1998). The artifact mainly spans cardiac harmonic frequencies ranging between 1 and 15 Hz, overlapping with the frequency of neural oscillations captured by EEG

Abbreviations: EEG-fMRI, simultaneous EEG-fMRI; GA, gradient artifact; BCG, ballistocardiographic; AAS, average artifact subtraction; OBS, Optimal Basis Set; ICA, Independent Component Analysis; PROJIC, PROJection onto Independent Components; IFE, Independent component analysis Feature Extraction; EC-EO, eyes closure-opening; rs, resting-state.

(Debener et al., 2008). Given its large variability between and within individuals, the BCG artifact represents a major challenge for EEG-fMRI. Several signal processing tools have been developed to deal with the BCG artifact and reduce its contribution from the recordings while preserving EEG signal properties (Bullock et al., 2021; Ebrahimzadeh et al., 2022). As summarized in Bullock et al. (2021) the most popular BCG correction approaches include Average Artifact Subtraction (AAS); (Allen et al., 1998), Optimal Basis Set (OBS); (Niazy et al., 2005), Independent Component Analysis (ICA); (Srivastava et al., 2005) and the combination of these: OBS-ICA (Debener et al., 2006) and AAS-ICA (Mayeli et al., 2021). Some other methods have also been recently proposed including the PROJIC-AAS and PROJIC-OBS methods (Abreu et al., 2016). Even though there have been studies comparing these methods, most of such studies have been based on assessing artifact reduction by comparing the amplitude of the artifact waveform (Mullinger et al., 2013b; Marino et al., 2018) or the reduction of its spectral components from the EEG signals before and after applying artifact correction (Abreu et al., 2016; Bullock et al., 2021). Although assessing artifact removal is important when validating BCG correction methods, the ultimate goal is to preserve the integrity of the functional properties of EEG signals. However, there are actually less studies focusing on signal preservation than artifact reduction (Marino et al., 2018; Bullock et al., 2021). Importantly, most of these studies have focused on evaluating the preservation of event related responses obtained from task paradigms (Debener et al., 2006; Assecondi et al., 2010; Vanderperren et al., 2010), where the high number of epochs used for averaging significantly increases the signal-to-noise ratio of the signals, as compared to continuous recordings (Schomer and da Silva, 2011). With a growing number of alternatives proposed to deal with the BCG artifact, there is a tremendous need to evaluate the efficiency of these methods to also preserve the spectral properties of spontaneous EEG oscillations recorded during resting-state and task paradigms, and to address the impact of BCG artifact residuals/EEG signal loss on multimodal data analysis results (Marino et al., 2019; Bullock et al., 2021). Therefore, the aim of this work was to characterize the impact of the BCG artifact on spontaneous EEG spectral power and to compare the effectiveness of the most commonly used BCG correction methods to remove the artifact while preserving underlying EEG signals. Specifically, we evaluated the spectral profile of resting-state EEG signals recorded during EEG-fMRI before and after BCG artifact correction, as compared to the spectral power of the EEG data recorded outside the scanner. We then assessed whether the functional reactivity of posterior EEG alpha power to a simple eyes closure-opening task was preserved after BCG removal and evaluated how the choice of BCG correction method affected the results from EEG-informed fMRI analysis.

2 Materials and methods

2.1 Participants

EEG and MRI data were collected from 20 healthy male individuals (mean age = 26 years; SD = 3.8 years) who were all graduate students from the Universidad Nacional Autónoma de México, campus Juriquilla (UNAM) community. Before enrolling participants into the study, the research protocol was explained to them both verbally and *via* an informed consent form. A psychologist with experience applying neuropsychological tests (JG) administered the Spanish version of the MINI International Neuropsychiatric Interview (Sheehan et al., 1997; Ferrando et al., 1998). Only cognitively healthy individuals who did not have diagnosis of any neurological or psychiatric disease or history of substance abuse were invited to participate in the study. As a last filter, participants were asked to fill in a brief checklist to corroborate the presence of counter-indications to perform the MR protocol. Individuals that fulfilled the requirements to be included in the sample and agreed to participate in the experiment signed the consent form and were recruited for the study. This research project was conducted in accordance with the principles of the Declaration of Helsinki for experiments involving human participants and was approved by the Bioethics Committee of the Instituto de Neurobiología, UNAM.

2.2 EEG data acquisition

Both EEG and MRI data were acquired in a single session lasting around 2.5 h. EEG data were recorded using a GES 400 MR system equipped with a 32-channel MR-compatible EEG cap (Electrical Geodesics Inc., Eugene, OR, USA). The sampling rate was 1000 Hz and Cz was used as the reference electrode. Electrode impedances were measured before starting the outside EEG recordings and all sensors were adjusted to keep impedance values below 50 k-ohms. A silk mesh was placed over the electrode cap and bandages were used to reduce electrode movement and improve EEG data quality (Ives et al., 1993; Bénar et al., 2003). Electrocardiogram (ECG) data was recorded using MR-compatible patch electrodes. The active electrode was placed over the heart (slightly to the left of the sternum bone) and the reference electrode was placed over the medial end of the left collarbone.

Electroencephalographic data were first recorded outside the MR environment, with the participant lying down in supine position, same as inside the MR-scanner. The outside EEG recording protocol consisted of 2 min of eyes-closed resting-state (Outside rs-EEG), 2 min of eyes-open EEG (not used in this study) and 2 min of an eyes closure-opening task (Outside EEG EC-EO) consisting of 20-s blocks, starting with eyes-closed. After the outside EEG recording, the participant was taken into the MR room. Once the participant was inside the scanner, EEG leads were carefully examined in search for loops and oriented in a straight line parallel to the B0 magnetic field, to reduce EEG artifacts and the risk of radiofrequency-induced heating of the sensors (Yan et al., 2009; Chowdhury et al., 2015; Assecondi et al., 2016). Sandbags and tape were used to minimize electrode leads movement and soft pads were placed between the receiving coil and the subject's head to reduce participant's movement (Bénar et al., 2003; Bullock et al., 2021; Ebrahimzadeh et al., 2022). The EEG amplifier was placed next to the bed of the scanner behind the 400 Gauss magnetic field iso-intensity line, in accordance with the safety guidelines provided by the vendor. Lights and ventilation systems were turned off during the entire session, to avoid further artifacts in the EEG signal (Mullinger et al., 2013a; Nierhaus et al., 2013; Rothlübbers et al., 2015). Due to our facility regulation protocols, the helium cooling pump remained active for all of the recordings.

After ensuring the participant was feeling comfortable inside the scanner, we recorded 2 min of eyes-closed EEG (Inside rs-EEG) without image acquisition. We then began the simultaneous EEG-fMRI protocol, which consisted of 10 min of eyes-closed resting-state (rs-EEG-fMRI) and 4 min of the eyes closure-opening (EEG-fMRI EC-EO) task.

2.3 MRI data acquisition

Brain magnetic resonance images were obtained with a Discovery MR750 3.0T scanner (General Electric, Milwaukee, WI, USA), equipped with a 32-channel array head coil. Blood-oxygen level-dependent (BOLD) contrast images were acquired for the rs-EEG-fMRI and EEG-fMRI EC-EO conditions using an echo-planar reconstruction (spatial resolution = 4 × 4 × 4 mm^3 voxels, TR = 2000 ms, TE = 40 ms). High resolution structural sagittal T1-weighted images (spoiled gradient-recalled sequence; resolution of 1 × 1 × 1 mm^3 voxels; TR = 8.1 ms; and TE = 3.2 ms) were collected following the simultaneous EEG-fMRI recordings, after the EEG cap was removed from the participant's head.

2.4 EEG preprocessing and BCG artifact removal

The Outside rs-EEG, Inside rs-EEG, rs-EEG-fMRI, Outside EEG EC-EO and EEG-fMRI EC-EO data were preprocessed separately following the same pipeline, with the exception of the artifact removal steps that were added to correct the gradient (rs-EEG-fMRI, EEG-fMRI EC-EO) and the BCG (inside rs-EEG, rs-EEG-fMRI, EEG-fMRI EC-EO) artifacts from the data acquired inside the MR-environment.

The GA removal was the first preprocessing step for the rs-EEG-fMRI and the EEG-fMRI EC-EO datasets and was implemented directly in the Net Station software (Electrical Geodesics Inc., Eugene, OR, USA). We applied AAS by averaging the signals aligned with the event markers automatically generated by the hardware synchronization between the EEG amplifier and the MR scanner clock, using a sliding-window consisting of 5 TR volumes to generate the template. The rest of the preprocessing for all datasets was performed using customized scripts calling EEGLAB (Delorme and Makeig, 2004) and MATLAB (The MathWorks, Inc., Natick, MA, USA) functions. EEG data from all conditions (.mff files) were imported into MATLAB following the EEGLAB data structure by using the MFFmatlabIO plugin. Only data from the eighteen 10–20 system electrodes (Fp1, Fp2, F3, F4, F7, F8, Fz, C3, C4, P3, P4, Pz, T3, T4, T5, T6, O1, O2; Cz was used as reference) were considered for the analysis. Channel locations were set using the corresponding Geodesic Sensor Net template from EEGLAB. Continuous EEG data were band-pass filtered (1–50 Hz) and then segmented into 2-s epochs. EEG signals were visually inspected, and epochs containing high amplitude artifacts related to the subject's movements or blinks were rejected. Additionally, in the case of the Inside rs-EEG, rs-EEG-fMRI and EEG-fMRI EC-EO datasets we corrected the BCG artifact using one of seven methods: (1) AAS, (2) OBS, (3) ICA, (4) OBS followed by ICA, (5) AAS followed by ICA, (6) PROJIC-AAS, and (7) PROJIC-OBS. The detection of QRS peaks and the implementation of the AAS and OBS-based correction approaches (Iannetti et al., 2005; Niazy et al., 2005) were performed using the EEGLAB FMRIB plug-in provided by the University of Oxford Centre for Functional MRI of the Brain. A constant delay of 210 milliseconds between the cardiac event markers and the main BCG peak was assumed for all AAS and OBS-based methods, which is the default value in the FMRIB plugin (Allen et al., 1998). For OBS-based corrections, the four principal components that explained most of the artifact's waveform variance were automatically selected and regressed-out from the data. ICA was implemented using EEGLAB's *runica* algorithm. A variable number of artifact-related independent components (ICs) were manually selected for each subject, based on criteria suggested in previous studies (Srivastava et al., 2005; Debener et al., 2006, 2008; Iannotti et al., 2014). Specifically, we removed ICs displaying all three of the following features: (1) time-series with rhythmic peaks that followed the ECG trace, (2) increased power showing multiple peaks at cardiac-related frequencies, and (3) topographical distribution of power showing either left-right or anterior-posterior polarity inversion. PROJIC-AAS and PROJIC-OBS methods were implemented using the code provided by Abreu et al. (2016). Both methods rely on applying the same functions from the FMRIB plugin, but in this case the AAS and OBS corrections are applied on the ICs timeseries before retrieving the original EEG time series by multiplying the EEG activations * mixing matrix W^{-1}, in contrast to applying the correction directly on the sensor timeseries as with the regular AAS and OBS approaches. For both PROJIC approaches, we used the recommended default parameters, with the only major difference that we used the same ICs we manually selected for the ICA approach, rather than using the PROJIC-ICA automatic selection of the BCG-related components (which failed to accurately identify the BGC-related ICs for many subjects).

2.5 Data analysis

2.5.1 Eyes-closed resting-state EEG

To evaluate the impact of the BCG artifact on EEG spectral power and to test if resting-state EEG spectral properties could be preserved after artifact correction, we compared the absolute and relative power from the corrected rs-EEG-fMRI signals vs. the Outside rs-EEG. The first available twenty-two (minimal number of clean epochs available per subject), 2-s non-overlapping clean epochs from the Outside rs-EEG and rs-EEG-fMRI conditions for each subject were selected for quantitative analysis. We computed the fast Fourier transform of the signals and then calculated the absolute and relative power across traditional EEG frequency bands (Schomer and da Silva, 2011), defined as follows: Delta = 1–3 Hz, Theta = 4–7 Hz, Alpha = 8–12 Hz, Slow beta = 13–17 Hz, Fast beta = 18–30 Hz, and Gamma = 31–50 Hz.

To obtain a qualitative measure of the BCG artifact contribution to each frequency band and visualize the variability across subjects before and after artifact correction, we calculated the percentage change in absolute power from the rs-EEG-fMRI relative to the outside rs-EEG [(rs-EEG-fMRI power/outside rs-EEG power) *100] −100, for each electrode of each subject. For the statistical analysis, we used one-way repeated measures ANOVAs to compare the log-transformed absolute power and the relative power values of the corrected rs-EEG-fMRI and the Outside rs-EEG. Each frequency band was considered independently. Bonferroni correction for multiple comparisons was applied to the p-values of the *post-hoc* test between the Outside rs-EEG and the seven corrected versions of the rs-EEG-fMRI data. Adjusted p-vales below 0.05 were considered to be significant. To discard the contribution of GA residuals and further validate our results, we repeated our analysis using the Inside rs-EEG instead of the rs-EEG-fMRI data (**Supplementary material**).

In addition to the absolute and relative spectral power analysis, we tested the reliability of the estimates of the individual alpha peak frequency and alpha center of gravity from the power spectrum of the resting-state signals collected during EEG-fMRI. Following the methods and using the code provided by Corcoran et al. (2018), we employed an automated approach based on applying a Savitzky–Golay filter (Klimesch et al., 1990) to calculate each individual's alpha peak frequency and center of gravity. We set the band-pass filter from 1 to 40 Hz and looked for the alpha peak in the range between 7 and 13 Hz. We set a value of 11 for the Savitzky–Golay filter frame width and a $k = 5$ for the polynomial order. For the statistical comparison, we used one-way repeated measures ANOVAs to compare the individual alpha peak frequency and center of gravity estimates from the corrected rs-EEG-fMRI and the Outside rs-EEG (**Supplementary material**). We applied Bonferroni correction to account for multiple comparisons and considered adjusted p-vales below 0.05 to be significant. Once again, we repeated our analysis using the Inside rs-EEG instead of the rs-EEG-fMRI data (**Supplementary material**).

2.5.2 Eyes closure-opening task EEG data

Given that the posterior alpha power reactivity to the eyes closure-opening task is one of the most prominent and consistent features observed in human EEG recordings (Berger, 1929; Barry et al., 2007; Klimesch et al., 2007; Barry and De Blasio, 2017), we selected this task to evaluate if alpha power functional reactivity was preserved in the EEG-fMRI signals corrected using different BCG removal approaches. The first available twenty, 2-s non-overlapping EEG epochs from the eyes-closed and eyes-open blocks of the Outside EEG EC-EO and the EEG-fMRI EC-EO conditions were selected for each subject and submitted to fast Fourier transform, as implemented previously. Besides the seven BCG correction methods used before, an eighth method consisted of using ICA as a feature extraction tool (IFE), aiming to isolate alpha power activity related to the task. In this case, instead of removing the ICs associated with the BCG artifact we only retained components with a time-series that showed clear alpha activity during the EC blocks, a peak around 10 Hz in its power spectrum, and a topographical distribution showing higher alpha power in posterior electrodes. To estimate a quantitative difference between the two physiological states, we calculated a ratio by using the signal from occipital O1 and O2 electrodes and dividing the alpha power of the EC over the EO condition (EC-EO alpha power ratio). We performed the statistical analysis on the EC-EO alpha power ratio values rather than the raw eyes-open and eyes-closed alpha power values given that absolute and relative alpha power is highly variable across individuals (Shaw, 2003). To assess signal preservation, a one-way repeated measures ANOVAs was performed to compare the EC-EO alpha power ratio between the Outside EEG EC-EO and the EEG-fMRI EC-EO corrected with different methods (AAS, OBS, ICA, OBS-ICA, AAS-ICA, PROJIC-AAS, PROJIC-OBS, and IFE). As before, Bonferroni-correction for multiple comparisons was applied to the p-values of the *post hoc* comparisons, and adjusted p-values > 0.05 were considered as significant.

2.5.3 EEG-informed fMRI analysis

Functional magnetic resonance imaging data preprocessing and analysis was performed using the FSL software (Jenkinson et al., 2012). Preprocessing included motion correction, slice timing (interleaved acquisition) correction, brain extraction, spatial smoothing using a Gaussian kernel (full-width-half-maximum = 6 mm), high-pass temporal filtering (cutoff frequency = 0.01 Hz), and registration to each subject's structural image followed by spatial normalization to the Montreal Neurological Institute standard space (MNI ICBM-152 template) using linear transformations with seven and twelve degrees of freedom, respectively. EEG-informed fMRI first-level analysis was performed using alpha power fluctuations to derive a BOLD signal predictor for each subject. To generate the predictors, we selected either the O1 or O2 channel timeseries (selected on an individual basis to obtain the best available predictor) and calculated the alpha band absolute power for each 2-s epoch. The values corresponding to epochs that were eliminated due to excessive movement or eye-related artifacts were replaced using a simple interpolation method (taking the average of the previous and following epoch). The resulting time series (60 timepoints) were convolved with a gamma hemodynamic response function in the GLM tool of FSL's FEAT to generate the BOLD signal predictors. We focused on the negative contrast, as our hypothesis was that alpha power fluctuations would be negatively correlated with BOLD signal. We first conducted this analysis on the full sample (n = 20), though no significant associations were found between the predictors and the BOLD signal using any method, due to some individuals that did not show any associations between the signals in the first-level analysis. We therefore repeated the analysis after removing these 5 individuals, which corresponds to the data presented here.

To assess the impact of BCG artifact residuals on preserving the EEG functional reactivity for multimodal integration, the EEG-informed fMRI analysis was repeated using the alpha power predictors derived from the same EEG signals, corrected using each method. Second-level analyses were performed using permutation-based inference (Nichols and Holmes, 2002) with threshold-free cluster enhancement to account for multiple comparisons (Smith and Nichols, 2009) as implemented by FSL's randomize function. Group-level statistical parametric maps obtained from the EEG-informed fMRI analyses were compared with conventional fMRI analysis results, performed using the task design to build the hemodynamic response predictor.

2.6 Data/code availability statement

All the data used in this study is available on an open data repository: "Simultaneous EEG-fMRI dataset," Mendeley Data, V1, doi: 10.17632/crhybxpdy6.1 (Gallego-Rudolf et al., 2022). All the preprocessing and analysis steps in this study used a combination of existing documented functions from MATLAB v.18b (The MathWorks, Inc., Natick, MA, USA) and EEGLAB v.14.1.2b (Delorme and Makeig, 2004) software packages. EEG-informed fMRI was conducted using the FSL software (Jenkinson et al., 2012). Statistical analysis was performed in R Studio, using R v.4.1.1 (R Core Team, 2022) and ggplot2 (Wickham, 2016) was used to generate the plots.

3 Results

3.1 Resting-state–BCG artifact reduction and preservation of EEG spectral features

Given that the GA is entirely dependent on the properties of the MR sequence, having an adequate synchronization between the EEG and MRI hardware and using the AAS approach allowed to effectively remove the GA from the signals of all participants. The first panel of **Figure 1** shows the comparison between the average power spectrum across all electrodes from all subjects from the Outside rs-EEG (black), the rs-EEG-fMRI data before GA correction (blue) and the rs-EEG-fMRI data after GA and before BCG artifact correction (red). The rest of the panels from **Figure 1** show the group average power spectra for the Outside rs-EEG (black, same for all panels), the rs-EEG-fMRI before BCG artifact correction (red, same for all panels) and its corrected version (green) using each BCG correction approach. The main contribution of the BCG artifact to the power spectrum can be observed as a generalized increase in spectral power, more pronounced in the theta and slow beta range (red power spectrum). In general ICA-based approaches (ICA, but specially OBS-ICA and AAS-ICA) performed better in reducing the BCG-induced absolute power increases, partially retrieving the characteristic shape of the eyes-closed rs-EEG spectrum. The number of components removed for each method (mean; SD; range) was 9.2; 1.8; 6–12 for ICA, 4.8; 1.2; 3–7 for OBS-ICA and 5.7; 1.4; 3–8 for AAS-ICA. **Supplementary Figure 1** shows that similar results were obtained when calculating the power spectrum from the Inside rs-EEG, instead of the rs-EEG-fMRI data.

Figure 2 shows the percentage change in power of the rs-EEG-fMRI relative to the Outside rs-EEG, after correcting the rs-EEG-fMRI signal with different BCG removal approaches. Each matrix corresponds to a particular method and frequency band and shows all electrodes (rows) for every subject (columns). As can be observed, a considerable increase in power across all bands was present in the uncorrected rs-EEG-fMRI data, with theta and slow beta being the most affected frequency bands. Again, ICA, OBS-ICA and AAS-ICA correction methods were more efficient in suppressing the contribution of the BCG artifact, by reducing the BCG-induced power increase especially in the delta, theta, and alpha bands. However, even when using ICA–based corrections, artifact residuals remained for most subjects, particularly in the fast beta and gamma bands.

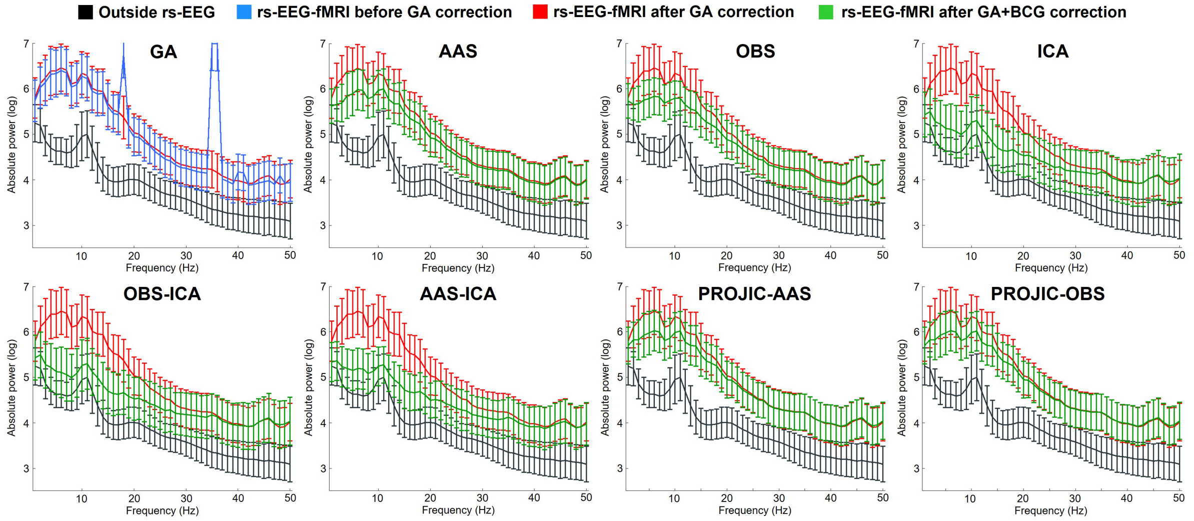


FIGURE 1
Average power spectrum (and standard deviation) computed from the resting-state eyes-closed EEG signal of all electrodes from all subjects. The first panel shows the Outside rs-EEG (black) and the rs-EEG-fMRI data before (blue) and after (red) GA correction. The rest of the panels show a comparison between the Outside rs-EEG (black line, repeated in all panels), the rs-EEG-fMRI after GA removal but before BCG artifact correction (red line, repeated in all panels) and its corrected version (green) after using one of seven BCG correction methods: Average Artifact Subtraction (AAS), Optimal Basis Set (OBS), Independent Component Analysis (ICA), OBS-ICA, AAS-ICA, PROJection onto Independent Components (PROJIC)-AAS or PROJIC-OBS. ICA-based corrections performed better in reducing the BCG artifact contribution and preserving the spectral profile of rs-EEG-fMRI signals, though power remained higher compared to the Outside rs-EEG.

Moreover, we also observed decreases in power compared to the outside rs-EEG, reflecting potential EEG signal losses produced during artifact correction.

The statistical analysis comparing the absolute power across the six frequency bands is shown in Figure 3. Significant statistical differences were found between the power of the Outside rs-EEG and the rs-EEG-fMRI for all frequency bands (delta $F_{3.46,65.67} = 61.34$, $p < 0.001$; theta $F_{3,56.98} = 165.68$, $p < 0.001$; alpha $F_{2.89,54.87} = 141.33$, $p < 0.001$; slow beta $F_{2.97,56.36} = 239.32$, $p < 0.001$; fast beta $F_{2.98,56.55} = 96.57$, $p < 0.001$; gamma $F_{2.58,48.94} = 82.62$, $p < 0.001$), regardless of the BCG correction method employed. Very similar results were obtained for the Inside rs-EEG data (Supplementary Figures 2, 3).

Even though the BCG-induced power increase across frequency bands remained significant after artifact correction, qualitatively the rs-EEG-fMRI data showed that the individual power estimates computed after applying ICA-based correction approaches displayed a more similar distribution compared to the Outside rs-EEG values. Therefore, we also analyzed the relative power of each frequency band and compared the Outside rs-EEG vs. the corrected versions of the rs-EEG-fMRI data (Figure 4). Delta relative power was significantly decreased, while slow beta remained significantly increased across all correction methods ($F_{2.9,55.18} = 79.33$, $p < 0.001$; $F_{2.48,47.11} = 65.84$, $p < 0.001$). Theta relative power from the ICA, OBS-ICA and AAS-ICA approaches was not significantly different compared the Outside rs-EEG, which was the case for all other methods ($F_{2.79,52.93} = 32.11$, $p < 0.001$), but in contrast only these approaches showed significant differences in the alpha relative power compared to the Outside rs-EEG ($F_{1.92,36.49} = 14.16$, $p < 0.001$). The only method in which fast beta relative power was different from the Outside rs-EEG was ICA ($F_{3.03,57.56} = 17.51$, $p < 0.001$) and for gamma relative power there were significant increases observed in the ICA, OBS-ICA, and AAS-ICA approaches ($F_{2.69,51.17} = 62.02$, $p < 0.001$). Once again, we found very similar results when using the Inside rs-EEG data (Supplementary Figure 4).

The analysis of the individual alpha peak frequency and center of gravity revealed that, even with fewer electrodes with sufficient quality for the estimation of these parameters in the rs-EEG-fMRI condition (Supplementary Table 1), there were no significant differences in the estimations of the alpha peak frequency and center of gravity when comparing the corrected (and uncorrected) rs-EEG-fMRI against the Outside rs-EEG data (Supplementary Figure 5).

3.2 Eyes closure-opening task–preservation of EEG functional reactivity

We then focused on evaluating if EEG functional reactivity to the EC-EO task could be preserved after BCG artifact removal. Figure 5 shows the group average power spectrum

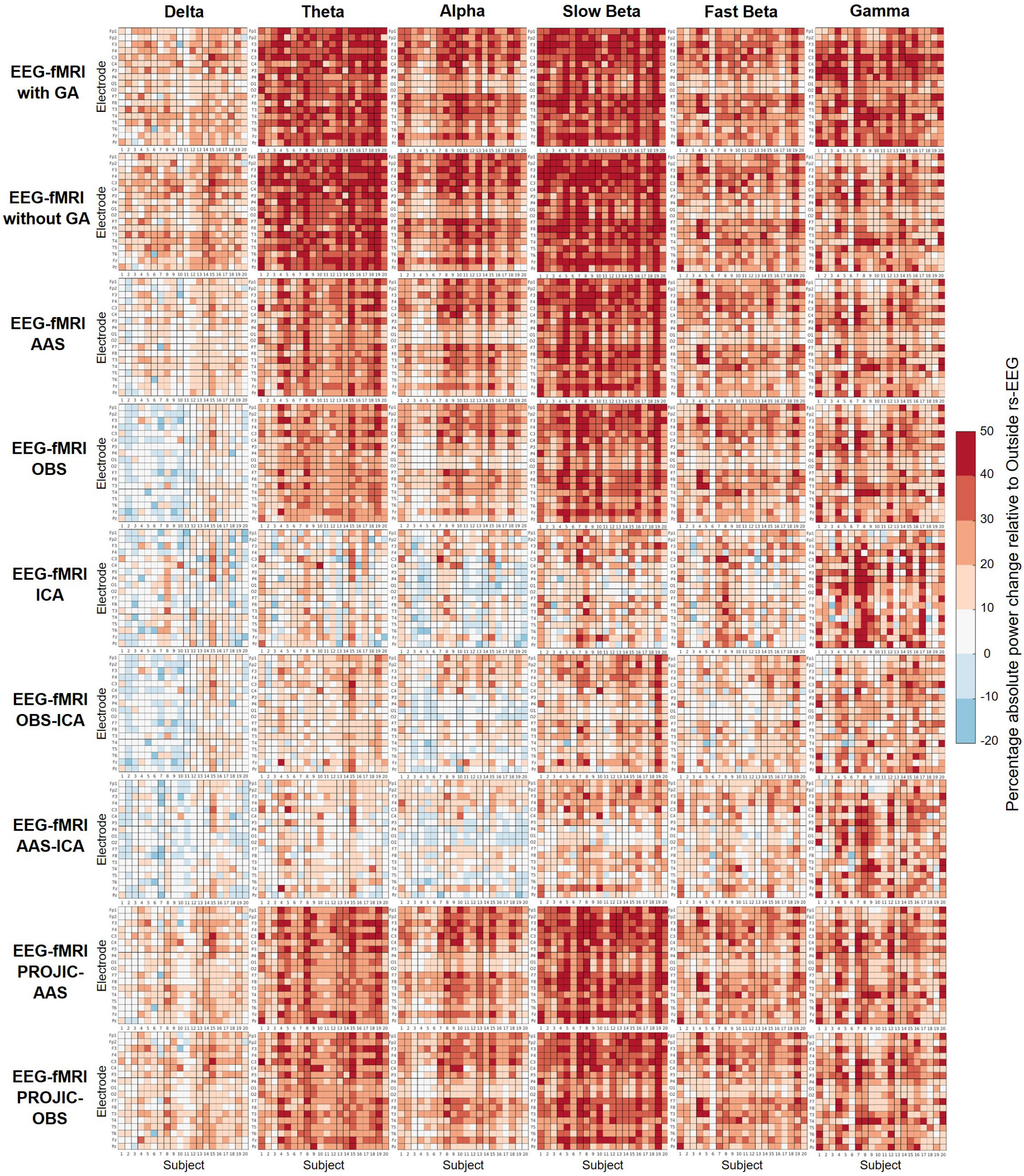


FIGURE 2
Percentage change (colorbar) in the absolute power of each frequency band of the rs-EEG-fMRI data before and after BCG artifact removal, relative to the Outside rs-EEG. Each matrix shows all electrodes (rows) for each subject (columns). A negative percentage indicates lower absolute power in the rs-EEG-fMRI compared to the outside rs-EEG. ICA-based corrections performed better in reducing the BCG artifact contribution and preserving the rs-EEG-fMRI spectral profile (especially for delta, theta, and alpha bands), though artifact residuals and/or absolute power decreases were evident for most subjects, across all frequency bands.

from the O1 electrode, obtained from the EEG signals collected during the eyes-closed (green) and eyes-open (red) conditions. For the Outside EEG EC-EO spectrum, a clear distinction between the two physiological states is observed as a higher-amplitude alpha power peak in the absolute power EC EEG spectrum, compared to the EO spectrum. This difference is completely masked by the BCG artifact. Although the difference between the two conditions was never as evident

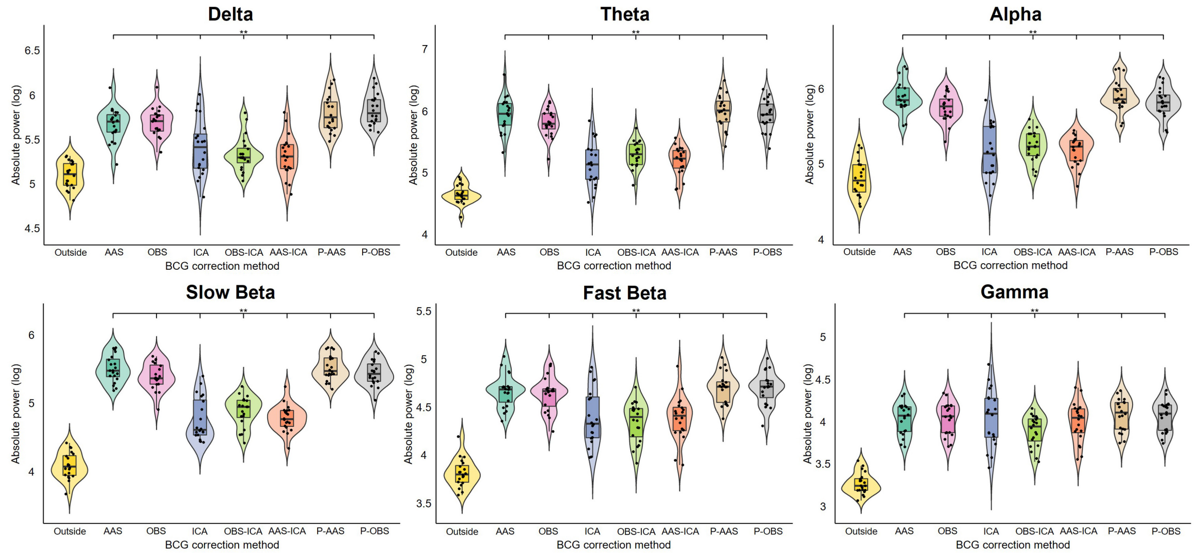


FIGURE 3
Results of the repeated measures ANOVAs comparing the average absolute power of all electrodes from all subjects between the Outside rs-EEG and the rs-EEG-fMRI data corrected using each of the seven BCG correction methods. Each frequency band was analyzed separately. The asterisks indicate significant statistical differences (p_{adj} < 0.05) between the corrected rs-EEG-fMRI and the Outside rs-EEG data. A generalized increase in absolute power across all frequency bands was observed for the data recorded simultaneously with fMRI, which remained significant after applying all BCG correction methods. Note that PROJIC-AAS and PROJIC-OBS were abbreviated as P-AAS and P-OBS, respectively.

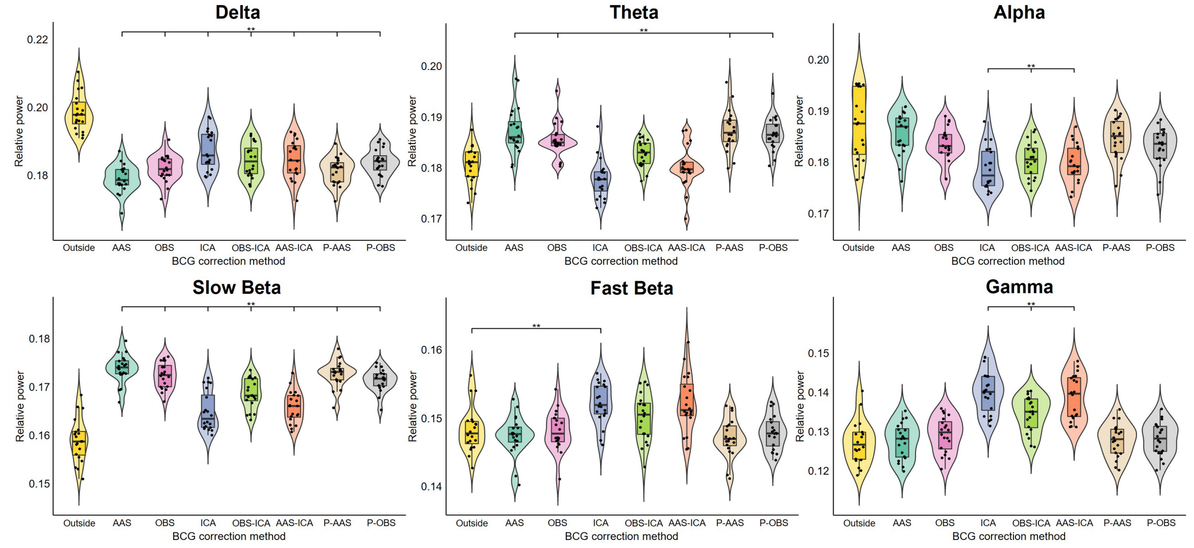


FIGURE 4
Results of the repeated measures ANOVAs comparing the average relative power of all electrodes from all subjects between the Outside rs-EEG and the rs-EEG-fMRI data corrected using each of the seven BCG correction methods. Each frequency band was analyzed separately. The asterisks indicate significant statistical differences (p_{adj} < 0.05) between the corrected rs-EEG-fMRI and the Outside rs-EEG data. Relative power was altered across all frequency bands for the data recorded simultaneously with fMRI. Some correction approaches rescued relative power for some frequency bands, but the overall spectral power profile remained altered across all BCG correction methods. Note that PROJIC-AAS and PROJIC-OBS were abbreviated as P-AAS and P-OBS, respectively.

as for the Outside EEG EC-EO data, the use of ICA, or the combination of OBS-ICA and AAS-ICA allowed to partially retrieve the difference between EC and EO states. The number of components removed per each method (mean; SD; range) was 9.6; 1.3; 8–11 for ICA, 6.7; 1.4; 4–10 for OBS-ICA and 5.8; 1.2; 4–8 for AAS-ICA. The rest of the correction approaches failed

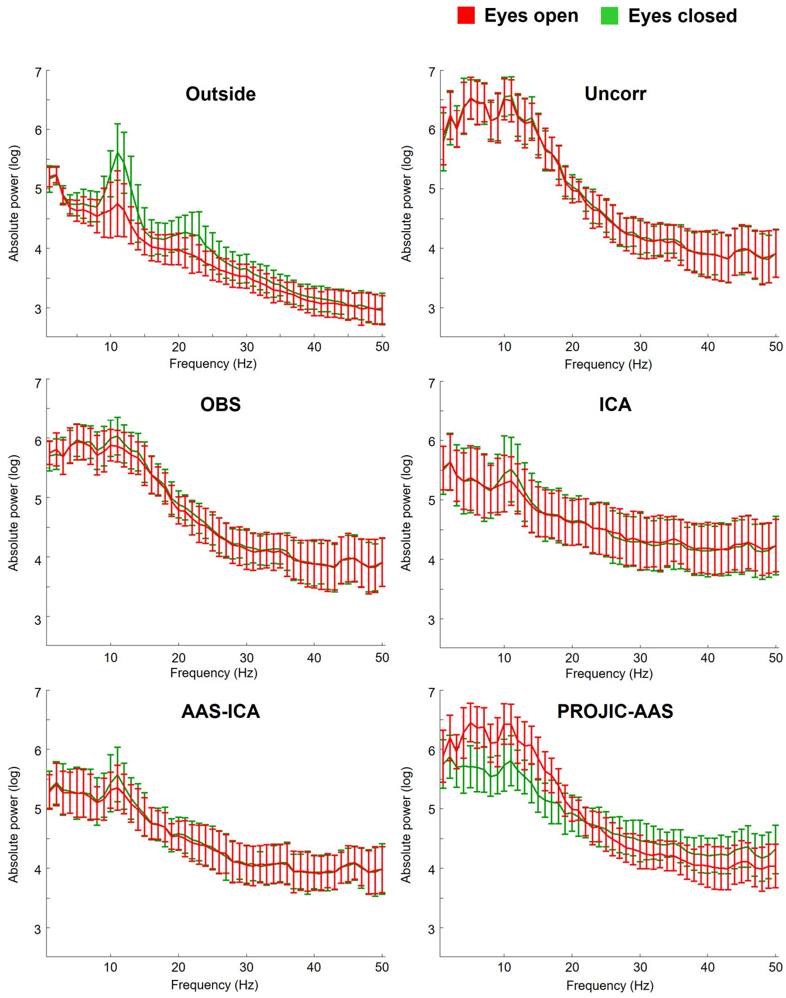


FIGURE 5
Group average power spectrum (with standard deviation) of the O1 electrode during the eyes-closed (EC; green) and eyes-open (EO; red) conditions of the EC-EO task. A comparison is shown between the spectra obtained from the Outside EEG EC-EO and the EEG-fMRI EC-EO data, before and after removing the BCG artifact with each correction method. ICA-based approaches performed better in retrieving the difference between EC and EO conditions (reflected as higher alpha power for the EC condition), though the difference was still attenuated when compared to the data acquired outside the scanner.

to retrieve a clear distinction in the alpha band between the two conditions.

This was confirmed by the statistical analysis shown in Figure 6, comparing the ratio obtained from dividing the alpha power of the EC condition by the alpha power of the EO condition. BCG artifact residuals resulted in a significant decrease in the alpha power ratio for all correction methods ($F_{2.77,38.75} = 35.52$, $p = > 0.001$). The only approach that allowed to retrieve an EC-EO power ratio that was not statistically different from the Outside EEG EC-EO data was the ICA feature extraction of the alpha power, indicating this strategy retrieved the functional reactivity of posterior alpha oscillations (Figure 6). The number of retained ICs related to alpha activity for the ICA feature extraction approach (mean; SD; range) was 2.1; 0.6; 1–3.

3.3 EEG-informed fMRI—impact of BCG artifact residuals on multimodal analysis

To evaluate if the BCG artifact biased multimodal data analysis results, we performed EEG-informed fMRI analysis using alpha power fluctuations derived from the EEG-fMRI EC-EO condition to generate the BOLD signal predictors. EEG predictors were obtained from the same EEG-fMRI EC-EO signals corrected with one of the eight approaches (AAS, OBS, ICA, OBS-ICA, AAS-ICA, PROJIC-AAS, PROJIC-ICA, and IFE). The results were compared to those obtained with conventional fMRI analysis, using the task design to build the hemodynamic response model (positive contrast). Figure 7

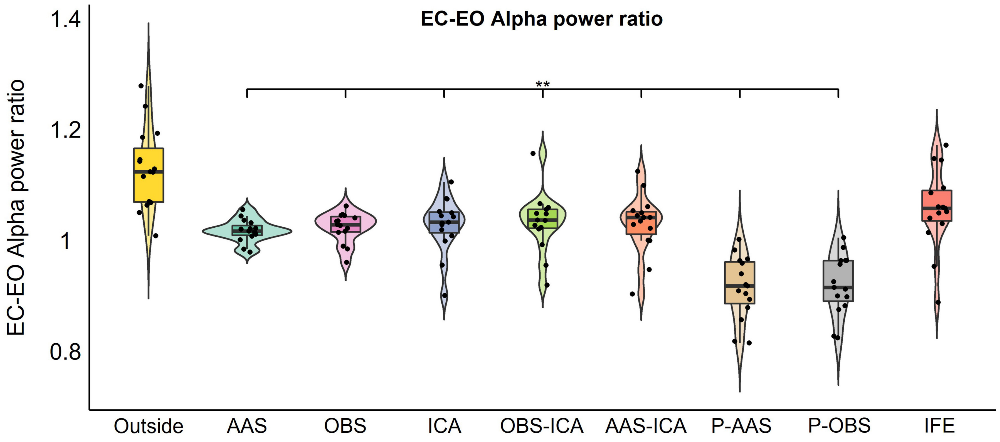


FIGURE 6
Results of the repeated measures ANOVAs comparing the eyes closure-opening (EC-EO) alpha power ratio calculated from O1 and O2 electrodes for the Outside EEG EC-EO and the EEG-fMRI EC-EO signals corrected with each BCG approach or Independent component analysis Feature Extraction (IFE). The asterisks indicate significant statistical differences (p_{adj} < 0.05) between the EEG-fMRI EC-EO and the Outside EEG EC-EO data. IFE was the only method that did not show significant differences in the EC-EO alpha power ratio when compared to the data recorded outside the scanner. Note that PROJIC-AAS and PROJIC-OBS were abbreviated as P-AAS and P-OBS, respectively.

displays the statistical parametric maps obtained using the task design and each of the EEG-derived predictors. As expected, in the task design predictor maps we observed BOLD signal increases during the EO condition and decreases during the EC condition within occipital and parietal cortices. BOLD signal changes were accurately predicted by EEG signals, as observed in the maps from the EEG-derived predictor generated after using IFE to extract alpha power fluctuations. Importantly, BCG residuals/signal loss that remained after implementing all the tested BCG correction approaches biased the results of the EEG-informed fMRI analysis, obscuring the associations between alpha power and BOLD signal fluctuations.

4 Discussion

In this study we aimed to characterize the impact of the BCG artifact on spontaneous EEG spectral power and to compare some of the most popular available BCG correction approaches. Our main focus was to assess the preservation of resting-state EEG spectral properties by statistically comparing the absolute and relative power changes in the EEG data simultaneously acquired with fMRI (corrected with different methods) with respect to the uncorrected data and the data obtained outside of the MR environment. We further investigated whether the functional information from EEG spectral power could be retrieved regardless of the presence of BCG artifact residuals, by evaluating the alpha power reactivity to an EC-EO task. Finally, we wanted to assess how the selection of the BCG artifact correction method influenced the results from EEG-informed fMRI analysis. Although several studies have previously compared different BCG correction methods to assess artifact reduction and signal preservation (Debener et al., 2006; Vanderperren et al., 2010; Marino et al., 2018; Bullock et al., 2021), ours is one of the few studies that: (1) Focus on the preservation of spontaneous brain oscillations rather than ERPs, (2) Characterize the impact of BCG artifact removal using seven state-of-the-art methods by using a specific task paradigm to test the functional reactivity of a particular spontaneous brain rhythm (alpha oscillations), and (3) Provide a direct side-by-side comparison of the impact of using different BCG correction approaches prior to multimodal EEG-informed fMRI analysis.

The uncorrected rs-EEG-fMRI showed a marked increase in absolute power across all frequency bands, more pronounced within the theta and slow beta bands. Relative power was also severely distorted, making uncorrected EEG signals unusable for any analysis purposes. We found that, even though a clear reduction of the artifact was observed on the power spectra of our rs-EEG-fMRI data, none of the BCG artifact removal approaches tested (AAS, OBS, ICA, OBS-ICA, AAS-ICA, PROJIC-AAS, PROJIC-OBS) entirely preserved the spectral profile of EEG signals, due to both artifact residuals and induced EEG signal losses. Overall, in line with previous reports (Srivastava et al., 2005; Debener et al., 2006; Mayeli et al., 2021), we found better results with ICA-based approaches, especially when used after AAS or OBS, as compared to the conventional AAS and OBS and PROJIC approaches. Additionally, large variability in the artifact correction outcomes was observed, with some subjects even showing decreased absolute power compared to their outside rs-EEG, which may be reflecting EEG signal losses after the artifact correction procedure (Ullsperger and Debener, 2010; Marino et al., 2018). To our surprise, the estimation of the individual alpha peak frequency and center of gravity

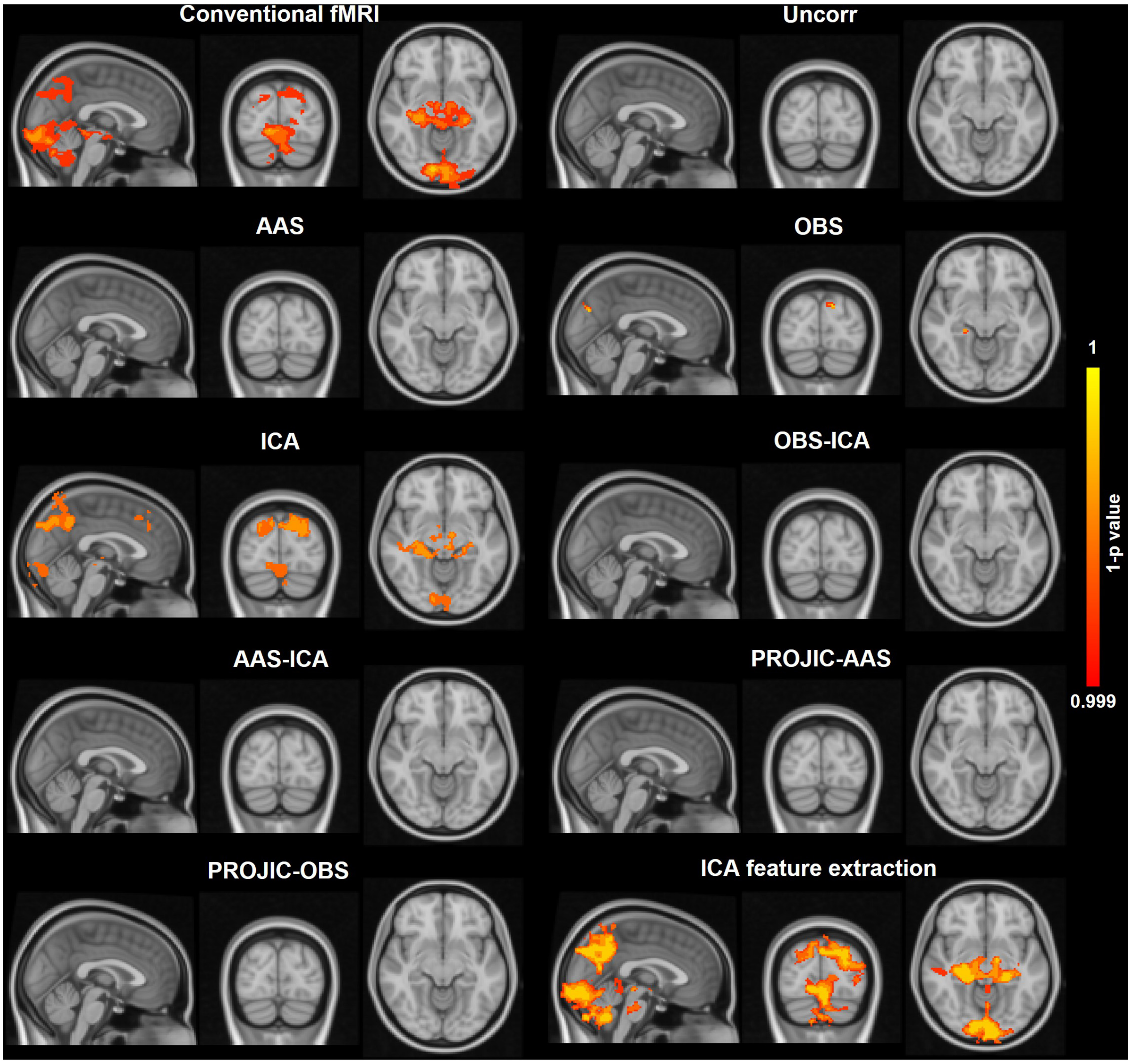


FIGURE 7
Corrected threshold-free cluster enhancement voxel-wise group-level statistical maps obtained from the EC-EO task fMRI data analysis (n = 15) using either the task design or the EEG alpha power fluctuations to generate the blood-oxygen level-dependent (BOLD) signal predictors used in the general linear model. For the conventional fMRI analysis (task design), the map shows the voxels that displayed a positive association with the model (higher BOLD signal in EO vs. EC conditions). For the EEG-informed fMRI analyses, the maps show the voxels where the BOLD signal exhibited a significant negative association with the EEG alpha power derived BOLD signal predictor. A comparison is shown between the maps obtained using the predictors derived from the same EEG signals, corrected using each BCG correction method or IFE. Only IFE preserved the negative relationship between alpha power fluctuations and the BOLD signal, providing very similar maps to those obtained from the conventional fMRI analysis. The color scale shows the 1-p statistical values with a threshold set at p < 001.

were preserved even in the uncorrected rs-EEG-fMRI data, suggesting that such features can be successfully extracted from EEG data recorded inside the MR environment. We replicated this finding on the Inside rs-EEG data, supporting the robustness of this approach (Klimesch et al., 1990; Corcoran et al., 2018) and suggesting that these features may be extracted from simultaneous EEG-fMRI studies, and could potentially be used as features for integrative analysis.

The severe distortions observed in the absolute and relative spectral power highlight the huge impact BCG artifact residuals have on the resting-state EEG signals and demonstrate that artifact residuals remain after applying all the tested BCG correction methods, impairing the preservation of spontaneous EEG signal properties, as opposed to what is observed in event related potential studies (Debener et al., 2006; Assecondi et al., 2010; Vanderperren et al., 2010). We were also interested in investigating if, despite the generalized distortions of the power spectrum, functional information from task-reactive spontaneous EEG signals could be retrieved. We selected our task considering that alpha power reactivity to eyes closure-opening is one of the most robust phenomena observed in human EEG (Berger, 1929; Adrian and Matthews, 1934; Barry and De Blasio, 2017) and that it is the most commonly used paradigm on simultaneous EEG-fMRI resting-state studies (Goldman et al., 2002; Moosmann et al., 2003; Munck and Maurits, 2006; de Munck et al., 2007). We found that the occipital alpha rhythm reactivity to the EC-EO task was retrieved when using IFE and, to a lesser extent, with the OBS-ICA, and AAS-ICA–based corrections. However, neither the AAS, OBS, ICA or the PROJIC-AAS and PROJIC-OBS approaches preserved a clear distinction between EC and EO states.

Considering the potential implications of our findings, we then evaluated how the choice of the BCG correction method impacted the generation of EEG alpha power derived BOLD signal predictors used for EEG-informed fMRI analysis. Only the data processed using IFE of the alpha rhythm showed a clear significant inverse relation between alpha power and the BOLD signal from the occipital and parietal cortices, yielding similar statistical parametric maps to those obtained with conventional fMRI analysis (**Figure 7**), and those reported in previous alpha power EEG-informed fMRI studies (Goldman et al., 2002; Laufs et al., 2003; Moosmann et al., 2003; de Munck et al., 2007). None of the seven BCG removal methods tested here allowed to preserve the EEG alpha fluctuations to the same extent, and no statistical associations with the BOLD signal were observed in the EEG-informed fMRI analysis. These results provide compelling evidence that BCG artifact residuals and/or EEG signal losses related to the artifact removal procedure severely impair data quality and mask the functional association between EEG alpha power and occipito-parietal BOLD signal, hampering our interpretations from multimodal EEG-fMRI integrative analyses (Goldman et al., 2002; Scrivener, 2021; Warbrick, 2022).

Overall, our results demonstrate that state-of-the-art BCG artifact correction approaches still have important limitations. Our work highlights the need for refining and standardizing existing methods, and to develop novel approaches to deal with the BCG artifact to fully benefit from the advantages provided by simultaneous EEG-fMRI. We also highlight the need to validate current and novel approaches by evaluating the preservation of spontaneous EEG brain rhythms and their impact on multimodal integrative analyses. We demonstrated that IFE was effective to rescue the alpha rhythm reactivity to the eyes closure-opening task, though future studies should design specific paradigms to test the reactivity of other brain rhythms.

Regarding new software implementations, interesting proposals have arisen among the EEG-fMRI community. Given that the delay between the ECG and the BCG peak may vary over time (Oh et al., 2014), the adaptative OBS method was proposed to improve the results obtained with conventional OBS, by adjusting the variable delay between the QRS peak and the main BCG artifact peak (Marino et al., 2018). Another set of promising alternatives are the machine learning-based approaches that employ different data learning algorithms to better identify and classify the BCG artifact (Abolghasemi and Ferdowsi, 2015; McIntosh et al., 2021; Ebrahimzadeh et al., 2022; Lin et al., 2022). Even with the development of new signal processing tools that allow to better characterize and correct the BCG artifact, it has become evident that the solution for the BCG artifact problem must come not from software but most likely from hardware-based approaches, that incorporate additional elements or change the configuration of the EEG-fMRI setup to measure and/or reduce the artifact during data acquisition (Ullsperger and Debener, 2010; Jorge et al., 2014; Ebrahimzadeh et al., 2022). Promising examples include modified EEG caps containing a reference adapting layer (Xia et al., 2014) or carbon-based wire loops (van der Meer et al., 2016) that record electrode motion and use this information to better model and subtract the BCG artifact from the data, and also modifications in the materials for electrodes and leads as well are their geometrical configuration (Chowdhury et al., 2015; Assecondi et al., 2016).

Our study also contributes to the field by providing a simple, easy-to-implement workflow to characterize the impact of the BCG artifact and assess the efficiency of BCG artifact removal methods to reduce the artifact and preserve EEG spectral properties, which may be useful when attempting to validate novel BCG artifact correction approaches in resting-state EEG data or implementing an EEG-fMRI protocol in a new facility. Also, by making our dataset available to the scientific community we hope to incentivize other groups to participate in EEG-fMRI research and take advantage of these data to explore and validate novel BCG removal approaches, aiming to increase the collective effort to solve this 30-year-old problem in the field of simultaneous EEG-fMRI.

4.1 Study limitations

The present study has many strengths as it is one of the few works characterizing the preservation of spectral properties of resting-state EEG and EEG reactivity to a task after BCG artifact correction, and its impact on multimodal EEG-informed fMRI analysis. We carefully selected a sample of young healthy adults to assess EEG data quality. Although the number of subjects was relatively small ($N = 20$), it is much higher than many previous studies assessing EEG data quality during simultaneous EEG-fMRI experiments. Additionally, we validated and replicated our main findings in the rs-EEG-fMRI data by also analyzing the data recorded inside the scanner without fMRI acquisition. Overall, we found very similar results, supporting the idea that GA residuals do not influence our results from the EEG-fMRI data and showing that there was a consistent pattern between the results obtained from two independent sets of data from the same subjects, further supporting our conclusions.

Several limitations should also be considered. For practical reasons, the outside-EEG was always recorded before the inside-EEG and simultaneous EEG-fMRI for all subjects. Not counterbalancing the conditions may bias EEG quantitative measures if the subject's mental state has changed due to vigilance fluctuations or fatigue. However, given that in this study the time between conditions was relatively short (around 15 min between outside and inside scanner EEG recordings) we do not expect this to significantly affect our findings. Additionally, only male subjects were included in this study, which impacts the generalizability of our results and poses the need of replicating these findings in a cohort of female participants. Although we put great effort into matching the experimental conditions across subjects, we also acknowledge that our results may be influenced by the subject's head position relative to the B0 magnetic field and the amount of movement during the recordings (Debener et al., 2008; Yan et al., 2010; Mullinger et al., 2013b), both of which increase the within and between-subject variability in the BCG artifact spectral profile.

Inconsistent results as compared to other studies may be attributed to differences in methodologies, such as the use of low-impedance, conductive paste EEG caps in other studies (Debener et al., 2008; Vanderperren et al., 2010; Mullinger et al., 2013b; Arrubla et al., 2014), which may have some advantages over high-impedance caps as the one used here, but also differ in terms of the length and geometrical arrangement of the EEG wires and placement of the EEG amplifier relative to the B0 magnetic field (Chowdhury et al., 2015; Assecondi et al., 2016). Our data also suggest that cardiac signal recording using conventional ECG montages is not ideal for EEG-fMRI studies, and therefore other measurements of cardiac activity (or ideally scalp measurements of the BCG artifact itself) should be used, given that low quality ECG data may result in a poor estimation of the QRS-peak, which impacts the efficiency of the BCG artifact correction process (Iannotti et al., 2014). Another aspect to consider is that the parameter tunning for each BCG correction approach may dramatically influence the results. Both AAS and OBS were implemented using a fixed delay between the QRS events and the BCG amplitude peak (210 ms) which actually has been proved to be very variable within and between individuals (Yan et al., 2010; Marino et al., 2018). The number of principal components used to implement the OBS-based correction approach was kept constant for all subjects, while some studies have shown that optimizing parameters for each subject improves the results from the artifact correction process (Abreu et al., 2016; Marino et al., 2018). The parameters used here were selected to match the default options of the fMRIB toolbox, which are also the parameters typically used in many EEG-fMRI studies. We should therefore keep in mind that there may be room for improving the artifact correction procedure by fine-tuning these parameters (Marino et al., 2018). For ICA–based corrections, the ICs corresponding to the BCG artifact were manually selected. Although we used standard criteria to select the artifact-related components this generates a potential bias, and future studies should try using automatic or semi-automatic independent component selection algorithms. It is also very plausible that having a higher number of electrodes would improve the spatial characterization of the artifact, facilitating the selection of components and improving signal preservation. The most recent PROJIC approaches (Abreu et al., 2016) might significantly improve by adjusting different parameters on an individual subject basis. Here we used the recommended parameters across all individuals, and therefore this question should be addressed in future studies.

Finally, we evaluated the preservation of EEG functional properties only by focusing on posterior alpha power reactivity to the EC-EO task. Although we demonstrated that ICA feature extraction allowed to retrieve the associations between alpha power and hemodynamic signals, future studies are needed to evaluate if other spontaneous brain rhythms can be preserved, especially considering that lower frequencies are even more affected by the BCG artifact harmonic frequencies, and that higher frequencies have a much lower amplitude compared to the artifact waveforms. To tackle this question, other study designs need to be implemented to evaluate the reactivity of these particular rhythms (i.e., cognitive tasks, spontaneous activity recording in other natural physiological states such as sleep).

5 Conclusion

Overall, our study provides strong evidence that the most commonly used BCG correction methods have important limitations and were not able to entirely preserve the spectral power features of resting-state eyes-closed EEG activity (excepting for the individual alpha peak frequency and center of gravity), nor the functional reactivity of EEG signals to a simple EC-EO task, using this particular EEG-fMRI setup. Importantly, the EEG signal distortions compromised the results from integrative multimodal data analysis, evidencing the imposed difficulty of reliably studying the relationship between spontaneous electrophysiological activity and hemodynamic brain responses without optimal EEG data quality. ICA feature extraction allowed to preserve EEG oscillations related to the EC-EO task and to obtain reliable predictors for EEG-informed fMRI analysis. Future studies assessing novel or adapted hardware and software strategies to deal with the BCG artifact are needed and should be validated by assessing the preservation of EEG signal properties as the main concern.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://data.mendeley.com/datasets/crhybxpdy6.

Ethics statement

The studies involving human participants were reviewed and approved by the Bioethics Committee of the Instituto de Neurobiología, Universidad Nacional Autónoma de México, Campus Juriquilla. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

JG-R: conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing – original draft, editing and review, visualization, and project administration. MC-C: conceptualization, methodology, validation, formal analysis, writing – review and editing, and supervision. LC: conceptualization, methodology, validation, investigation, resources, writing–review and editing, and funding acquisition. JR-G: conceptualization, methodology, and validation. EP-A: conceptualization, methodology, validation, investigation, resources, data curation, supervision, project administration, and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Laboratorio Nacional de Imagenología por Resonancia Magnética, which has received funding from CONACyT and by the Instituto de Neurobiología, UNAM. JG-R received a scholarship from CONACyT (716272/634009). LC was partially funded by UNAM-DGAPA (IG200117 and IN204720) and Conacyt (1782).

Acknowledgments

We thank the personnel at the Unidad de Resonancia Magnética, UNAM, the Laboratorio Nacional de Imagenología por Resonancia Magnética and the Instituto de Neurobiología, UNAM for supporting this study. We also thank Dr. Sarael Alcauter and Dr. Markus Müller for their advice on data analysis. Finally, we thank Jessica González Norris for reviewing the English version of this 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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2022.951321/full#supplementary-material

References

Abolghasemi, V., and Ferdowsi, S. (2015). EEG-fMRI: Dictionary learning for removal of ballistocardiogram artifact from EEG. *Biomed. Signal Process. Control.* 18, 186–194.

Abreu, R., Leite, M., Jorge, J., Grouiller, F., van der Zwaag, W., Leal, A., et al. (2016). Ballistocardiogram artefact correction taking into account physiological signal preservation in simultaneous EEG-fMRI. *Neuroimage* 135, 45–63. doi: 10.1016/j.neuroimage.2016.03.034

Adrian, E. D., and Matthews, B. H. C. (1934). The interpretation of potential waves in the cortex. *J. Physiol.* 81, 440–471.

Allen, P., Josephs, O., and Turner, R. (2000). A method for removing imaging artifact from continuous EEG recorded during functional MRI. *Neuroimage* 12, 230–239.

Allen, P., Polizzi, G., Krakow, K., Fish, D., and Lemieux, L. (1998). Identification of EEG events in the MR scanner: The problem of pulse artifact and a method for its subtraction. *Neuroimage* 8, 229–239. doi: 10.1006/nimg.1998.0361

Arrubla, J., Neuner, I., Dammers, J., Breuer, L., Warbrick, T., Hahn, D., et al. (2014). Methods for pulse artefact reduction: Experiences with EEG data recorded at 9.4 T static magnetic field. *J. Neurosci. Methods* 232, 110–117. doi: 10.1016/j.jneumeth.2014.05.015

Assecondi, S., Lavallee, C., and Jovicich, J. (2016). Length matters: Improved high field EEG-fMRI recordings using shorter EEG cables. *J. Neurosci. Methods* 269, 74–87. doi: 10.1016/j.jneumeth.2016.05.014

Assecondi, S., Vanderperren, K., Novitskiy, N., Ramautar, J., Fias, W., Staelens, S., et al. (2010). Effect of the static magnetic field of the MR-scanner on ERPs: Evaluation of visual, cognitive and motor potentials. *Clin. Neurophysiol.* 121, 672–685. doi: 10.1016/j.clinph.2009.12.032

Barry, R., and De Blasio, F. (2017). EEG differences between eyes-closed and eyes-open resting remain in healthy ageing. *Biol. Psychol.* 129, 293–304.

Barry, R., Clarke, A., Johnstone, S., Magee, C., and Rushby, J. (2007). EEG differences between eyes-closed and eyes-open resting conditions. *Clin. Neurophysiol.* 118, 2765–2773.

Bénar, C., Aghakhani, Y., Wang, Y., Izenberg, A., Al-Asmi, A., Dubeau, F., et al. (2003). Quality of EEG in simultaneous EEG-fMRI for epilepsy. *Clin. Neurophysiol.* 114, 569–580.

Berger, H. (1929). Uber das elektrenkephalogramm des menshen. *Arch. Psychiatr. Nervenkr.* 87, 527–570.

Bullock, M., Jackson, G., and Abbott, D. (2021). Artifact reduction in simultaneous EEG-fMRI: A systematic review of methods and contemporary usage. *Front. Neurol.* 12:622719. doi: 10.3389/fneur.2021.622719

Chowdhury, M., Mullinger, K., and Bowtell, R. (2015). Simultaneous EEG – fMRI: Evaluating the effect of the cabling configuration on the gradient artefact. *Phys. Med. Biol.* 60, 241–251.

Corcoran, A., Alday, P., Schlesewsky, M., and Bornkessel-Schlesewsky, I. (2018). Toward a reliable, automated method of individual alpha frequency (IAF) quantification. *Psychophysiology* 55, 1–21. doi: 10.1111/psyp.13064

de Munck, J., Gonçalves, S., Huijboom, L., Kuijer, J., Pouwels, P., Heethaar, R., et al. (2007). The hemodynamic response of the alpha rhythm: An EEG / fMRI study. *Neuroimage* 35, 1142–1151.

Debener, S., Mullinger, K., Niazy, R., and Bowtell, R. (2008). Properties of the ballistocardiogram artefact as revealed by EEG recordings at 1.5, 3 and 7 T static magnetic field strength. *Int. J. Psychophysiol.* 67, 189–199. doi: 10.1016/j.ijpsycho.2007.05.015

Debener, S., Strobel, A., Sorger, B., Peters, J., Kranczioch, C., Engel, A., et al. (2006). Improved quality of auditory event-related potentials recorded simultaneously with 3-T fMRI: Removal of the ballistocardiogram artefact. *Neuroimage* 34, 587–597. doi: 10.1016/j.neuroimage.2006.09.031

Delorme, A., and Makeig, S. (2004). EEGLAB an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21.

Ebrahimzadeh, E., Saharkhiz, S., Rajabion, L., Oskouei, H., Seraji, M., Fayaz, F., et al. (2022). Simultaneous electroencephalography-functional magnetic resonance imaging for assessment of human brain function. *Front. Syst. Neurosci.* 16:934266. doi: 10.3389/fnsys.2022.934266

Ferrando, L., Bobes, J., Gibert, M., Soto, M., and Soto, O. (1998). *M.I.N.I. Mini international neuropsychiatric interview. Versión en español*. Madrid: Instituto IAP.

Figley, C., and Stroman, P. (2011). The role(s) of astrocytes and astrocyte activity in neurometabolism, neurovascular coupling, and the production of functional neuroimaging signals. *Eur. J. Neurosci.* 33, 577–588. doi: 10.1111/j.1460-9568.2010.07584.x

Gallego-Rudolf, J., Corsi-Cabrera, M., Concha, L., Ricardo-Garcell, J., and Pasaye-Alcaraz, E. (2022). *Simultaneous EEG-fMRI dataset. Mendeley data*. Available online at: https://data.mendeley.com/datasets/crhybxpdy6 (accessed November 1, 2022).

Goldman, R., Stern, J., Engel, J., and Cohen, M. (2002). Simultaneous EEG and fMRI of the alpha rhythm. *Neuroreport* 13, 2487–2492.

Huettel, S., Song, A., and McCarthy, G. (2004). *Functional magnetic resonance imaging*, 3rd Edn. Sunderland, MA: Sinauer Associates, Inc.

Huster, J., Debener, S., Eichele, T., and Herrmann, C. (2012). Methods for simultaneous EEG-fMRI: An introductory review. *J. Neurosci.* 32, 6053–6060. doi: 10.1523/JNEUROSCI.0447-12.2012

Iannetti, G., Niazy, R., Wise, R., Jezzard, P., Brooks, J., Zambreanu, L., et al. (2005). Simultaneous recording of laser-evoked brain potentials and continuous, high-field functional magnetic resonance imaging in humans. *Neuroimage* 28, 708–719. doi: 10.1016/j.neuroimage.2005.06.060

Iannotti, G., Pittau, F., Michel, C., and Vulliemoz, S. (2014). Pulse artifact detection in simultaneous EEG – fMRI recording based on EEG map topography. *Brain Topogr.* 28, 21–32. doi: 10.1007/s10548-014-0409-z

Ives, J., Warach, S., Schmitt, F., Edelman, R., and Schomer, D. (1993). Monitoring the patient's EEG during echo planar MRI. *Electroencephalogr. Clin. Neurophysiol.* 87, 417–420.

Jenkinson, M., Beckmann, C., Behrens, T., Woolrich, M., and Smith, S. (2012). FSL. *Neuroimage* 62, 782–790.

Jorge, J., van der Zwaag, W., and Figueiredo, P. (2014). EEG-fMRI integration for the study of human brain function. *Neuroimage* 102, 24–34. doi: 10.1016/j.neuroimage.2013.05.114

Klein, C., Hänggi, J., Luechinger, R., and Jäncke, L. (2015). MRI with and without a high-density EEG cap-what makes the difference? *Neuroimage* 106, 189–197. doi: 10.1016/j.neuroimage.2014.11.053

Klimesch, W., Sauseng, P., and Hanslmayr, S. (2007). EEG alpha oscillations: The inhibition-timing hypothesis. *Brain Res. Rev.* 53, 63–88.

Klimesch, W., Schimke, H., and Ladurner, G. (1990). Alpha frequency and memory performance. *J. Psychophysiol.* 4, 381–390.

Krakow, K., Allen, P., Symms, M., Lemieux, L., Josephs, O., and Fish, D. (2000). EEG recording during fMRI experiments: Image quality. *Hum. Brain Mapp.* 10, 10–15.

Laufs, H. (2012). A personalized. *Neuroimage* 62, 1056–1067.

Laufs, H., Kleinschmidt, A., Beyerle, A., Eger, E., Salek-haddadi, A., Preibisch, C., et al. (2003). EEG-correlated fMRI of human alpha activity. *Neuroimage* 19, 1463–1476.

Lemieux, L., Allen, P., Krakow, K., Symms, M., and Fish, D. (1999). Methodological issues in EEG-correlated functional MRI experiments. *Int. J. Bioelectromagn.* 1, 87–95.

Lin, G., Zhang, J., Liu, Y., Gao, T., Kong, W., Lei, X., et al. (2022). Ballistocardiogram artifact removal in simultaneous EEG-fMRI using generative adversarial network. *J. Neurosci. Methods* 371:109498. doi: 10.1016/j.jneumeth.2022.109498

Luo, Q., and Glover, G. (2011). Influence of dense-array EEG cap on fMRI signal. *Magn. Reson. Med.* 68, 807–815. doi: 10.1002/mrm.23299

Mandelkow, H., Halder, P., Boesiger, P., and Brandeis, D. (2006). Synchronization facilitates removal of MRI artefacts from concurrent EEG recordings and increases usable bandwidth. *Neuroimage* 32, 1120–1126. doi: 10.1016/j.neuroimage.2006.04.231

Marino, M., Arcara, G., Porcaro, C., and Mantini, D. (2019). Hemodynamic correlates of electrophysiological activity in the default mode network. *Front. Neurosci.* 13:1060. doi: 10.3389/fnins.2019.01060

Marino, M., Liu, Q., Castello, M., Del Corsi, C., Wenderloth, N., and Mantini, D. (2018). Heart-brain interactions in the MR environment: Characterization of the ballistocardiogram in EEG signals collected during simultaneous fMRI. *Brain Topogr.* 31, 337–345. doi: 10.1007/s10548-018-0631-1

Mayeli, A., Al Zoubi, O., Henry, K., Wong, C., White, E., Luo, Q., et al. (2021). Automated pipeline for EEG artifact reduction (APPEAR) recorded during fMRI. *J. Neural Eng.* 18:0460b4. doi: 10.1088/1741-2552/ac1037

McIntosh, J., Yao, J., Hong, L., Faller, J., and Sajda, P. (2021). Ballistocardiogram artifact reduction in simultaneous EEG-fMRI using deep learning. *IEEE Trans. Biomed. Eng.* 68, 78–89. doi: 10.1109/TBME.2020.3004548

Moosmann, M., Ritter, P., Krastel, I., Brink, A., Thees, S., Blankenburg, F., et al. (2003). Correlates of alpha rhythm in functional magnetic resonance imaging and near infrared spectroscopy. *Neuroimage* 20, 145–158. doi: 10.1016/s1053-8119(03)00344-6

Mulert, C., and Lemieux, L. (eds) (2010). *EEG-FMRI: Physiological basis, technique and applications*, 1st Edn. Berlin: Springer.

Mullinger, K., Castellone, P., and Bowtell, R. (2013a). Best current practice for obtaining high quality EEG data during simultaneous fMRI. *J. Vis. Exp.* 76:e50283. doi: 10.3791/50283

Mullinger, K., Havenhand, J., and Bowtell, R. (2013b). Identifying the sources of the pulse artefact in EEG recordings made inside an MR scanner. *Neuroimage* 71, 75–83. doi: 10.1016/j.neuroimage.2012.12.070

Mullinger, K., Debener, S., Coxon, R., and Bowtell, R. (2008). Effects of simultaneous EEG recording on MRI data quality at 1.5, 3 and 7 tesla. *Int. J. Psychophysiol.* 67, 178–188. doi: 10.1016/j.ijpsycho.2007.06.008

Munck, J., and Maurits, N. (2006). Correlating the alpha rhythm to BOLD using simultaneous EEG / fMRI. *Neuroimage* 30, 203–213.

Niazy, R., Beckmann, C., Iannetti, G., Brady, J., and Smith, S. (2005). Removal of FMRI environment artifacts from EEG data using optimal basis sets. *Neuroimage* 28, 720–737.

Nichols, T., and Holmes, A. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Hum. Brain Mapp.* 15, 1–25. doi: 10.1002/hbm.1058

Nierhaus, T., Gundlach, C., Goltz, D., Thiel, S., Pleger, B., and Villringer, A. (2013). Internal ventilation system of MR scanners induces specific EEG artifact during simultaneous EEG-fMRI. *Neuroimage* 74, 70–76. doi: 10.1016/j.neuroimage.2013.02.016

Oh, S., Han, Y., Lee, J., Yun, S., Kang, J., Lee, E., et al. (2014). A pulse artifact removal method considering artifact variations in the simultaneous recording of EEG and fMRI. *Neurosci. Res.* 81–82, 42–50. doi: 10.1016/j.neures.2014.01.008

R Core Team (2022). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rothlübbers, S., Relvas, V., Leal, A., Murta, T., Lemieux, L., and Figueiredo, P. (2015). Characterisation and reduction of the EEG artefact caused by the helium cooling pump in the MR environment: Validation in epilepsy patient data. *Brain Topogr.* 28, 208–220. doi: 10.1007/s10548-014-0408-0

Schomer, D., and da Silva, F. (2011). *Niedermeyer's electroencephalography: Basic principles, clinical applications, and related fields*, 6th Edn. Philadelphia, PA: Lippincott Williams & Williams.

Scrivener, C. (2021). When is simultaneous recording necessary? A guide for researchers considering combined EEG-fMRI. *Front. Neurosci.* 15:636424. doi: 10.3389/fnins.2021.636424

Shaw, J. (2003). "Inter-individual differences I. The classic studies," in *The brain's alpha rhythm and the mind*, ed. J. C. Shaw (Amsterdam: Elsevier), 125–143.

Sheehan, D. V., Lecrubier, Y., Harnett-Sheehan, K., Janavs, J., Weiller, E., Bonora, L., et al. (1997). Reliability and validity of the MINI international neuropsychiatric interview (mini): According to the SCID-P. *Eur. Psychiatry* 12, 232–241.

Smith, S., and Nichols, T. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage* 44, 83–98. doi: 10.1016/j.neuroimage.2008.03.061

Srivastava, G., Crottaz-Herbette, S., Lau, K., Glover, G., and Menon, V. (2005). ICA-based procedures for removing ballistocardiogram artifacts from EEG data acquired in the MRI scanner. *Neuroimage* 24, 50–60. doi: 10.1016/j.neuroimage.2004.09.041

Ullsperger, M., and Debener, S. (2010). *Simultaneous EEG and FMRI*, 1st Edn. Oxford: Oxford University Press.

van der Meer, J., Pampel, A., Van Someren, E., Ramautar, J., van der Werf, Y., Gomez-Herrero, G., et al. (2016). Carbon-wire loop based artifact correction outperforms post-processing EEG/fMRI corrections-A validation of a real-time simultaneous EEG/fMRI correction method. *Neuroimage* 125, 880–894.

Vanderperren, K., De Vos, M., Ramautar, J., Novitskiy, N., Mennes, M., Assecondi, S., et al. (2010). Removal of BCG artifacts from EEG recordings inside the MR scanner: A comparison of methodological and validation-related aspects. *Neuroimage* 50, 920–934. doi: 10.1016/j.neuroimage.2010.01.010

Warbrick, T. (2022). Simultaneous EEG-fMRI: What have we learned and what does the future hold? *Sensors* 22:2262. doi: 10.3390/s22062262

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer-Verlag.

Xia, H., Ruan, D., and Cohen, M. (2014). Separation and reconstruction of BCG and EEG signals during continuous EEG and fMRI recordings. *Front. Neurosci.* 8:163. doi: 10.3389/fnins.2014.00163

Yan, W., Mullinger, K., Brookes, M., and Bowtell, R. (2009). Understanding gradient artefacts in simultaneous EEG/fMRI. *Neuroimage* 46, 459–471.

Yan, W., Mullinger, K., Geirsdottir, G., and Bowtell, R. (2010). Physical modeling of pulse artefact sources in simultaneous EEG/fMRI. *Hum. Brain Mapp.* 31, 604–620. doi: 10.1002/hbm.20891

TYPE Original Research
PUBLISHED 04 January 2023
DOI 10.3389/fnins.2022.1011895

OPEN ACCESS

EDITED BY
Zheng Liu,
The University of British Columbia - Okanagan Campus, Canada

REVIEWED BY
Brian S. Eastwood,
MBF Bioscience LLC, United States
Shangbin Chen,
Huazhong University of Science and Technology, China
Philipp Boehm-Sturm,
Charité – Universitätsmedizin Berlin, Germany

*CORRESPONDENCE
G. Allan Johnson
✉ gjohnson@duke.edu

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 04 August 2022
ACCEPTED 12 December 2022
PUBLISHED 04 January 2023

CITATION
Tian Y, Cook JJ and Johnson GA (2023) Restoring morphology of light sheet microscopy data based on magnetic resonance histology. *Front. Neurosci.* 16:1011895. doi: 10.3389/fnins.2022.1011895

Restoring morphology of light sheet microscopy data based on magnetic resonance histology

Yuqi Tian, James J. Cook and G. Allan Johnson*

Department of Radiology, Duke University School of Medicine, Durham, NC, United States

The combination of cellular-resolution whole brain light sheet microscopy (LSM) images with an annotated atlas enables quantitation of cellular features in specific brain regions. However, most existing methods register LSM data with existing canonical atlases, e.g., The Allen Brain Atlas (ABA), which have been generated from tissue that has been distorted by removal from the skull, fixation and physical handling. This limits the accuracy of the regional morphologic measurement. Here, we present a method to combine LSM data with magnetic resonance histology (MRH) of the same specimen to restore the morphology of the LSM images to the in-skull geometry. Our registration pipeline which maps 3D LSM big data (terabyte per dataset) to MRH of the same mouse brain provides registration with low displacement error in ~10 h with limited manual input. The registration pipeline is optimized using multiple stages of transformation at multiple resolution scales. A three-step procedure including pointset initialization, automated ANTs registration with multiple optimized transformation stages, and finalized application of the transforms on high-resolution LSM data has been integrated into a simple, structured, and robust workflow. Excellent agreement has been seen between registered LSM data and reference MRH data both locally and globally. This workflow has been applied to a collection of datasets with varied combinations of MRH contrasts from diffusion tensor images and LSM with varied immunohistochemistry, providing a routine method for streamlined registration of LSM images to MRH. Lastly, the method maps a reduced set of the common coordinate framework (CCFv3) labels from the Allen Brain Atlas onto the geometrically corrected full resolution LSM data. The pipeline maintains the individual brain morphology and allows more accurate regional annotations and measurements of volumes and cell density.

KEYWORDS

mouse brain imaging, magnetic resonance histology, light sheet microscopy, cross-modality registration, tissue clearing

1. Introduction

Combining mesoscopic structural information of the brain and histology at the cytoarchitectural scale has been a focus in recent years to reveal the bridge between tissue morphological alternations and disease (Casanova et al., 2009; Vemuri and Jack, 2010; Zhang et al., 2012), brain insult (Tuor et al., 2014; Fornito et al., 2015; Weishaupt et al., 2016) and aging (Eylers et al., 2016; Schmitz et al., 2018). There is clear evidence that morphological disruptions underlie brain dysfunctions at both the meso- and microscopic scale; for example the corpus callosum volume reduction in autism (Egaas et al., 1995; Hardan et al., 2000; Tepest et al., 2010; Loomba et al., 2021) and neuronal death following ischemic insult (Weishaupt et al., 2016). Merging structural changes in specific brain regions at the mesoscale with corresponding quantitative cellular measurements at the microscopic scale will open an entirely new window into understanding the brain.

Diffusion tensor imaging (DTI) provides particularly unique insight into brain morphology and connectivity (Fornito et al., 2015). However, extension of DTI to more basic studies in the mouse is challenging because the mouse brain @ 435 mg is about 3,000 times smaller than the human brain. Through a series of innovations, the Duke Center for *in vivo* Microscopy (CIVM) has extended the spatial resolution of magnetic resonance imaging (MRI)/DTI by more than 500,000 times that of routine clinical scans in perfusion fixed post mortem specimens (e.g., MRH) (Johnson et al., 1993; Johnson et al., 2007). Recent work has pushed the resolution of DTI to 15 × 15 × 15 μm^3 and accelerated the acquisition with compressed sensing, which enables routine acquisition of high-resolution multidimensional whole mouse brain images (Wang et al., 2018a; Johnson et al., 2019, 2022). These high-fidelity mesoscale MRH data now enable correlation between the MRH metrics and the tissue cytoarchitecture.

The development of tissue clearing and LSM have allowed neuroscientists to routinely image whole cleared mouse brains at cellular resolution (Erturk et al., 2012). Continued innovation in clearing (SHIELD) (Park et al., 2019) and immunohistochemistry (SWITCH) (Murray et al., 2015) has enabled staining of varied cell types (neuron, oligodendrocyte, microglia), structural proteins (myelin) and pathologies (a-beta and tau proteins).

Merging MRH and LSM data from the same specimen will capture the best of both. MRH with DTI is a non-destructive and multi-contrast imaging method which preserves accurate brain morphology since the scanning can be done with the brain in the skull. DTI with high angular sampling provides maps of whole brain connectivity (Johnson et al., 2019). Multiple scalar images provide exquisite tissue contrast differentiating brain subunits. Post processing pipelines can exploit these multi-contrast images to automatically label more than 300 different sub-regions (Johnson et al., 2022). LSM provides cellular resolution but requires the removal of the brain from the skull and tissue clearing, which induces tissue swelling. Dissection of the brain from the skull frequently results in tissue loss or tearing (**Figure 1**). Labeling is not always as uniform as one might hope. Mapping LSM to MRH restores the tissue geometry and allows automated labeling of the sub-regions in the LSM data.

Finally, the most common method for labeling cleared brain images (Kutten et al., 2016; Goubran et al., 2019; Tappan et al., 2019; Perens et al., 2021) involves registration to the Allen Brain Atlas which has been constructed from 2D serial sections acquired at 100 µm intervals averaged from ∼1,600 young adult C57BL/6J mice (Wang et al., 2020). Mapping the cleared brain images from another strain at a different age to the ABA may obscure regional volume changes that might be important image phenotypes for the study.

Our long-range goal is development of the infrastructure to support routine, comprehensive morphologic phenotyping of the mouse brain using combined MRH and LSM to map the genetic impact on cells and circuits. Those familiar to registration methods will appreciate that registration of images into a common space requires recognition of the challenges that are unique to the task and adapting the code to those challenges. Those challenges are: (1) The sources of contrast in MRH and LSM are wildly different. (2) Each modality has many different contrasts, e.g., 11 different scalar images in MRH and even greater number of contrasts in immune histochemistry for LSM. (3) The geometric distortion in the LSM data can exceed 40% and there is frequent tissue loss. (4) The data volumes are large approaching a terabyte for a single specimen. In this paper we have addressed a these challenges, developed a process for optimizing the software, and highlighted some of the limitations in combining MRH/LSM of the same brain routine.

2. Materials and methods

2.1. MRH histology and LSM

All animal procedures were conducted under guidelines approved by the Duke Institutional Animal Care and Use Committee. Specimens were perfusion fixed using an active staining method that has been described in detail previously (Johnson et al., 2019). Warm saline to flush out blood was perfused through a catheter in the left ventricle. This was followed by a formalin/Prohance (Gadoteridol) mixture titrated to reduce the spin lattice relaxation time (T1) of the tissue enabling accelerated scanning. The MRH scanning was performed on a 9.4T vertical bore magnet with a Resonance Research Inc. (Billerica, Md) gradient coil yielding peak gradients up to 2,500 mT/m. The scanner is controlled by an Agilent console running VnmrJ 4.0. The acquisition was accelerated using compressed sensing (Wang et al., 2018b;

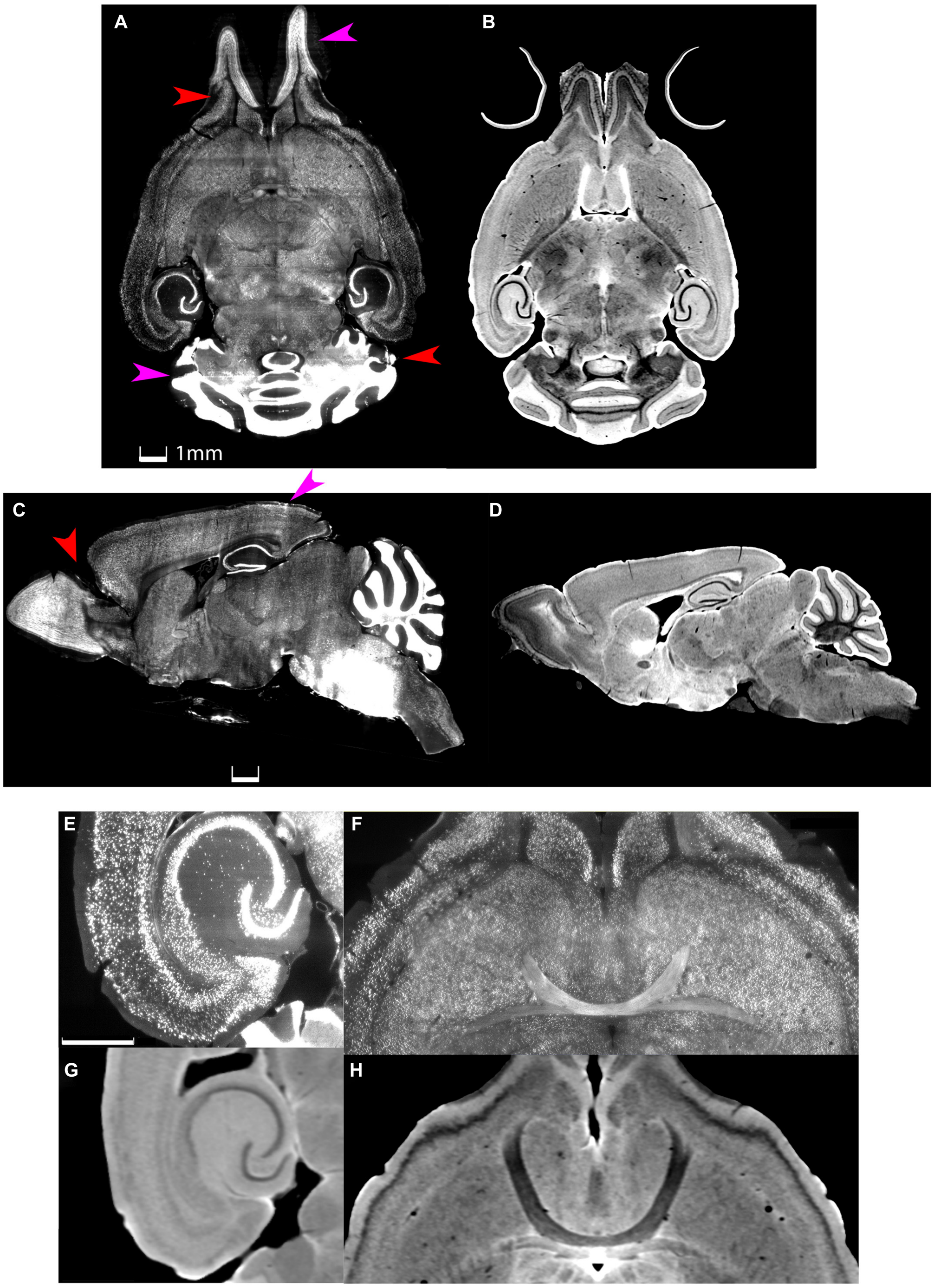


FIGURE 1
Distortion and tissue tearing in light sheet microscopy (LSM) compared to magnetic resonance histology (MRH). A comparison between LSM images of a mouse brain stained with NeuN **(A,C,E,F)** and a diffusion weighted MRH image of the same specimen **(B,D,G,H)** highlights some of the challenges and opportunities. Red arrows indicate the tissue tearing. Purple arrows indicate the swelling (specimen 200316). Scale bar: 1 mm.

Johnson et al., 2019). Diffusion tensor images were acquired using a protocol described in detail in Johnson et al. (2022). The protocol employed a Stesjkal Tanner spin echo sequence with *b*-values of 3,000 s/mm^2, 108 angular samples spaced uniformly on the unit sphere, a compression factor of 8 × yielding a large (252 GB) 4D volume with isotropic resolution of 15 μm. A baseline (b$_0$) image was acquired after every 10th angular sample, yielding 18 baseline volumes. These volumes were averaged together to create a template to which all other volumes were registered (ANTs) to correct for residual eddy currents. A MATLAB script produced a diffusion weighted image (DWI) by averaging the 108 diffusion images together. The 4D data volume was processed through DSI Studio[1] using both the DTI and GQI algorithms (Yeh et al., 2010) which yields eleven different scalar images (see **Supplementary Table 3**). We explored the use of the following DTI scalar images to drive the registration: axial diffusivity (AD), diffusion weighted (DWI), fractional anisotropy (FA) and radial diffusivity (RD). Two scalar data sets (DWI and FA) were used to registered labels to the MRH volumes (and thence to the LSM) using the Small Animal Multivariate Brain Analysis (SAMBA) an pipeline described fully in Anderson et al. (2019).

Five specimens from Johnson et al. (2022) were included in this study. They are summarized in **Table 1**. Specimen 200316, a 90 day male C57/B6 mouse was used as a reference atlas. It provides a modified version of the Common Coordinate Frame (CCFv3) from the Allen Brain Atlas (Wang et al., 2020). The CCFv3 defines regions of interest (ROIs) for 461 structures. Many of these structures are so small that reliable alignment is challenging. The reduced CCFv3 (rCCFv3) is a set of 180 labels/hemisphere generated by combining some of the regions in CCFv3 that are too small to transfer accurately in the registration pipeline. The full summary of the rCCFv3 can be found in Johnson et al. (2022).

Following the MRH scans, the brains were removed from the skulls and sent to LifeCanvas Technology[2] for tissue clearing and LSM imaging. The brains were cleared using SHIELD (Park et al., 2019) and stained using SWITCH (Murray et al., 2015) and scanned on a selective plane illumination microscope (SPIM) yielding three channel whole brain images at a resolution of 1.8 × 1.8 × 4.0 μm. Each of the three channels yields a nearly isotropic volume at a different wavelength of ∼ 300GB. The aggregate dataset for one specimen (MRH and 3 channels of LSM) is typically ∼ 1 TB. **Table 1** lists immuno histochemistry stains that were used to test the pipeline.

2.2. Multiple stages of the workflow

Initial attempts at registration with popular registration algorithms (Avants et al., 2008; Klein et al., 2010) were particularly unsuccessful in cerebellum and olfactory bulb both of which are prone to significant distortion after removal from the skull (**Figure 2**). Our workflow employs an initial manual initialization followed by an automated multistep registration based on ANTs (Avants et al., 2008). The manual initialization is applied to all specimens to correct the most challenging distortions. It uses sparse landmarks (15∼20) with many concentrated in olfactory bulb and brain stem where the tissue distortion in the LSM are the greatest. Landmarks are placed in pairs, on both LSM and MRH. The landmark locations are 4 landmarks on olfactory bulbs, 2–3 landmarks on vessels on both sides between cortex and striatum, 3 landmarks on cerebellum, 2 landmarks on dentate gyrus, 2 landmarks on hippocampus and 2 landmarks on brain stem (as shown in **Supplementary Figures 5B, D**). The second automated step is described in detail below.

2.3. Quantitative loss function

The goal of registration is to transform the image of interest, M i.e., the image that is being moved (the LSM volume) into the space defined by the fixed reference image F (MRH volume). Our pipelines use a series of transforms applied successively with a loss function to evaluate each stage of transformation. For a single transform stage n, the transformation T$_n$ can be obtained from optimizing the loss function:

$$L_n\,(M, F) \;=\; S(T_n \circ M, F) \tag{1}$$

in which S is the similarity between F and transformed M. Common similarity metrics include mutual information (MI), cross correlation (CC), mean square error (MSE), which capture how well the two images are matched based on the joint histogram or signal intensities. Since we may use these metrics during registration, using the same metric repetitively for evaluation is unacceptable. At the same time, MSE, CC, global MI etc., by their intensity-based or histogram-based principles will not generate a stable predictability map between LSM and MRH due to the wildly different contrasts. The further explanation can be seen with the MI equation in the section "2.4 Optimization and validation." Therefore, we need to devise a different loss function.

The initialized LSM data is warped to MRH space with a combination of registration steps built on ANTs (Avants et al., 2008). Our workflow encompasses multiple types of registration, and each type has different settings of metrics for optimization and multi-resolution coarse-to-fine refinement. The loss function should evaluate the cumulative consequences of each of these steps. We devised a loss function based on a large group (50–200) of fiducials to optimize the pipeline and evaluate its stability (see **Table 1**). We emphasize that these fiducials were used only in the evaluation of our pipeliness and are not required for routine use. These fiducials were generated by an experienced researcher on five different specimens (see **Table 1**) and consisted of matched pairs of points in MRH and LSM. Assuming the composite transform generated from our workflow is T, applying T to the fiducials in the space of LSM transforms these fiducials to the MRH space. The distance between one MRH fiducial (r_{mr}) and its corresponding transformed LSM fiducial ($T(r_{lst})$) in the space of MRH is

[1] https://dsi-studio.labsolver.org/

[2] https://lifecanvastech.com/

TABLE 1 Test specimens for combined magnetic resonance histology (MRH)/light sheet microscopy (LSM) registration.

Specimen	Strain/Age	Fiducial	NeuN	Syto	MBP	IBA1	AutoF
191209	C57/90 d	175	X	X	X		
200302	C57/90 d	50	X		X		X
200316	C57/90 d	200	X		X		X
190108	BXD89/111 d	52	X	X		X	
200803	BXD89/687 d	51	X	X		X	

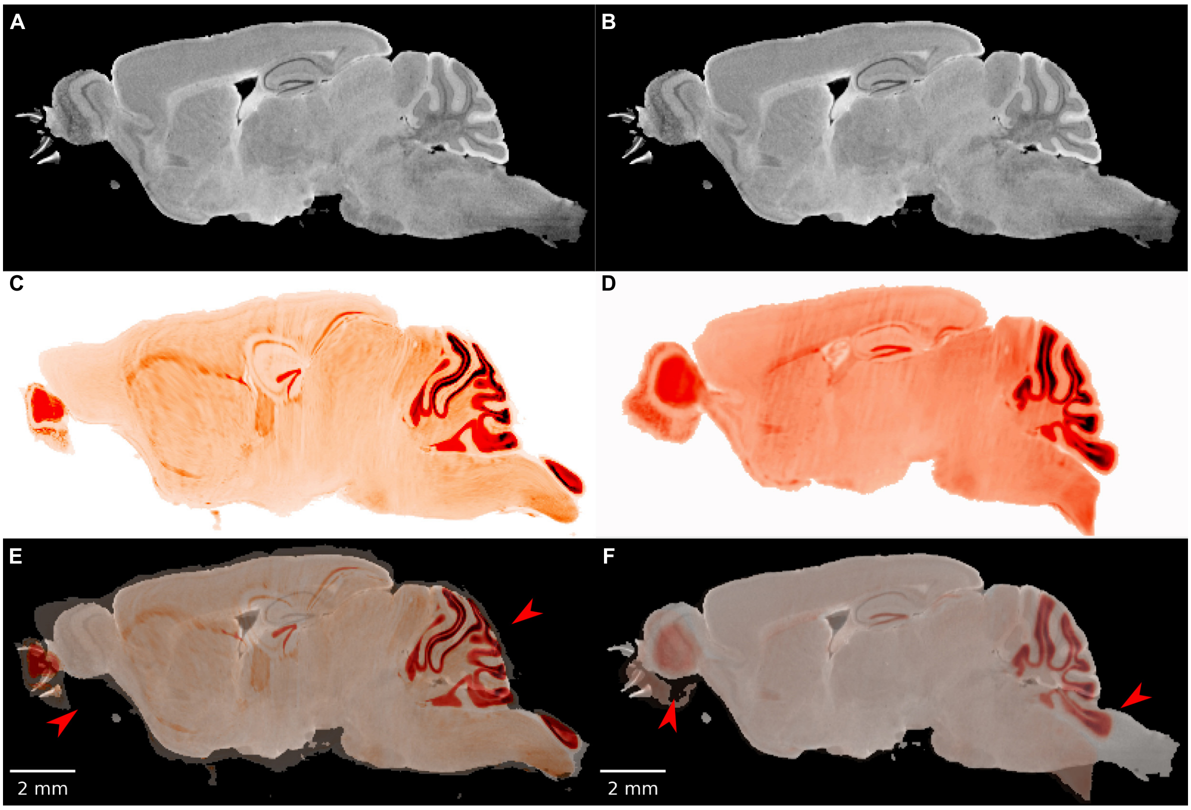


FIGURE 2
The failure of existing registration algorithms in the cerebellum and olfactory bulb. **(A,B)** DWI; **(C,D)** NeuN image after registration; **(E,F)** overlaid DWI/NeuN (specimen 191209). The left hand column shows the result of Elastix (Klein et al., 2010) with rigid and b-spline registration and default settings. The registration errors in the olfactory bulb and brain stem are reduced but the errors in the dentate gyrus and cerebellum are significant (arrows in panel **E**). The right hand column shows the result of ANTs (Avants et al., 2008) with affine and SyN and default settings. There is a reasonable overlap in the dentate gyrus but significant mismatch in the cerebellum and olfactory bulb (arrows in panel **F**).

regarded as displacement from ground truth, and the average displacement i.e., L2 norm is used as the loss score, i.e.

$$\mathrm{L2} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{r}_{\mathrm{mr,i}} - \mathrm{T}(\mathrm{r}_{\mathrm{lst,i}})\right)^2}{n} \qquad (2)$$

2.4. Optimization and validation

The registration transform can be separated into linear and non-linear stages. To reduce the computation, a complicated registration should start from the linear transforms to adjust the position, orientation, and scaling of the moving image to coarsely and globally match the fixed and moving images. Then, application of non-linear transforms will deform the grid to locally match the fine details of fixed and moving images. From the popular options of non-linear transforms, we choose b-spline and symmetric diffeomorphic normalization (SyN) registration methods based on their efficiency on large datasets with complicated geometry.

B-spline relies on the control points to adjust local transform until reaching the minima of the loss function. The curve defined by b-spline is a conjunction of multiple polynomial curves which only depends on a local group of the control points. Based on the zero-order parametric continuity of B-spline, changing one control point will only influence the local neighborhood on the grid instead of propagating further. Therefore, b-spline can generate localized deformations flexibly and is computationally efficient when dealing with many control points. The conventional b-spline method applies free-form deformation to the image. In this study, the reversal form of the deformation is also required when transforming images between the fixed and moving spaces. Hence, we adopt the b-spline with the explicit symmetry i.e., b-spline Syn (Tustison and Avants, 2013) in the actual practice.

SyN, as a representation of diffeomorphic algorithms, generates voxel-wise transformation based on symmetrical and invertible displacements and velocity fields. SyN is implemented on the Insight ToolKit platform and based on Large Deformation Diffeomorphic Metric Matching (LDDMM) principles. As an improvement, it develops the symmetry between the fixed and moving images, i.e., instead of maximizing the similarity between $T \circ M$ and F, SyN maximizes the similarity between $\varphi_1 \left(m, t\right) M$ and $\varphi_2 \left(f, 1 - t\right) F$, in which $t \in [0, 1]$, m and f are the respective identity positions of M and F, and φ_1, φ_2 are the respective correspondence maps from M to F, and from F to M. Based on the backward and forward symmetry, $t = 0.5$. The optimization problem is then based on the equation:

$$\begin{aligned} & E\left(F, M\right) \\ & = \inf_{\varphi_1} \inf_{\varphi_2} \int_{t=0}^{0.5} \{||\upsilon_1 \left(x, t\right)||_L^2 + ||\upsilon_2 \left(x, t\right)||_L^2\} dt \\ & + S_\Omega(|F\left(\varphi_1 \left(0.5\right)\right) - M\left(\varphi_2 \left(0.5\right)\right)|) \end{aligned} \tag{3}$$

to minimize both the pixel displacement and the difference between $F\left(\varphi_1 \left(0.5\right)\right)$ and $M\left(\varphi_2 \left(0.5\right)\right)$, in which υ_1 and υ_2 are velocity fields in the opposite directions, S_Ω is the similarity measurements across the whole x surface. The advantage of SyN is the low computational cost and the preservation of the image topology.

An additional factor influencing the registration is the selection of the similarity metrics. The most common similarity metrics include cross correlation (CC) and mutual information (MI).

A common definition of CC is

$$CC\left(F, M\right) = \frac{\sum_{i,j} \left(F_{i,j} - \overline{F}\right)\left(M_{i,j} - \overline{M}\right)}{\sqrt{\sum_{i,j} \left(F_{i,j} - \overline{F}\right)^2} \sqrt{\sum_{i,j} \left(M_{i,j} - \overline{M}\right)^2}} \tag{4}$$

CC is very sensitive to significant rotation and scale changes and any intensity difference, which limits its performance on cross modality registration evaluation, but including local neighborhood CC into the optimization penalty may still help with matching the contours of cross modality images.

MI defined by:

$$\begin{aligned} MI\left(F, M\right) & = H\left(M\right) - H\left(M \mid F\right) = H\left(M\right) + H\left(F\right) \\ -H\left(FM\right) & = \sum_{m \in M} \sum_{f \in F} p(f, m) log \frac{p(f, m)}{p\left(f\right) p(m)} \end{aligned} \tag{5}$$

originates from information theory and measures how much information of one image can be predicted correctly from another image which is already known. In this equation, H is the entropy, p(f, m) is the joint probability density function of the fixed reference atlas F and the moving image M that is being mapped into that reference, and p(f) and p(m) are the marginal probability density functions of F and M.

MI is commonly used for cross-modality registration because it is based on intensity probability distribution instead of pure intensity. However, for registering MRH and LSM, only employing MI may be risky. As shown in **Figure 1**, e.g., DWI and NeuN, in regions like cerebellum and olfactory bulbs, the intensity of gray matter in DWI is relatively low while in NeuN is high; meantime, in the central parts of the brain and the cortex, the intensity in DWI is relatively high while in NeuN is low. With the definition of MI, the joint histogram of F and M is scattered and the MI in this case is low, with the minimum being 0 which means no mutual information between two images. MI is a good measurement for Image F,M when the joint histogram of F and M consists of one or multiple condensed distributions, but may not be a good similarity measurement for MRH+LSM as the local contrast distribution is wildly different. Therefore, if the loss function calculated by MI is high, we do not know whether it is induced by the geometric mismatch because of the failed registration, or just the local contrast difference between MRH and LSM.

Table 2 describes the steps for optimizing the registration between an MRH and LSM. In our initial tests we used the DWI and Syto16 images from specimen 191209, because they both present abundant landmarks with some similarities, though the contrasts are different. In later studies, we used DWI and NeuN because NeuN and Syto16 have similar contrast and the NeuN stain from LifeCanvas was more consistent. **Table 2** lists multiple stages starting with the global alignment progressing to local higher resolution details. At each stage multiple variations of the ANTs modules appropriate for that task are compared. We refer to a collection as a "pipe" e.g., P1_01 is one combination of ANTs modules to perform global registration. The pipe with the lowest L2 norm is chosen for the final pipeline. The output of this pipe is the starting point for the next stage. The Syto LSM image was initialized using the coarse (20 point) landmark initialization correcting the large distortions in brainstem and olfactory bulb. The optimization described in **Table 2** was performed on data that had been down sampled to 45 μm to allow a broad search of parameters. In each stage, we employ the multi-resolution method, which

TABLE 2 Pipeline optimization pyramid @ 45 μ m resolution.

Experiments	Optimization composition	Score
Stage 1 Global	To optimize the combination of multiple transforms	
P1_01	Affine (Default) + Syn (Default)	0.3467
P1_02	Affine (Default) + B-spline Syn (Default) + Syn (Default)	0.303
P1_03	Rigid (Default) + Affine (Default) + Syn (Default)	0.4269
P1_04	Rigid (Default) + Affine (Default) + B-spline Syn (Default) + Syn (Default)	0.3644
P1_05	Affine (Default) + B-spline Syn (Default)	0.3333
P1_06	B-spline Syn (Default) + Syn (Default)	0.3131
Stage 2 Similarity	To optimize the similarity metrics	
P2_01	Affine (MI) + B-spline Syn (CC) + Syn (MI)	0.303
P2_02	Affine (MI) + B-spline (CC) + Syn (CC)	0.3606
P2_03	Affine (MI) + B-spline (MI) + Syn (MI)	0.3385
P2_04	Affine (MI) + B-spline (MI) + Syn (CC)	0.3752
P2_05	Affine (CC) + B-spline (CC) + Syn (MI)	0.3186
Stage 3 B-spline	To tune the multiresolution setting in b-spline stage	
P3_11	--shrink-factor 10--smoothing 5	0.332
P3_12	--shrink-factor 1--smoothing 5	0.323
P3_13	--shrink-factor 1--smoothing 1	0.326
P3_21	--shrink-factor 10 × 1--smoothing 2 × 1	0.280
P3_22	--shrink-factor 10 × 1--smoothing 10 × 2	0.285
P3_23	--shrink-factor 10 × 1--smoothing 10 × 10	0.350
P3_24	--shrink-factor 2 × 1--smoothing 2 × 1	0.308
P3_31	--shrink-factor 10 × 5 × 1--smoothing 3 × 2 × 1	0.277
P3_32	--shrink-factor 10 × 5 × 1--smoothing 10 × 5 × 1	0.312
P3_33	--shrink-factor 10 × 5 × 1--smoothing 10 × 10 × 10	0.383
P3_34	--shrink-factors 3 × 2 × 1--smoothing 3 × 2 × 1	0.300
P3_41	--shrink-factor 10 × 7 × 4 × 1--smoothing 1 × 1 × 1 × 1	0.274
P3_42	--shrink-factor 10 × 7 × 4 × 1--smoothing 4 × 3 × 2 × 1	0.268
P3_43	--shrink-factor 10 × 7 × 4 × 1--smoothing 10 × 7 × 4 × 1	0.362
P3_44	--shrink-factor 10 × 7 × 4 × 1--smoothing 10 × 10 × 10 × 10	0.495
P3_45	--shrink-factor 4 × 3 × 2 × 1--smoothing 4 × 3 × 2 × 1	0.278
P3_51	--shrink-factor 9 × 7 × 5 × 3 × 1--smoothing 9 × 7 × 5 × 3 × 1	0.292
P3_52	--shrink-factor 9 × 7 × 5 × 3 × 1--smoothing 5 × 4 × 3 × 2 × 1	0.355
Stage 4 B-spline distance	To tune b-spline spline distance	
P4_00	Spline distance default to 26	0.268
P4_01	Spline distance = 10	0.341
P4_02	Spline distance = 40	0.268
P4_03	Spline distance = 60	0.268
Stage 5 Syn	Tuning the multiresolution setting in SyN stage	

(Continued)

TABLE 2 (Continued)

Experiments	Optimization composition	Score
P5_01	--smoothing 3 × 2 × 1 × 0--shrink 4 × 3 × 2 × 1	0.2679
P5_02	--smoothing 10 × 7 × 4 × 1--shrink 10 × 7 × 4 × 1	0.3543
P5_03	--smoothing 10 × 7 × 4 × 1--shrink 4 × 3 × 2 × 1	0.3084
P5_04	--smoothing 1 × 1 × 1 × 1--shrink 4 × 3 × 2 × 1	0.2703
P5_05	--smoothing 0 × 0 × 0 × 0--shrink 4 × 3 × 2 × 1	0.2637
P5_06	--smoothing 0 × 0 × 0 × 0--shrink 6 × 4 × 2 × 1	0.2626
P5_07	--smoothing 0 × 0 × 0 × 0--shrink 10 × 7 × 4 × 1	0.2625
P5_08	--smoothing 0 × 0 × 0 × 0 × 0--shrink 20 × 15 × 10 × 5 × 1	0.2633
P5_09	--smoothing 3 × 2 × 1 × 0--shrink 10 × 7 × 4 × 1	0.2656

Specimen is 191209-1-1. The steps and parameters for the pipes that were tested are summarized for each stage. For each stage only the parameters to be optimized will change, and one optimal pipe will be selected among the pipes within one stage. The aim of the pipeline initialization is to select an optimal registration variables for certain contrasts in MRH/LSM. The pipeline optimization has been performed using one specimen. The application to additional specimens and contrast combinations has been demonstrated in **Supplementary Table 2**.

initially performs the registration at a lower resolution with fewer control points and then samples the control points to a higher resolution following convergence of the loss function without consuming large computing resources.

The optimization pyramid (**Table 2**) includes:

- Stage 1 focuses on optimizing large *global* details. Each pipe employs linear registration (rigid and affine) followed by non-linear registration (b-spline syn and syn). Each pipe uses the same default parameters. In stage 1, P1_02 i.e., Affine (Default) + B-spline Syn (Default) + Syn (Default) yielded the lowest loss score so its output served as the input for stage 2.
- Stage 2 focuses on *similarity metrics*, i.e., mutual information or cross correlation.
- Stage 3 adjusts the *b-spline multi-resolution* settings with number of layers, shrink factors (i.e., down-sampling) and smoothing sigmas (i.e., the radius of Gaussian filter).
- Stage 4 adjusts the *b-spline distance*, an additional parameter in b-spline syn.
- Stage 5 alters the *syn multi-resolution* settings with different number of layers, shrink factors and smoothing sigmas.

The pipe with the lowest L2 norm is labeled in green at each stage.

2.5. Registration validation

Registration with the five specimens was evaluated using the fiducials recorded in **Table 1**. The use of fiducials facilitates the comparison of different pipes and image combinations explained in the section "3.1 Optimization of pipes" and "3.2 Pipeline performance with varied image combinations." **Supplementary Figure 5** shows the dense collection of fiducials used to optimize the pipes (specimen: 191209). We performed an initial evaluation on specimen 200316 with an equally dense set of fiducials. At this point, it was clear that a sparser set would be adequate for validation in the other specimens.

The precision of a given registration was measured using Imaris[3] which allows one to load multiple 3D volumes of different spatial resolution as layers. Vascular landmarks were identified using the three-plane view. Imaris allows one to toggle between an LSM image and a companion MRH image while interactively moving a 3D cross hair. One initially identifies a vessel in cross section in the LSM and moves the plane until one encounters a bifurcation. At this point the 3-dimensional coordinates are recorded. The process is repeated in the MRH and the Euclidean distance is measured. **Supplementary Figure 3** shows the magnified cross section of a vessel in the NeuN image. The plane of the vessel cross section was adjusted until the bifurcation was evident and a fiducial was marked. The RD image provides high contrast for the same vessel where the same vessel bifurcation is visible.

2.6. Data and code availability

We have made the data for experiments 1–3 available under creative commons by NC-SA at https://civmimagespace.civm.duhs.duke.edu/login.php/client/4. The data is stored in H5 format to enable interactive examination using Neuroglancer.[4] Reviewers can log in with the following credentials. Viewers will remain anonymous.
cr371@duke.edu
Password: mrmicroscopy

The code is available in github.[5] The code provided is implemented in Perl and bash (which are available on windows/macos) and based on Ants.[6]

[3] https://imaris.oxinst.com/products/imarisessentials
[4] https://github.com/google/neuroglancer
[5] https://github.com/YuqiTianCIVM/MRH_LSM_registration
[6] https://github.com/ANTsX/ANTs

When applying this method, please follow the procedures described in the accompanying instructions for installation and in the method section. The processing time will depend on the computing resource. Please use a high-performance computing resource paired with high memory and page faulting, especially if the input data is hundreds of GB.

3. Results

3.1. Optimization of pipes

Figure 3A plots the rank ordered L2 norm for each pipe. Visual comparison are provided in Figures 3B–I. Figures 3B, F, the starting point for all the comparisons shows the initialization using ∼20 manual landmarks. The comparison between a 45 μm pipe that is less accurate (e.g., p2_02, L2 = 0.361) and the optimal pipe @ 45 μm e.g., (p3_42, 0.268), is shown in Figures 3C, D, G, H. The improvement is evident (see white arrows in Figure 3G).

The parameters derived from Stages 1–3 had significant impact on the L2 norm. Changing the b spline distance and Syn in Stages 4 and 5 had less impact so the default settings were used in P3_042 as the starting point for experiments conducted with the full resolution (15 μm data) outlined in Table 3. The variable of interest for this stage of optimization is the shrink factor. This last stage is more nuanced depending on compute time and the combination of LSM/MRH contrasts (e.g., DWI/Syto, FA/NeuN) which is discussed in more detail in the section "3.2 Pipeline performance with varied image combinations." The optimization @ 15 μm is started from pipe P6_01, which has the same registration setting with the optimal pipe @ 45 μm (P3_042). Table 3 demonstrates that the shrink factor has an enormous impact on compute time but the L2 norm remains relatively unchanged. Inspection of the results shows more subtle impact of the shrink factor. P6_01H overfits the data and is 27 times slower. P6_07H does not overfit and it can be executed in a modest time. Comparison between the best pipe at 45 μm (P3_042) and P6_07H optimized on 15 μm is shown in Figures 3D, E, H, I.

The L2 norm is also shown separately for the cerebellum (CB), olfactory bulb (OB), central section of the brain (C), and brain stem (BS). Each region poses unique challenges to the algorithm. The contrast is very high between the white matter and the intensely stained granular cell layer in the cerebellum in both the NeuN and Syto images, and there is comparable strong contrast in the DWI. Thus, the L2 norm for this cerebellar region converges to a low value for all the pipes. In the central part of the brain, the dentate gyrus, fimbria, and corpus callosum all provide unambiguous landmarks and fine tuning the pipeline leads to gradual improvement in the score. The olfactory bulb shows a similar effect, but the score does not converge to as low *a*-value. This may be because the olfactory bulb is one of the most distorted regions of the brain, and there are frequent tissue tears (e.g., the top red arrow in Figure 1). Finally, the brain stem is the most challenging region for registration as evidenced by high L2 norm and the high variability between different pipes. The cause of this is again evident on inspection of the sagittal LSM and MRH imaged in Figures 1C, 2D, F. The spinal cord in the LSM is grossly misplaced from its natural position forcing the algorithm into large displacements.

The transform obtained from the 15 μm registration was applied to the full resolution LSM data through the python interface of 3D Slicer, in the order of their generation. The time to apply transforms to full resolution LSM data (∼300GB) was ∼2 h.

3.2. Pipeline performance with varied image combinations

The registration success depends on the similarity between the anatomical features that are evident in the fixed and moving volumes. The initial work described above varied the pipes while registering Syto16 to DWI using specimen 191209-1-1. This section of the manuscript uses a fixed pipe (p6_07H) to explore the success of several specific combinations of LSM/MRH images in another specimen (200316) to demonstrate the approach more broadly. The DTI pipeline produces eleven different scalar images, each highlighting different diffusion properties (see Supplementary Table 3). The anatomic landmarks in the LSM vary widely depending on the immunohistochemistry used. There are an enormous number of combinations. Figures 4A–F show representative comparisons derived from specimen 200316 to help justify the comparisons we chose. The auto fluorescence (AutoF) image (Figure 4A) is frequently used to drive registration to the AutoF image in the ABA. NeuN (Figure 4B) and Myelin basis protein (i.e., MBP, Figure 4C) are of particular interest to our work in aging. The DWI (Figure 4D) is created by averaging all the (registered) diffusion weighted images producing high contrast to noise with many anatomic landmarks throughout the volume. Cortical layer definition and contrast in the dentate gyrus are particularly high in this volume. There are strong similarities between NeuN (Figure 4B) and DWI (Figure 4D). The FA image (Figure 4E) is a logical choice as it highlights white matter. The RD image (Figure 4F) is a putative marker of myelin integrity that might map well to the MBP.

3.2.1. Comparison of p6_03H and p6_07H

Two pipes were chosen for more careful comparison: p6_03H and p6_07H. Because of the similarities between NeuN

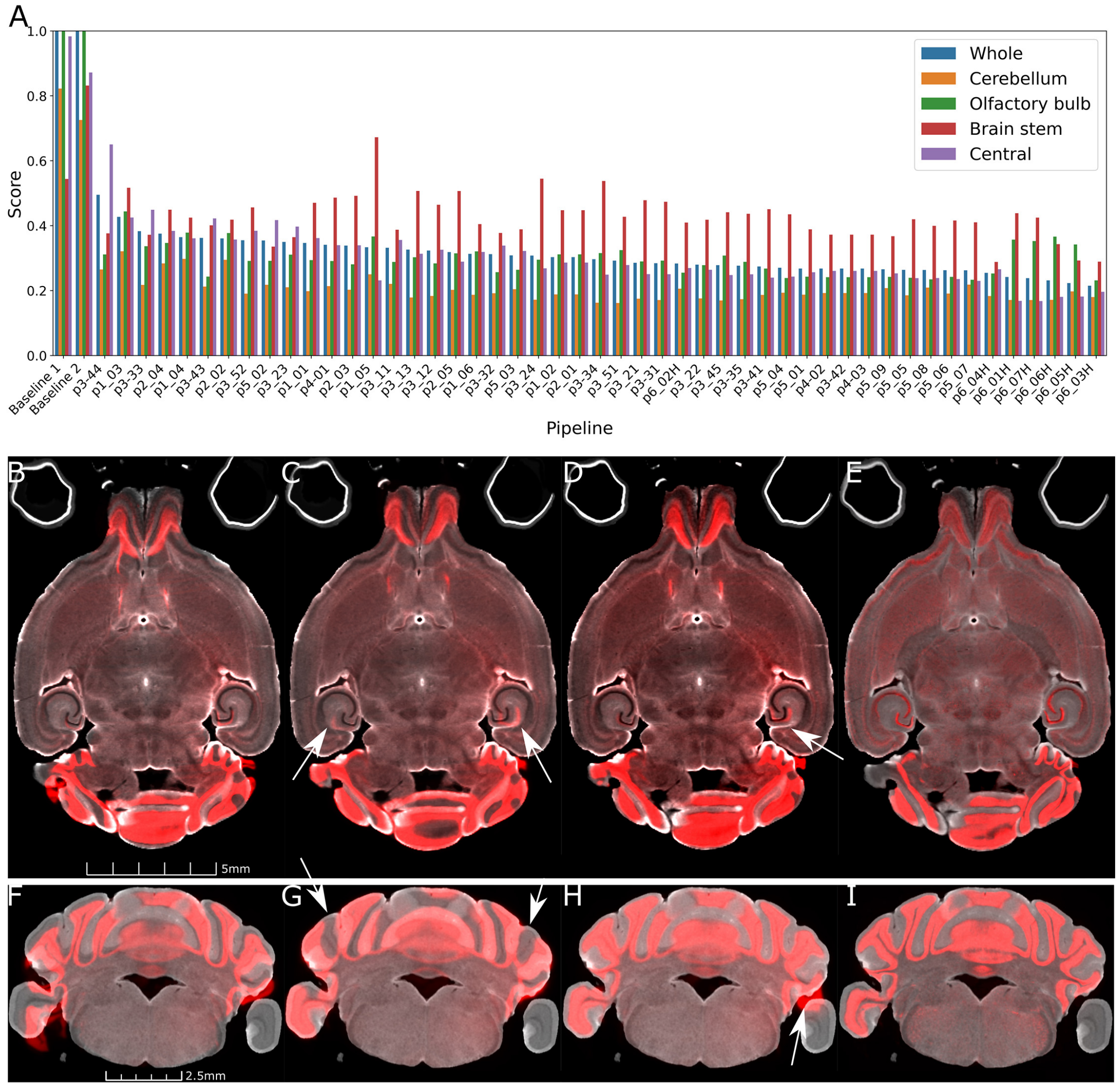


FIGURE 3
Demonstration of the range of results derived from the varied pipes. **(A)** Shows the L2 norm for the pipes listed in Tables 2, 3 registering Syto16 to DWI (specimen 191209). **(B)** Shows that the initialization results in reasonable alignment in the central slice. But **(F)** shows that initialization fails in the distal slices in the cerebellum. **(C,D,G,H)** Show results at 45 μm with L2 norms of 0.361 using pipe P2_02. There are still significant errors in the cerebellum (arrows in panel **G**). **(D,H)** With pipe P3_42 performs better with a lower L2 norm of 0.268. Finally, a comparison of panels **(D,H)** (@ 45 μm) and **(E,I)** (@ 15 μm) with pipeline P6_07_H demonstrates the utility of performing the registration using the higher resolution data. The cerebellar slice in panels **(F–I)** highlights a frequent problem i.e., loss of the paraflocuIoss from handling. The broken symmetry in the data gives rise to asymmetric misalignment (arrows in panels **C,D,H**).

and DWI, this combination was chosen to evaluate these two pipes in three different specimens. Supplementary Figure 1 and Supplementary Table 1 summarize the comparison. P6_03H is faster than p6_07H and for one specimen (191209) yielded a lower L2 norm. The resulting volumes were imported into Imaris to allow interactive review of the relative success of the registration across the entire volume. Supplementary Figure 1 demonstrates that p6_03H yields consistent subtle misregistration in the dentate gyrus that is absent in p6_07H.

3.2.2. Relative success of multiple combinations

Supplementary Table 2 summarizes an exhaustive comparison of p6_07H across five specimens with 15 different pairs of images. Specimen 200316 with the largest number (200) of fiducials was run twice with different initializations. Specimens 190108 and 191209 are from the BXD series providing a strain with different anatomy than the B6. Comparison of the L2 norms between specimens is not

TABLE 3 Optimization of pipeline @ 15 μ m resolution.

Pipeline	Composition	Score	Time
Stage 6			
P6_01_H	--shrink-factor 10 × 7 × 4 × 1	0.2419	3 d 17 h
P6_02_H	Coarser affine --shrink-factor 30 × 21 × 12 × 1	0.2834	6 d 12 h
P6_03_H	--shrink-factor 30 × 21 × 12 × 3	0.2147	2 h 29 m
P6_04_H	--shrink-factor 30 × 21 × 12 × 1	0.2544	6 d 12 h
P6_05_H	--shrink-factor 40 × 28 × 16 × 4	0.2232	2 h 43 min
P6_06_H	--shrink-factor 20 × 14 × 8 × 2	0.2314	18 h 12 m
P6_07_H	--shrink-factor 10 × 7 × 4 × 2	0.2382	10 h 29 m

The shrink factors in the b-spline and SyN stages are the main variables to be optimized.

appropriate since each specimen has a different set of fiducials. This highlights some of the limitations in using fiducials as a quantitative metric for comparison of the quality of a registration. The precision of fiducial pairs will be biased by the reader placing the pairs. This results in a lower (nonzero) level which will vary between specimens that is dependent on the reader/fiducial e.g., an average error of 135 μm for the NeuN/DWI combination for specimen 200803 with 51 fiducials and 235 μm for specimen 191209 with 175 fiducials. However, comparison of the L2 norms across the different registration combinations within a specimen can provide useful insight into which pairs provide the best registration. For example, mapping MBP to RD is one of the least successful combinations. Mapping NeuN to DWI or Syto to DWI yields one of the lower L2 norms for all the specimens. The duplicate comparison for specimen 200316 highlights the stochastic nature of the registration with a 12% difference in the L2 norm (NeuN+DWI) between the two runs, but the relative scores of varied combinations of mapping remain unchanged.

One of the more surprising results is the success of the AutoF/DWI combination. **Supplementary Figure 2** shows the results of registration using the pipe p6_07H with two image combinations: AutoF to DWI and NeuN to DWI with specimen 200316. The transforms generated with the AutoF to DWI registration was then applied to the NeuN. The registered pairs (NeuN to DWI) for both transforms were interactively reviewed in Imaris to discern areas in which the transforms differed. The target image (DWI) is displayed in yellow, and the moving image (NeuN) is displayed in green. In **Supplementary Figure 2A** (NeuN to DWI) there are subtle errors in alignment in the cerebellum that are not evident in the autoF/DWI pair. Yet the internal structures e.g., the dentate gyrus seem to be comparable. Comparison of the moving images C) NeuN or D) AutoF, highlight the high contrast granular layer in the NeuN image and the relatively flat contrast in the AutoF image. The high contrast in this granular layer dominates the registration since the NeuN stain in the outer edge of the brain is nonexistent. Registration using the AutoF is more successful since the contrast in the cerebellum is quite flat. This highlights one of the most challenging aspects of this task i.e., the registration of two volumes with completely different sources of contrast.

The NeuN/DWI combination has become our standard method since many of our planned studies require insight into neuronal density. Landmark comparison of the vessels in the NeuN to DWI registration was undertaken using Imaris as described in the section “2.5 Registration validation” to gauge the quality of registration away from the edges. The process was executed on 11 different vessels spread throughout the brain. The mean displacement was 22 ± 14 μ m.

3.3. Volume corrections to LSM

The most common way of delineating brain regions on an cleared brain image is *via* registration to an atlas (Kutten et al., 2016; Tappan et al., 2019; Perens et al., 2021) or registration of the atlas to the volume under study (Goubran et al., 2019). The most commonly used atlas is the ABA i.e., the CCFv3 3D template constructed from a population of 1,675 young adult B6 brains using AutoF (Wang et al., 2020). In **Figure 5**, we used our MRH atlas to estimate the regional volume changes in the LSM images from tissue swelling in specimen 190108. This specimen (111 day BXD 89) is representative of our broader interest- understanding the genetic basis for age related changes in the BXD family (Ashbrook et al., 2021). We registered the NeuN to DWI for specimen 190108 using the final registration pipeline. Labels were registered to the DWI of specimen 190108 from our reference B6 atlas (200302) using our MRH registration pipeline (Anderson et al., 2019). The transform that was generated was inverted to transform the labels on the DWI back to the uncorrected NeuN volume. **Figures 5A, B** shows the NeuN volume before and after correction, respectively. Note the changes in the width is larger than the change in length highlighting the nonuniform distortion. This is even more apparent in **Figures 5C, D** which shows a sagittal cross section before and after correction.

Figure 6 summarizes the change in volume for the 50 largest regions of interest. We have used the reduced set of labels (rCCFv3) defined in Johnson et al. (2022). The nomenclature is consistent with CCFv3. The magnitude and variability are significant. The olfactory bulb (OB) is nearly 80% larger in the uncorrected data while the corpus callosum (cc) is ∼10% smaller. The problem is compounded when comparing specimens as the differential swelling varies, and it varies considerably between different clearing methods. These variations must impact the shape of the structures.

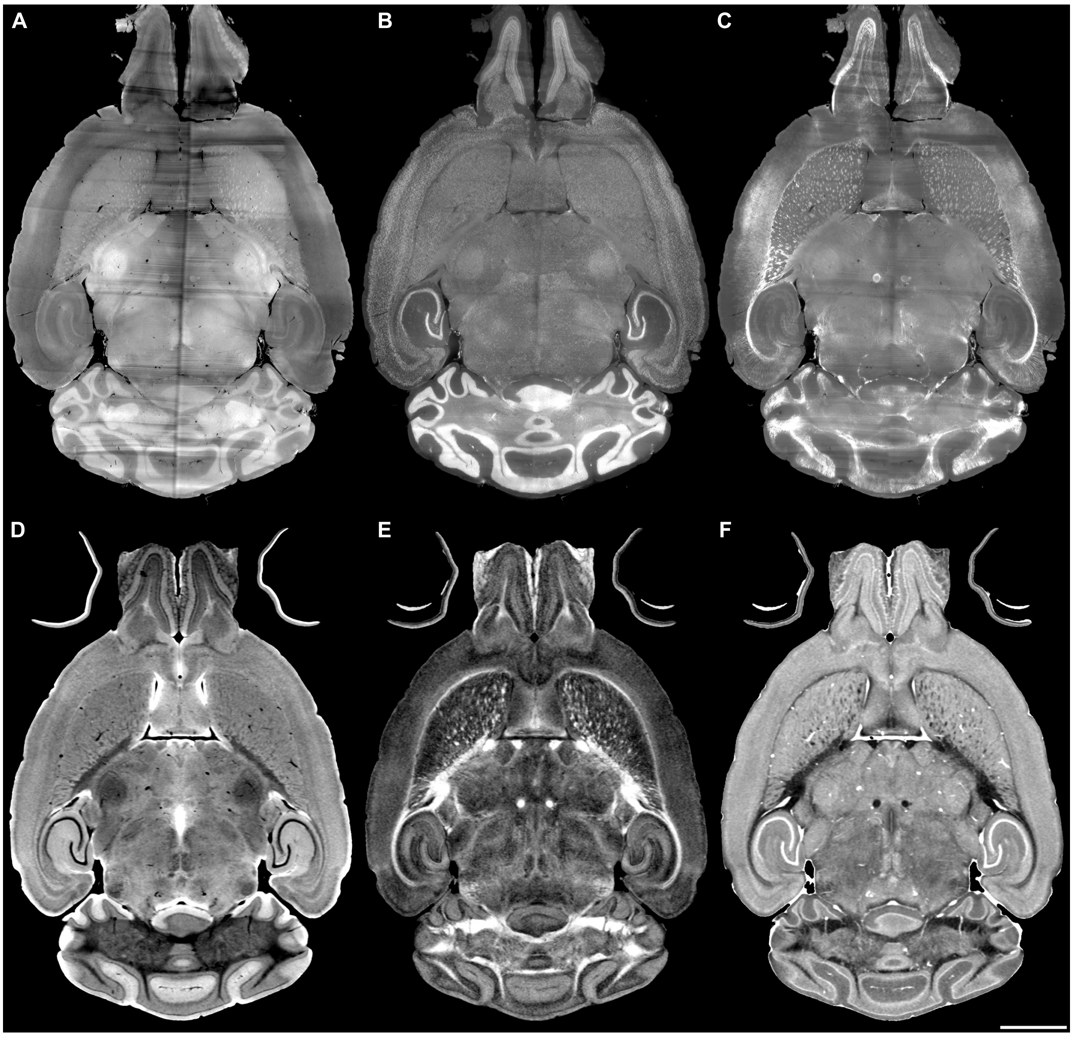


FIGURE 4
Light sheet microscopy of different stains and MRH of different contrasts. **(A)** Auto fluorescent, **(B)** NeuN, **(C)** MBP, **(D)** DWI, **(E)** FA, **(F)** RD scale bar is 2 mm (specimen 200316).

Supplementary Figure 4 demonstrates the impact on the non-uniform distortion on the hippocampus, a region of particular interest in age related neurodegeneration (Sabuncu et al., 2011; Katabathula et al., 2021). **Supplementary Figure 6** demonstrates the variability of deformation in 30 brain regions across multiple specimens.

4. Limitations

Registration of LSM to the MRH of the same specimen improves the geometric accuracy over existing methods of registration to the Allen Brain Atlas as demonstrated in **Figure 6**. But there are limitations. While the MRH data are acquired with the brain in the skull they are not a perfect match to the *in vivo* scan. Ma et al. (2005, 2008) have compared *in vivo* and *ex vivo* scans. They are significant with volume difference between *in vivo* and *ex vivo* (out of skull) varying from +60% (fimbria) to −79% (ventricles). The majority of this difference arises from removing the brain from the cranial vault. Our images have been acquired with the brain in the skull which reduces this problem. But the ventricles are collapsed and there may be shrinkage due to fixation. Inspection of the data before skull stripping has demonstrated no measurable separation of the brain surface from the skull so the shrinkage from fixation is limited. But ventricle distortion remains a limitation. An additional source of uncertainty arises from the transfer of the label from our canonical MRH atlas to any new MRH data using our SAMBA pipeline (Anderson et al., 2019). The accuracy and precision of the pipeline are dependent on the tuning parameters of the pipeline and the morphologic differences between the unknown specimen to

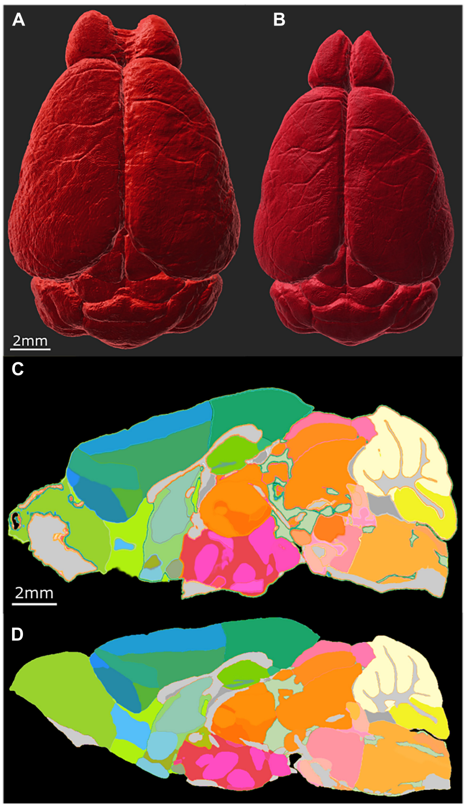


FIGURE 5
Distortion correction of the LSM data by registration to the MRI of the same specimen (190108). **(A)** Surface rendering of uncorrected LSM volume and **(B)** corrected LSM volume. The scale bar is 2 mm. **(C)** Midsagittal section of the labels on the LSM data before correction; **(D)** midsagittal section after correction: the scale bar in panels **(C,D)** is 2 mm. The distortion is present both within the plane of section and across the plane making it difficult to define identical planes. The highlighted edges in panel **(C)** are an interpolation artifact.

which labels are mapped and the canonical atlas. We are confronted with the fact that the atlas is constructed from a B6 as is the ABA. But the tests performed in validating the atlas included a systematic variation of inputs using a synthetic model with varied anatomy and a real world source of variation based on a model of stroke causing significant volume changes in several structures in the brain. With appropriate selection of the SAMBA registration parameters ROC analysis showed area under the curve (AUC) better than 0.99.

5. Discussion

This work was initiated to enable combined analysis of cells and circuits from MRH and LSM in the same specimen. We have developed a method to register the LSM images which allow us to count cells to MRH, which maintains brain morphology inside the skull more closely approximating that in a live animal. Transferring labels from the MRH to the corrected LSM data allows us to measure regional cell densities with much greater accuracy than previous methods.

We addressed several challenges in correcting the significant and irregular distortion in the LSM; registration between fundamentally different images with significant differences in contrast; registration of very large volumes (300 GB). We have employed an initialization involving $\sim$ 20 landmarks followed by pipeline with multiple stages of transformations and metrics to minimize a user customized L2 norm score.

From the optimization, we selected the registration workflow with a combined consideration on accuracy and time. The optimized workflow (pipeline p6_07) takes an average of 7.5 h on a computer with 2 64-core processors and 2TB RAM with page faulting, with the L2 norm of 135 μm. The workflow shows robustness in multiple specimens. Our approach takes advantage of the high spatial and contrast resolution in the MRH images to provide internal landmarks the drive the registration locally across the whole brain which is evident from the small mean displacement (~22 μm) of fiducials, which are picked at the junctures of vessels in both MRH and LSM.

As both MRH and LSM include varied contrasts (**Figure 4**), we did experiments to find the best combination of different

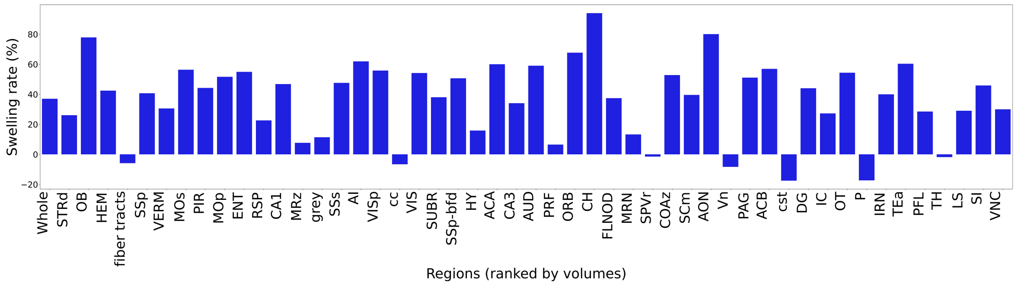


FIGURE 6
Bar plot of ratio of the volume before and after registration. The regions are ranked by the ROI volume. The ratio is obtained by $\frac{V_{before\ reg} - V_{after\ reg}}{V_{MR}}$. The abbreviations of the regions are based on rCCFv3 (Johnson et al., 2022) with a labeling convention consistent with CCFv3.

diffusion scalar images and immunohistochemistry with LSM. A surprising conclusion is that registrations between DWI and AutoF or NeuN are similarly good. The practical consequence for our use is that we will not have to acquire an AutoF image freeing up a channel in the LSM for a more useful cytoarchitectural measure i.e., NeuN.

Multiple groups have developed methods for automated labeling of 3D optical images from cleared mouse brains (Kutten et al., 2016; Perens et al., 2021). These approaches rely on the Allen Brain Atlas as the reference (Wang et al., 2020). We are interested in mapping the age-related changes across multiple strains (for both genders). Registration of these data to the young adult male C57 that is the core of the ABA could obscure the morphologic changes of interest. Renier et al. (2016) used MRI of a fixed mouse brain to measure the degree of distortion from tissue processing with iDisco but their MRH images were of a half brain taken with a relatively low contrast gradient echo out of the skull. Labeling relied on mapping the autofluorescence image to the ABA. The MRI was not used in this step. Goubran et al. (2019) have developed a pipeline that is similar to that which we report here. Our work differs from their approach in four ways. Our dMRI protocols acquire data @ 15 µm vs 200 µm i.e., a difference in voxel volume of 2370 X with the commensurate challenge of larger image arrays. As demonstrated in **Figures 3E, I**, registration with the full resolution MRH (15 µm) makes a difference. **Supplementary Table 2** provides an excellent starting point for evaluation of many of the alternatives. Finally, our pipeline takes advantage of a truly isotropic 3D MRH atlas of the brain in the skull to which rCCF3 labels have been mapped. Our approach provides an efficient method for segmenting brain regions in LSM data mapped in the MRH space of the same specimen which will allow quantitative study of cytoarchitecture e.g., cell density along with connectivity. The contrast study also would be a fruitful area for the further work. For example, a broader study could consider synthesizing synthetic contrast from combinations of scalar dMRI images that might contain complementary information or using machine learning to transferring the contrast from LSM to MRH to reduce the registration difficulty due to different contrast distributions (Sedghi et al., 2021). Artificial intelligence may well provide new avenues to improve the registration quality and efficiency (Fu et al., 2020; Sedghi et al., 2021).

Data availability statement

The original contributions presented in this study are included in this article/**Supplementary material**, further inquiries can be directed to the corresponding author.

Ethics statement

The animal study was reviewed and approved by the Duke Institutional Animal Care and Use Committee.

Author contributions

YT, JC, and GJ contributed to the conception and design of the study. YT performed the investigation and implementation and wrote the original draft. JC and GJ organized the database. GJ reviewed and edited the submitted version and supervised and funded the study. All authors contributed to the manuscript revision and approved the submitted version.

Funding

This work was supported by the National Institute of Aging R01AG070913 (to GJ and Robert W. Williams).

Acknowledgments

We are grateful to Leonard White Ph.D. in the Duke Department of Neurobiology and Robert W. Williams Ph.D. University of Tennessee Health Science Center Chair of Genetics for helpful guidance in neuroanatomy. We thank Tatiana Johnson for technical assistance in preparing 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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2022.1011895/full#supplementary-material

References

Anderson, R. J., Cook, J. J., Delpratt, N., Nouls, J. C., Gu, B., McNamara, J. O., et al. (2019). Small animal multivariate brain analysis (SAMBA)–a high throughput pipeline with a validation framework. *Neuroinformatics* 17, 451–472. doi: 10.1007/s12021-018-9410-0

Ashbrook, D. G., Arends, D., Prins, P., Mulligan, M. K., Roy, S., Williams, E. G., et al. (2021). A platform for experimental precision medicine: The extended BXD mouse family. *Cell Syst.* 12, 235–247.e9. doi: 10.1016/j.cels.2020.12.002

Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* 12, 26–41. doi: 10.1016/j.media.2007.06.004

Casanova, M. F., El-Baz, A., Mott, M., Mannheim, G., Hassan, H., Fahmi, R., et al. (2009). Reduced gyral window and corpus callosum size in autism: Possible macroscopic correlates of a minicolumnopathy. *J. Autism Dev. Disord.* 39, 751–764. doi: 10.1007/s10803-008-0681-4

Egaas, B., Courchesne, E., and Saitoh, O. (1995). Reduced size of corpus callosum in autism. *Arch. Neurol.* 52, 794–801. doi: 10.1001/archneur.1995.00540320070014

Erturk, A., Becker, K., Jahrling, N., Mauch, C. P., Hojer, C. D., Egen, J. G., et al. (2012). Three-dimensional imaging of solvent-cleared organs using 3DISCO. *Nat. Protoc.* 7, 1983–1995. doi: 10.1038/nprot.2012.119

Eylers, V. V., Maudsley, A. A., Bronzlik, P., Dellani, P. R., Lanfermann, H., and Ding, X. Q. (2016). Detection of normal aging effects on human brain metabolite concentrations and microstructure with whole-brain MR spectroscopic imaging and quantitative MR imaging. *AJNR Am. J. Neuroradiol.* 37, 447–454. doi: 10.3174/ajnr.A4557

Fornito, A., Zalesky, A., and Breakspear, M. (2015). The connectomics of brain disorders. *Nat. Rev. Neurosci.* 16, 159–172. doi: 10.1038/nrn3901

Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., and Yang, X. (2020). Deep learning in medical image registration: A review. *Phys. Med. Biol.* 65:20TR01. doi: 10.1088/1361-6560/ab843e

Goubran, M., Leuze, C., Hsueh, B., Aswendt, M., Ye, L., Tian, Q., et al. (2019). Multimodal image registration and connectivity analysis for integration of connectomic data from microscopy to MRI. *Nat. Commun.* 10:5504. doi: 10.1038/s41467-019-13374-0

Hardan, A. Y., Minshew, N. J., and Keshavan, M. S. (2000). Corpus callosum size in autism. *Neurology* 55, 1033–1036. doi: 10.1212/WNL.55.7.1033

Johnson, G. A., Ali-Sharief, A., Badea, A., Brandenburg, J., Cofer, G., Fubara, B., et al. (2007). High-throughput morphologic phenotyping of the mouse brain with magnetic resonance histology. *Neuroimage* 37, 82–89. doi: 10.1016/j.neuroimage.2007.05.013

Johnson, G. A., Benveniste, H., Black, R. D., Hedlund, L. W., Maronpot, R. R., and Smith, B. R. (1993). Histology by magnetic resonance microscopy. *Magn. Reson. Q.* 9, 1–30.

Johnson, G. A., Tian, Y., Cofer, G. P., Cook, J. C., Gee, J. C., Hall, A., et al. (2022). HiDiver: A suite of methods to merge magnetic resonance histology, light sheet microscopy, and complete brain delineations. *bioRxiv* [Preprint]. doi: 10.1101/2022.02.10.479607

Johnson, G. A., Wang, N., Anderson, R. J., Chen, M., Cofer, G. P., Gee, J. C., et al. (2019). Whole mouse brain connectomics. *J. Comp. Neurol.* 527, 2146–2157. doi: 10.1002/cne.24560

Katabathula, S., Wang, Q. Y., and Xu, R. (2021). Predict Alzheimer's disease using hippocampus MRI data: A lightweight 3D deep convolutional network model with visual and global shape representations. *Alzheimers Res. Ther.* 13:104. doi: 10.1186/s13195-021-00837-0

Klein, S., Staring, M., Murphy, K., Viergever, M. A., and Pluim, J. P. (2010). elastix: A toolbox for intensity-based medical image registration. *IEEE Trans. Med. Imaging* 29, 196–205. doi: 10.1109/TMI.2009.2035616

Kutten, K. S., Vogelstein, J. T., Charon, N., Ye, L., Deisseroth, K., and Miller, M. I. (2016). Deformably registering and annotating whole CLARITY brains to an atlas via masked LDDMM. *arXiv* [Preprint]. doi: 10.1117/12.2227444

Loomba, N., Beckerson, M. E., Ammons, C. J., Maximo, J. O., and Kana, R. K. (2021). Corpus callosum size and homotopic connectivity in Autism spectrum disorder. *Psychiatry Res. Neuroimaging* 313:111301. doi: 10.1016/j.pscychresns.2021.111301

Ma, Y., Hof, P. R., Grant, S. C., Blackband, S. J., Bennett, R., Slatest, L., et al. (2005). A three-dimensional digital atlas database of the adult C57BL/6J mouse brain by magnetic resonance microscopy. *Neuroscience* 135, 1203–1215. doi: 10.1016/j.neuroscience.2005.07.014

Ma, Y., Smith, D., Hof, P. R., Foerster, B., Hamilton, S., Blackband, S. J., et al. (2008). In vivo 3D digital atlas database of the adult C57BL/6J mouse brain by magnetic resonance microscopy. *Front. Neuroanat.* 2:1. doi: 10.3389/neuro.05.001.2008

Murray, E., Cho, J. H., Goodwin, D., Ku, T., Swaney, J., Kim, S. Y., et al. (2015). Simple, scalable proteomic imaging for high-dimensional profiling of intact systems. *Cell* 163, 1500–1514. doi: 10.1016/j.cell.2015.11.025

Park, Y. G., Sohn, C. H., Chen, R., McCue, M., Yun, D. H., Drummond, G. T., et al. (2019). Protection of tissue physicochemical properties using polyfunctional crosslinkers. *Nat. Biotechnol.* 37, 73–83. doi: 10.1038/nbt.4281

Perens, J., Salinas, C. G., Skytte, J. L., Roostalu, U., Dahl, A. B., Dyrby, T. B., et al. (2021). An optimized mouse brain atlas for automated mapping and quantification of neuronal activity using iDISCO plus and light sheet fluorescence microscopy. *Neuroinformatics* 19, 433–446. doi: 10.1007/s12021-020-09490-8

Renier, N., Adams, E. L., Kirst, C., Wu, Z., Azevedo, R., Kohl, J., et al. (2016). Mapping of brain activity by automated volume analysis of immediate early genes. *Cell* 165, 1789–1802. doi: 10.1016/j.cell.2016.05.007

Sabuncu, M. R., Desikan, R. S., Sepulcre, J., Yeo, B. T., Liu, H., Schmansky, N. J., et al. (2011). The dynamics of cortical and hippocampal atrophy in Alzheimer disease. *Arch. Neurol.* 68, 1040–1048. doi: 10.1001/archneurol.2011.167

Schmitz, B., Wang, X., Barker, P. B., Pilatus, U., Bronzlik, P., Dadak, M., et al. (2018). Effects of aging on the human brain: A proton and phosphorus MR spectroscopy study at 3T. *J. Neuroimaging* 28, 416–421. doi: 10.1111/jon.12514

Sedghi, A., O'Donnell, L. J., Kapur, T., Learned-Miller, E., Mousavi, P., and Wells, W. M. III (2021). Image registration: Maximum likelihood, minimum entropy and deep learning. *Med. Image Anal.* 69:101939. doi: 10.1016/j.media.2020.101939

Tappan, S. J., Eastwood, B. S., O'Connor, N., Wang, Q., Ng, L., Feng, D., et al. (2019). Automatic navigation system for the mouse brain. *J. Comp. Neurol.* 527, 2200–2211. doi: 10.1002/cne.24635

Tepest, R., Jacobi, E., Gawronski, A., Krug, B., Moller-Hartmann, W., Lehnhardt, F. G., et al. (2010). Corpus callosum size in adults with high-functioning autism and the relevance of gender. *Psychiatry Res.* 183, 38–43. doi: 10.1016/j.pscychresns.2010.04.007

Tuor, U. I., Morgunov, M., Sule, M., Qiao, M., Clark, D., Rushforth, D., et al. (2014). Cellular correlates of longitudinal diffusion tensor imaging of axonal degeneration following hypoxic-ischemic cerebral infarction in neonatal rats. *Neuroimage Clin.* 6, 32–42. doi: 10.1016/j.nicl.2014.08.003

Tustison, N. J., and Avants, B. B. (2013). Explicit B-spline regularization in diffeomorphic image registration. *Front. Neuroinform.* 7:39. doi: 10.3389/fninf.2013.00039

Vemuri, P., and Jack, C. R. Jr. (2010). Role of structural MRI in Alzheimer's disease. *Alzheimers Res. Ther.* 2:23. doi: 10.1186/alzrt47

Wang, N., Anderson, R. J., Badea, A., Cofer, G., Dibb, R., Qi, Y., et al. (2018a). Whole mouse brain structural connectomics using magnetic resonance histology. *Brain Struct. Funct.* 223, 4323–4335. doi: 10.1007/s00429-018-1750-x

Wang, N., Cofer, G., Anderson, R. J., Qi, Y., Liu, C., and Johnson, G. A. (2018b). Accelerating quantitative susceptibility imaging acquisition using compressed sensing. *Phys. Med. Biol.* 63:245002. doi: 10.1088/1361-6560/aaf15d

Wang, Q., Ding, S. L., Li, Y., Royall, J., Feng, D., Lesnar, P., et al. (2020). The allen mouse brain common coordinate framework: A 3D reference atlas. *Cell* 181, 936–953.e20. doi: 10.1016/j.cell.2020.04.007

Weishaupt, N., Zhang, A., Deziel, R. A., Tasker, R. A., and Whitehead, S. N. (2016). prefrontal ischemia in the rat leads to secondary damage and inflammation in remote gray and white matter regions. *Front. Neurosci.* 10:81. doi: 10.3389/fnins.2016.00081

Yeh, F. C., Wedeen, V. J., and Tseng, W. Y. I. (2010). Generalized q-sampling imaging. *IEEE Trans. Med. Imaging* 29, 1626–1635. doi: 10.1109/TMI.2010.2045126

Zhang, J. H., Badaut, J., Tang, J., Obenaus, A., Hartman, R., and Pearce, W. J. (2012). The vascular neural network–a new paradigm in stroke pathophysiology. *Nat. Rev. Neurol.* 8, 711–716. doi: 10.1038/nrneurol.2012.210

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 09 January 2023
DOI 10.3389/fnins.2022.1062889

OPEN ACCESS

EDITED BY
Xun Chen,
University of Science and Technology of China, China

REVIEWED BY
Juan Cheng,
Hefei University of Technology, China
Jing Liu,
Hebei Normal University, China

*CORRESPONDENCE
Xin Zhang
✉ xin_zhang_bme@tju.edu.cn
Dong Ming
✉ richard-ming@tju.edu.cn

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 06 October 2022
ACCEPTED 02 December 2022
PUBLISHED 09 January 2023

CITATION
Li Y, Zhang X and Ming D (2023) Early-stage fusion of EEG and fNIRS improves classification of motor imagery. *Front. Neurosci.* 16:1062889. doi: 10.3389/fnins.2022.1062889

Early-stage fusion of EEG and fNIRS improves classification of motor imagery

Yang Li[1], Xin Zhang[2,3]* and Dong Ming[2,3]*

[1]Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin, China, [2]The Laboratory of Neural Engineering and Rehabilitation, Department of Biomedical Engineering, School of Precision Instruments and Optoelectronics Engineering, Tianjin University, Tianjin, China, [3]The Tianjin International Joint Research Center for Neural Engineering, Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin, China

Introduction: Many research papers have reported successful implementation of hybrid brain-computer interfaces by complementarily combining EEG and fNIRS, to improve classification performance. However, modality or feature fusion of EEG and fNIRS was usually designed for specific user cases, which were generally customized and hard to be generalized. How to effectively utilize information from the two modalities was still unclear.

Methods: In this paper, we conducted a study to investigate the stage of bi-modal fusion based on EEG and fNIRS. A Y-shaped neural network was proposed and evaluated on an open dataset, which fuses the bimodal information in different stages.

Results: The results suggests that the early-stage fusion of EEG and fNIRS have significantly higher performance compared to middle-stage and late-stage fusion network configuration ($N = 57$, $P < 0.05$). With the proposed framework, the average accuracy of 29 participants reaches 76.21% in the left-or-right hand motor imagery task in leave-one-out cross-validation, using bi-modal data as network inputs respectively, which is in the same level as the state-of-the-art hybrid BCI methods based on EEG and fNIRS data.

KEYWORDS

EEG, fNIRS, hybrid-BCI, modality fusion, motor imagery

1. Introduction

Brain–computer interfaces (BCIs) are communication systems that utilize control signals generated by the brain to interact with the surrounding environment without the participation of the peripheral nervous system and muscles (Nicolas-Alonso and Gomez-Gil, 2012). These years have witnessed thriving progress in the field of BCI. Motor imagery (MI) is one of the common paradigms in BCI research (Kaiser et al., 2011), which is accomplished by imagining performing the given task (Jeannerod, 1995), such as grabbing (Herath and Mel, 2021), lifting (Kasemsumran and Boonchieng, 2019), and so on. MI-BCIs are widely used to aid patients with motor function impairments caused by stroke (Ang et al., 2010), amyotrophic lateral sclerosis (Lulé et al., 2007), spinal cord injury (Cramer et al., 2007), and so on, either for daily-life assistance or rehabilitative training. Since motor imagery tasks induce event-related desynchronization and synchronization (ERD/ERS) in EEG (Jeon et al., 2011), various feature extraction algorithms have been designed to detect ERD/ERS activities in EEG (Kee et al., 2017; Selim et al., 2018; Sadiq et al., 2019; Dagdevir and Tokmakci, 2021). However, due to its nonstationary nature, EEG is considered as bio-signals of extremely low signal-to-noise ratio with spatial ambiguity and distortion (Hallez et al., 2007). EEG feature extraction process, which is highly dependent on prior knowledge, is challenged by its high time complexity, imposing the risk of information loss (Zhang et al., 2018, 2021). Many researchers turned to deep learning methods for EEG feature extraction. For example, Schirrmeister et al. (2017) proposed an end-to-end learning network called ConvNets that was able to learn the spectral power modulation of different frequency bands and produce accurate spatial mapping for learned features. Lawhern et al. (2018) proposed a compact convolutional neural network to accurately decode EEG recorded from various paradigms.

The low spatial resolution characteristic of EEG leads to challenges in the accurate localization of cortical activation sources despite the fact that EEG signals are the most widely used bio-signals in BCIs (Liu et al., 2021). Due to its disadvantage in spatial resolution, some researchers attempted to incorporate the information from functional near-infrared spectroscopy (fNIRS) data to improve the performance of BCIs (Pfurtscheller, 2010; Fazli et al., 2012; Buccino et al., 2016). fNIRS measures oxygenated and deoxygenated hemoglobin (HbO and HbR) using near-infrared light (Fazli et al., 2012). On the one hand, the fusion of EEG and fNIRS has technical support because the electrophysiological signal and the inner edge light signal are not affecting each other. On the other hand, fNIRS-based BCIs are most commonly of the active type, where users react purposefully and independently (Khan and Hong, 2017). Therefore, plenty of mental tasks exploit fNIRS signals to assess brain status, which have proven to be effective in previous studies (Hong et al., 2015). Yin et al. introduced joint mutual information (JMI) to combine features and optimize BCIs, which was used to classify MI tasks with different strengths and speeds when clenching a fist. JMI reached an accuracy of 89 ± 2% with 1–5% improvement compared to using EEG or fNIRS alone. Al-Shargie et al. applied canonical correlation analysis to decode EEG-fNIRS and maximized the correlation between EEG and fNIRS to classify the influence of psychological stress on the prefrontal cortex (Al-Shargie et al., 2017). Sun et al. used tensor fusion and p-order polynomial fusion with deep learning technologies, which improved the accuracy at the cost of increased computational complexity and reduced the stability (Sun et al., 2020).

There are relatively mature methods and a relatively clear consensus for dealing with multimodal fusion problems in the field of computer vision. Depending on those methods, the researchers combined features in the early or late stage to achieve the best results. For example, Aygün et al. (2018) adapted various fusion methods, which were previously used in video recognition problems, to solve the brain tumor segmentation problem and conducted the related experiments in the BRATS dataset in the early, middle, and late fusion methods. A Y-shape network is widely used in tasks with multimodal inputs. The multimodal models usually have their own encoders on each modality. For example, the image encoder and the language encoder form a twin tower structure model that is used for loss calculation in CLIP, which is a training structure of language–image multimodal fusion (Radford et al., 2021). Lan et al. (2019) used a Y-shaped network to combine two encoders with the path of one decoder and extract more information from raw data. As a result, the Y-shape network is extremely helpful for data reconstruction and multimodal fusion. However, in the field of biomedical signal processing, there is no consensus on the processing of physiological signals from different modalities. Fusion of EEG and fNIRS information is conducted mostly arbitrarily at the feature level, which has been proven to be suitable for several specific user cases. When and how to effectively combine the bimodality data is still unclear. This study conducted experiments on an open dataset. A compact Y-shaped ANN architecture has been proposed and validated to investigate the EEG-fNIRS fusion methods and strategies. The main framework of EEGNet is used in the EEG processing branch, which is a proven successful framework for EEG data analysis. As the temporal resolution of fNIRS is low and minimal frequency information is present, only the second and third modules of EEGNet are used in the fNIRS processing branch. The results suggest that neural networks with EEG-fNIRS features integrated at an early stage demonstrated statistically higher accuracy. The final classification accuracy of the proposed method reaches 76.21%, which is at the same level compared to the state-of-the-art on the investigated open dataset in discriminating left and right motor imagery.

This article is organized as follows. In the "Materials and methods" section, the dataset is briefly introduced, and the preprocessing method and the proposed framework are demonstrated in detail. In the "Results" section, the results are presented. In the "Discussion" section, an in-depth discussion is presented. In the "Conclusion" section, conclusions are presented.

2. Materials and methods

2.1. Datasets

Shin et al. released two publicly available datasets of EEG-fNIRS multimodal, which were Dataset A, left-hand motor imagery and right-hand motor imagery, and Dataset B, mental arithmetic and relax imagery (Shin et al., 2017). The primary focus of this study was MI classification, and Dataset A was used to conduct a series of experiments and analyze further in this study.

For Dataset A, there were 29 participants (14 men and 15 women), all of whom had minimal experience with motor imagery experiments. In the experiment, a black arrow pointing to the left or right was shown in the middle of the screen for the first 2 s. Then, the arrow disappeared and a fixed black cross was shown on the screen for 10 s. All the participants were instructed to perform kinesthetic motor imagery at a speed of approximately 1 repetition per s, such as imagining a designated hand opening and closing as if they were grasping a ball, followed by a rest period of 10–12 s. Finally, there were 30 trials for each task of each participant. Common spatial pattern (CSP) features of EEG data and the mean and slope values of fNIRS signals were extracted from the data as features. A sliding window was used to conduct 10 × 5-fold cross-validation on the unimodal data and bimodal data, respectively, with window size set to 3 s, step size set to 1 s, and the range of sliding window set to between 5 s before the cue and 20 s after the cue. sLDA was used as a classifier to classify data between left and right motor imagery tasks.

In their article, the average classification accuracy of the 10 × 5-fold cross-validation under each window was considered as the classification accuracy of this window. In addition, the maximum classification accuracy among all the windows was regarded as the final classification accuracy for each participant. The highest classification accuracy of EEG-only was about 65%, and the highest classification accuracy of unimodal classification was HbO-fNIRS, which was approximately 57% according to the resulting figure.

2.2. Pre-processing

For dataset A, the EEG was recorded using a BrainAmp EEG amplifier, with the sampling rate set to 200 Hz in the original dataset. First, the data were downsampled from 200 to 128 Hz, and the channels related to EOG were removed for later analysis. Then, the EEG data were re-referenced to the common average reference. A band-pass filter with a frequency range of 8–25 Hz was applied to remove noise, leaving the μ-band and low-β band data unmodified. Since we wanted to focus on channels related to the sensorimotor cortex and maintain the correspondence with fNIRS optical channels, eight relevant electrodes were chosen around the sensorimotor cortex, namely, FCC5 h, FCC3 h, CCP5 h, CCP3 h, FCC4 h, FCC6 h, CCP4 h, and CCP6 h (shown in Figure 1). The amplitude of the signals was normalized to [−1, 1] for subsequent processing.

The sampling rate of the fNIRS signal was set to 10 Hz in the original dataset. First, the data were up-sampled from 10 Hz to 128 Hz to be consistent with EEG data (Abtahi et al., 2020). We chose eight optical channels (6 emitters and 6 detectors with 3-cm optrode separation) around the sensorimotor cortex, i.e., FC3-FC5, FC3-FC1, C5-C3, C1-C3, FC4-FC2, FC4-FC6, C2-C4, and C6-C4 (shown in Figure 1), whose positions corresponded to the selected EEG channel locations, to ensure spatial consistency of the recorded data. The modified Beer–Lambert law was used to convert the raw light intensity data to the relative oxyhemoglobin and deoxyhemoglobin concentrations. Then, a band-pass filter with a frequency range of 0.01–0.1 Hz was used to remove the effect of physiological noises such as heartbeat, breath, and other artifacts. We extracted 10 s data during the task period, and data from 5 s to 2 s before the visual cue were used to remove the baseline. Finally, the amplitude of the signals was normalized to [−1, 1] for subsequent processing.

2.3. Fusion network

The basic network structure is inspired by the EEGNet (Lawhern et al., 2018). The original EEGNet is composed of three modules. The first module is a temporal-domain convolution layer through which the time-frequency features of the signals are constructed. The kernel size is set to (1, *fs*//2), where *fs* is the sample rate of signals, and the sign // denotes the rounding operation. The second module is depth-wise convolution through which spatial filters are generated and more task-related channels are selected by the convolution kernel, where the kernel size is set to (N_{chan}, 1), where N_{chan} denotes the number of EEG channels. Average pooling is used to down-sample the feature dimension. The third module is a separable convolution layer, which consists of depth-wise convolution and pointwise convolution.

In this study, we followed EEGNet architecture with three complete modules for EEG data. For fNIRS, we only used the second and third modules since the fNIRS signals did not contain much information in the frequency domain due to the low sampling rate, and the features are mostly extracted from the time domain. At the end of the Y-shaped network, a SoftMax layer was used as a classifier to generate the outputs. The complete network architecture is shown in Figure 2.

In the literature, one of the commonly used fusion methods is the concatenation of features from each modality (Baltrusaitis et al., 2019). The network architecture is shown in Figure 2. In this study, three similar networks are proposed to investigate the effect of fusing bimodal features in different stages, i.e., before depth-wise convolution (referred to as E_0_Net and E_1_Net, please see Figure 2A), before separable convolution (referred to as M_0_Net and M_1_Net, please see Figure 2B), and before flatten layer (referred to as L_0_Net and L_1_Net, please see Figure 2C), where E, M, and L represent early stage fusion, middle-stage fusion, and late-stage fusion, respectively, and numbers 0 or 1 represent concatenation fusion performed at the depth-dimension or the channel-dimension. In this study, to maintain the integrity of the original design of EEGNet, the proposed network used the same hyperparameters as proposed in the original EEGNet paper (Lawhern et al., 2018); only the kernel size of the first layers in EEG branch was tuned as we had

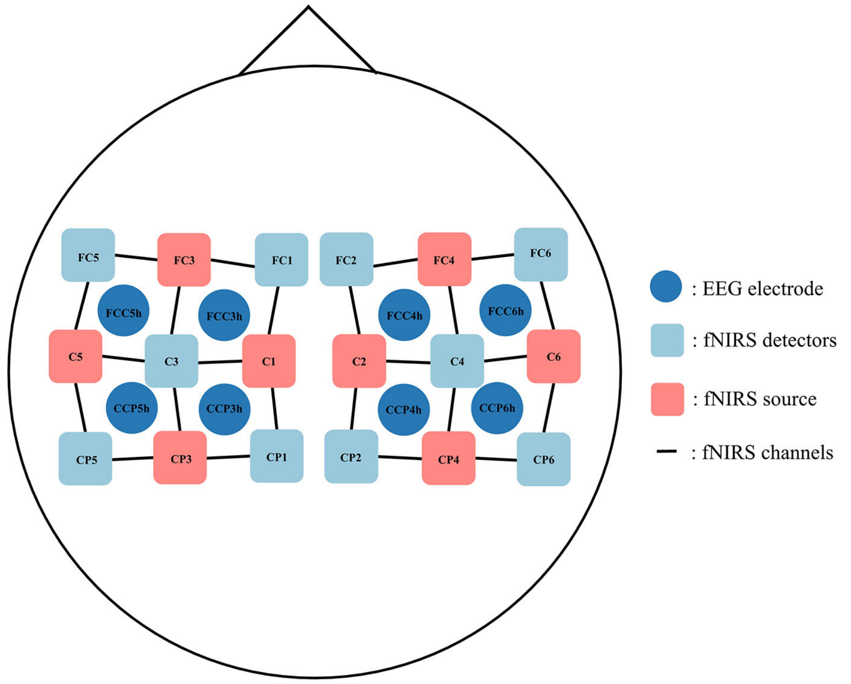


FIGURE 1
The placement of EEG electrodes, fNIRS sources, and fNIRS detectors.

a very limited number of trials in the investigated open dataset and a larger kernel size created more parameters to be learned. We tried kernel sizes of (1, 34), (1, 44), (1, 54), and (1, 64) to perform temporal-domain convolution on EEG data to have the best model performance during the training process.

Table 1 summarizes the number of network parameters with different fusion strategies. The numbers of neural network parameters for different fusion methods are similar, except for M_0_Net. For the open dataset used in this study, the amount of data from a single participant in one particular type of task is too small when using networks with a large number of parameters. Therefore, the use of a lightweight network can alleviate overfitting to a certain extent.

2.4. Model training

Early stopping is a form of regularization that prevents overfitting by stopping the iteration number. When training error decreases quickly, we hope that the model continues to be trained and that the generalization losses have a higher chance of being "repaired". In this study, we used an early stopping criterion that assumes that overfitting does not begin until the error decreases slowly. The algorithm is shown in Equations (1) and (2), referred to from Prechelt (2012). In this study, we did not use a validation dataset and we used an early stopping strategy to reduce jitter.

$$P_k = (t)\ 1000\cdot\left(\frac{\sum_{t'=t-k+1}^{t} E_{tr}\left(t'\right)}{k\cdot\min_{t'=t-k+1}^{t} E_{tr}\left(t'\right)} - 1\right) \qquad (1)$$

$$P_k\,(t) < \alpha \qquad (2)$$

where k is the training strip, Etr is the training error, and α is the threshold value. When $Pk(t)$ is less than α, we think that it is time to stop. In this study, k is 10 and α is 0.001.

Due to the limitation in the amount of data, for each participant, there were only 30 trials for each motor imagery task in the open dataset. A data augmentation method designed for long-interval EEG-fNIRS hybrid BCI applications was used to expand the size of the dataset. Due to the limitation in response time of fNIRS signals, the time intervals between experiment tasks were more than 10 s. Therefore, data augmentation can be achieved by repetitively sampling sub-trials from a single trial.

In this study, two training strategies were adopted. For training Strategy A, the window size was set to a 3-s time window, and the step size was set to 3 s. Then, each 10-s trial was divided into 3 sub-trials without overlapping. Therefore, the number of trials for each participant from one task was expanded to 90 trials and was used for neural network training. All the sub-trials were randomly shuffled before the train-test segmentation of data. The data were then randomly divided into an 80% training set and a 20% testing set. The proposed neural

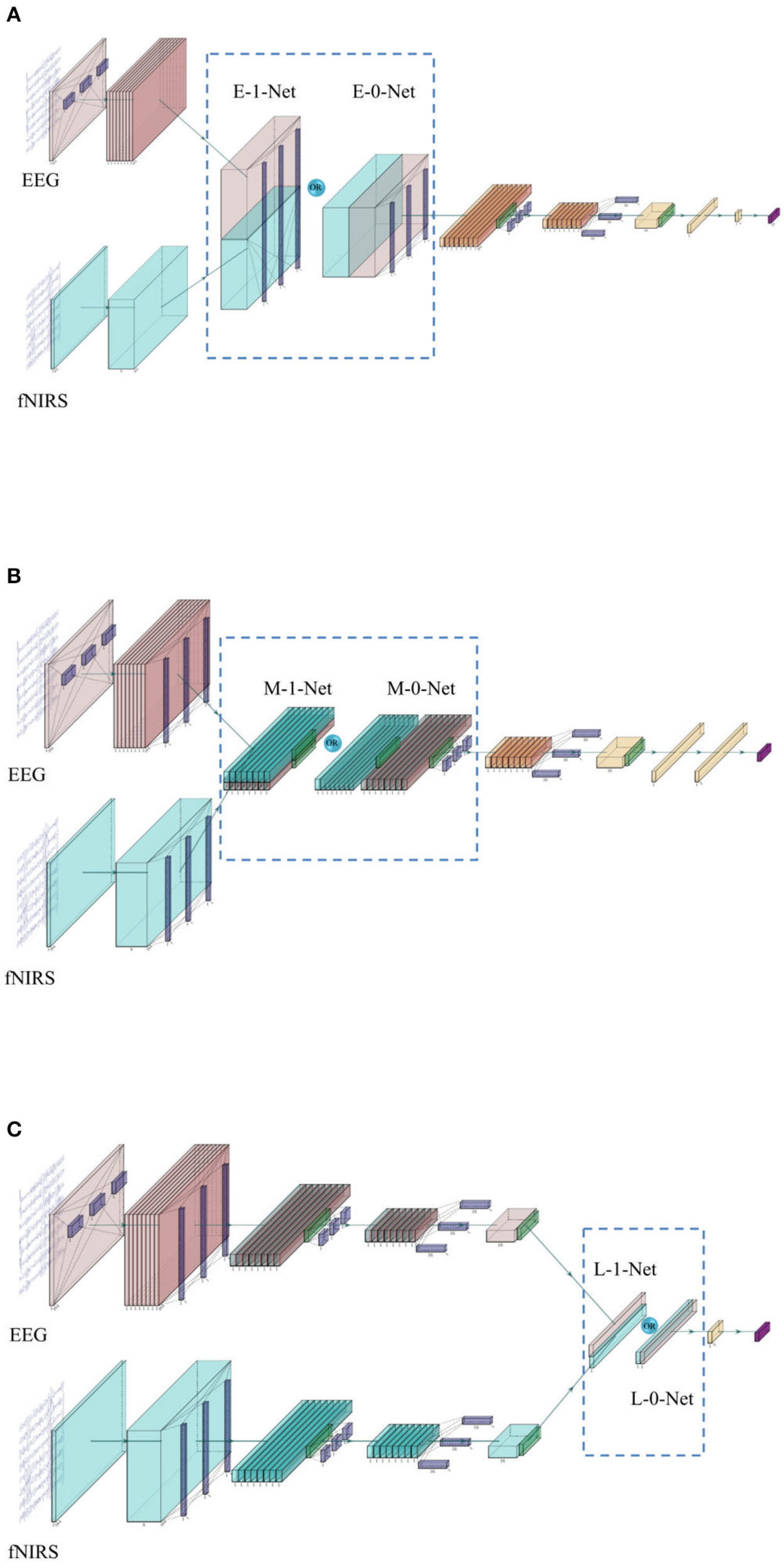


FIGURE 2
(A) The network architecture of early stage fusion, which is referred to as E-0-Net and E-1-Net in the following contents, depends on whether the concatenation was performed on the 1st or 2nd dimension. **(B)** The network architecture of middle-stage fusion, which is referred to as M-1-Net and M-0-Net in the following contents. **(C)** The network architecture of late-stage fusion, which is referred to as L-1-Net and L-0-Net in the following content.

TABLE 1 The number of all parameters.

Method	Number of parameters
E_0_Net	3,792
E_1_Net	3,792
M_0_Net	8,880
M_1_Net	4,172
L_0_Net	4,176
L_1_Net	4,176

networks either were trained for 500 epochs or met the early stopping criteria.

For training Strategy B, we used leave-one-out cross-validation for each participant. The voting method was used to train the network with the idea of decision fusion. We divided each trial with a window size of 3 s and a step size of 1 s. Each trial was divided into 8 sub-trials. In the training set and testing set, two overlapping sub-trials from the same trial did not appear at the same time. The data from the open dataset were further expanded without the training set leakage. During the training process, all data were randomly shuffled. The proposed neural networks were either trained for 500 iterations or met the early stopping criteria.

2.5. Voting mechanism

Ensemble learning is one of the most popular research topics (Wozniak et al., 2014). It extracts a set of features through a diversity of projections on data using multiple machine learning algorithms and performs various transformations of features. Then, various classification algorithms are used to generate prediction results based on the extracted features. Information from the abovementioned results is integrated to achieve better performances than information obtained from any stand-alone algorithm (Dong et al., 2020). For classification tasks, the voting method is often used to improve the final results (Zhou, 2012). One of the commonly used voting combinations is the majority voting combination, where the predicted results of most are considered as the final output. The voting algorithm is shown in Equation (3).

$$\hat{y}=\begin{cases}1, & n_{\hat{y}i}>n_{\hat{y}0}\\ 0, & n_{\hat{y}i}<n_{\hat{y}0}\end{cases} \tag{3}$$

where $n_{\hat{y}i}$ is the number of test samples with its predicted results being 1. $n_{\hat{y}0}$ is the number of test samples with its predicted results being 0, and $\hat{y}$ is the final predicted result of this trial.

In this study, we used a sliding window of 3 s with a step size of 1 s. Therefore, each trial is divided into 8 sub-trials. Then, a leave-one-out cross-validation scheme was used to test the model performance for each actual trial after data augmentation from each participant. The predicted results of the majority voting combination of 8 sub-trials were the final prediction results of one trial.

3. Results

3.1. Data augmentation

Deep convolutional neural networks have achieved outstanding performance in many areas, which is driven by improvements both in computational power and the availability of large datasets. However, it is extremely difficult to acquire or collect large datasets for lots of application fields, such as datasets of physiological signals. If a small dataset was used to train a model with a large number of parameters, overfitting would happen, resulting in poor generalization performance. In the related studies on computer vision, overfitting can be alleviated by data augmentation, such as geometric transformation, random cropping, feature space manipulation, adversarial training, and so on, to improve the model performance and expand its limited dataset (Shorten and Khoshgoftaar, 2019).

In the dataset investigated in this study, there were only 30 trials in each task for each participant in this open dataset, and each trial is 10 s long. Data augmentation was used in the model generation to improve the model performance. The data from one trial of 10 s were truncated to three trials as 0–3 s, 3–6 s, and 6–9 s without overlap. Through the augmentation process, the original dataset was expanded to three times its original size.

We selected unimodal data (EEG-only and HbO-only) without data augmentation from different time windows using the method from the original dataset study (Shin et al., 2017) and chose the average accuracy of all participants among different time windows as average accuracy for statistical analysis. The CSP algorithm was used to extract features from augmented EEG data, the mean and slope features were extracted from HbO-fNIRS, and sLDA was used as a classifier to generate a classification model. The highest classification accuracy of left-right motor imagery classification with only EEG data (referred to as EEG-only in the following content) was 66.09%, and the highest classification accuracy of left-right motor imagery classification with only HbO data (referred to as fNRIS-only in the following content) was 54.31%, which were similar to the results of the original study. After data augmentation, the highest average accuracy for EEG-only reached 69.25% and for fNIRS-only reached 58.33%. The average accuracy improved for both EEG-only and fNIRS-only. It can be seen from Figure 3 that the classification accuracy of 65.52% of participants improved for EEG and that 72.41% of participants improved for fNIRS compared with the original data. This method of data augmentation not only expands the dataset but also improves the classification performance.

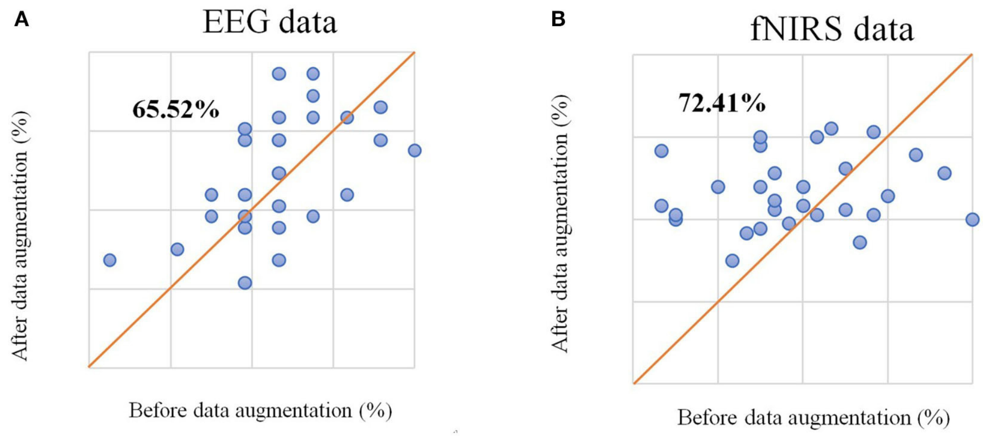


FIGURE 3
(A) The scatter plot of classification accuracy of each participant based on EEG signals before and after data augmentation. **(B)** The scatter plot of classification accuracy of each participant based on the fNIRS signal before and after data augmentation.

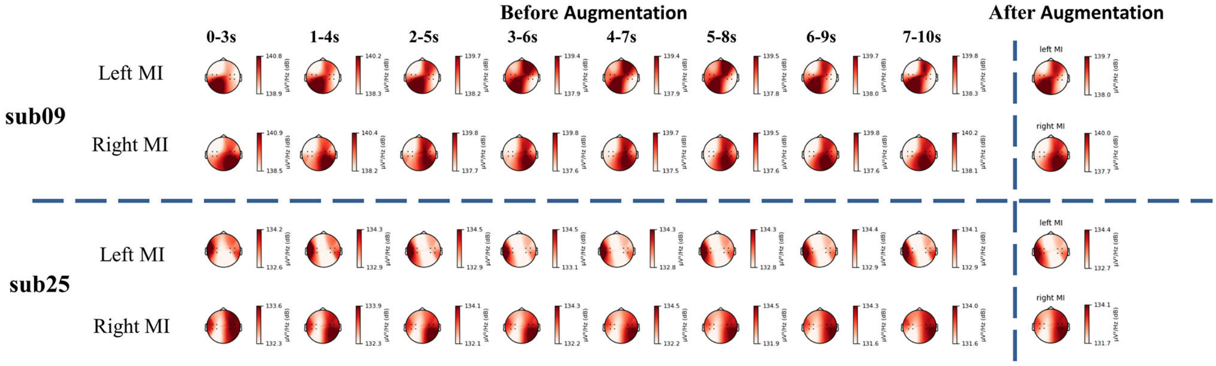


FIGURE 4
The PSD analysis of EEG signal. The PSD analysis of sub09 is shown above the horizontal dashed line and that of sub25 is shown below the horizontal dashed line.

Similarly, we used artificial neural networks to classify EEG and fNIRS data from different tasks. The average accuracy of all participants is 65.00%. Sub01, sub09, sub16, sub25, sub26, and sub27 demonstrated good classification performance using EEG data with a classification accuracy of more than 80%. Participants with top model classification performance (sub09 and sub25) were analyzed with power spectral density (PSD) (shown in Figure 4). Clear EEG power lateralization was identified both before data augmentation and after data augmentation. The proposed method of data augmentation can maintain the original temporal-spatial characteristic in the EEG data.

For fNIRS-HbO, the average accuracy of the lightweight network of all participants is 63.13%. Sub09, sub19, sub20, sub21, sub24, and sub29 were able to demonstrate good classification performance using fNIRS, with the classification accuracy reaching more than 70%. We used the cerebral oxygen exchange (COE, where COE value = HbO – HbR) (Naseer and Hong, 2015) as input and selected the participants with the top performances (sub20 and sub21) for temporal-domain analysis. As shown in Figure 5, before data augmentation, clear lateralization of the COE values can be identified in both left-hand and right-hand motor imagery tasks: the COE values of the left channels were significantly higher than that of the right channels during the left-hand motor imagery, and COE values of the right channels were significantly higher than that of the left channels during the right-hand motor imagery, which was consistent with the results presented in the literature (Asahi et al., 2004; Hétu et al., 2013). At the same time, data augmentation with a sliding window of 3 s with a 1 s step size also demonstrated similar lateralization characteristics, as shown in Figure 5. The proposed data augmentation method did not

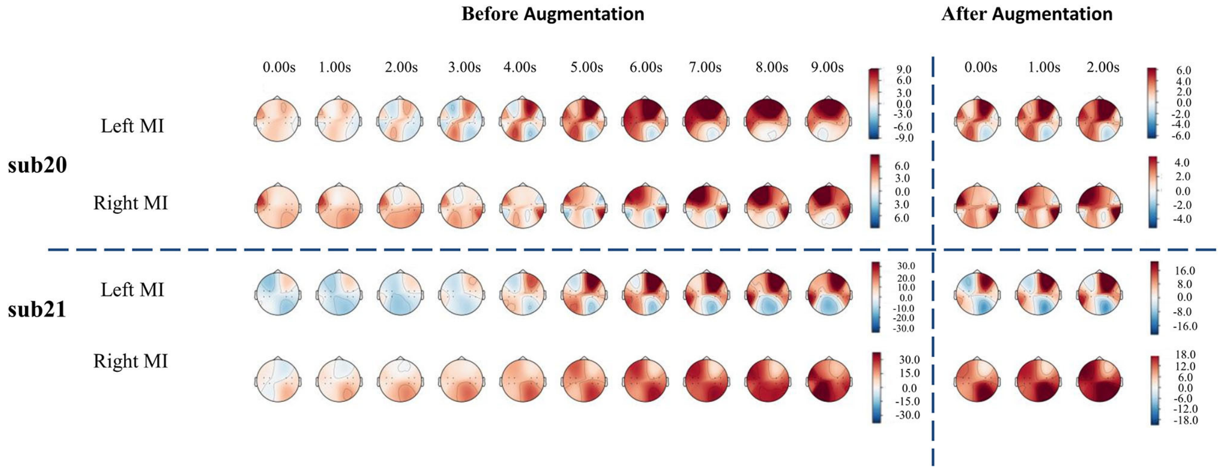


FIGURE 5
The temporal-domain analysis of the fNIRS signal. The temporal-domain analysis of sub20 is above the horizontal dashed line, and that of sub21 is below the horizontal dashed line.

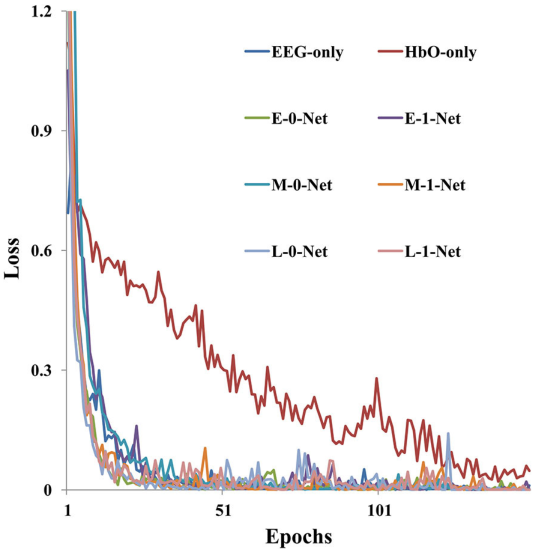


FIGURE 6
The model training loss (sub09 as an example) varies with the number of epochs under different network architectures.

disturb the temporal-spatial characteristics of original fNIRS data and maintained good consistency.

3.2. Model generation

As shown in Figure 6, model training loss varied with the number of iterations under different network architectures. It was clear that the value of training loss reduced as the number of epochs increased. In addition, the convergence speed was the lowest when fNIRS-only data were used for model generation, which required 150 epochs before convergence. However, for EEG-only data, the convergence speed was faster than fNIRS-only data, and the training reached convergence within 50 epochs. The model converged faster with a bimodal fusion network than with a single-modality network.

3.3. Test results

Due to the obvious temporal characteristic difference between EEG and fNIRS, we attempted different sizes of kernels (parameters used for temporal-domain convolution) during the comparison of fusion results at different stages. We divided the results into four groups at different stages, namely, (1, 34), (1, 44), (1, 54), and (1, 64). In Table 2 for different kernel sizes, the classification accuracies of the two methods of early stage fusion were significantly higher than that of other fusion methods. In addition, the accuracies of the two early fusion methods were both within the range of 69–70%. For middle-stage fusion methods, the classification accuracies ranged from 65 to 66%. For late-stage fusion methods, the classification accuracies ranged from 62 to 63% (see Figure 7). Since none of these results conformed to a normal distribution, the Wilcoxon signed-rank test was adopted to investigate the statistical significance. Table 2 summarizes the results of the significance analysis of different fusion methods. We observed that *p*-values between early stage fusion and middle-stage fusion or for late-stage fusion were all below 0.05 regardless of the kernel size, which represents the statistical significance of the performance difference between the early stage fusion method and other stage fusion methods.

TABLE 2 Statistical analysis results between the different fusion methods.

HbO	Kernel size	Early-mid (*N* = 57)	Early-late (*N* = 57)	Mid-late (*N* = 57)	Dim_0-Dim_1 [E-M] (*N* = 28)
P-values	(1,34)	0.0007	0.0011	0.3550	0.0566
	(1,44)	0. 0003	0.0013	0.7033	0.2801
	(1,54)	0.0007	0.0007	0.5872	0.2864
	(1,64)	2.8884	0.0001	0.2066	0.2594
HbR	**Kernel size**	**Early-mid (*N* = 57)**	**Early-late (*N* = 57)**	**Mid-late (*N* = 57)**	**Dim_0-Dim_1 [E-M] (*N* = 28)**
P-values	(1,34)	0.0024	0.0014	0.5217	0.0540
	(1,44)	0.0033	0.0045	0.3226	0.0257
	(1,54)	0.0025	0.0034	0.6294	0.3470
	(1,64)	0.0062	0.0008	0.4265	0.0809

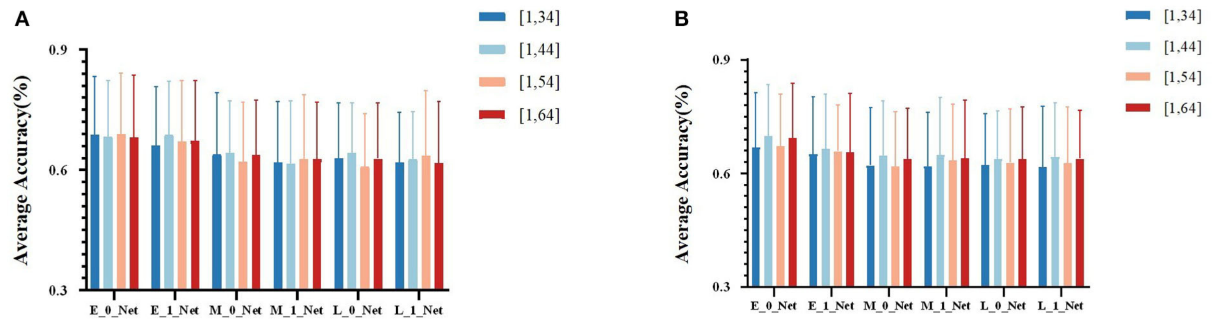


FIGURE 7
(A) The average classification accuracy of all participants using EEG-HbO in different fusion strategies. **(B)** The average classification accuracy of all participants using EEG-HbR in different fusion strategies.

The performance of early stage fusion was significantly higher than late-stage fusion. We also optimized the proposed bimodal fusion network to achieve the best classification performance further. We further optimized the size of the pooling layer and the number of convolution filters to optimize the model performance. For the pooling layer, we searched from (1, 4) to (1, 16), with (1, 4) as the step size and four options in total. Other hyperparameters were still the same as in the original paper of EEGNet. The optimal parameters are shown in Table 3. In addition, the optimal average accuracy was 71.60% and the standard deviation was 1.42% using EEG-HbO with E-N-0. The optimal average accuracy was 71.21%, and the variance was 1.88% using EEG-HbR with E-N-0.

3.4. Ablation analysis

The proposed fusion network architecture consisted of a temporal convolution layer, spatial convolution layer, and separable convolution layer, where the temporal convolution layer learned the time-frequency feature of each channel, the spatial convolution layer selected and extracted the spatial pattern of interesting channels, and separable convolution layer extracted global joint features and facilitated the design of a relatively lightweight network for small datasets. Feature fusion was conducted in these three modules, through which early fusion, middle fusion, and late fusion were configured and investigated. Ablation analysis was conducted to further explore the significance of multimodal fusion. We conducted the ablation experiments based on training strategy A and training strategy B, respectively.

First, we optimized the proposed bimodal fusion network to achieve the best classification performance. For training strategy A, we can conclude that, when using EEG-only, the average accuracy of all participants was 65.00% and the standard deviation was 2.11% using EEGNet with the same hyperparameters related to EEG in the bimodal process. When using HbO-fNIRS, the average accuracy was 63.13% and the standard deviation was 0.57% using ANN (consists of spatial convolution layer and separable convolution layer)

TABLE 3 Parameter table.

Block	Layer	#filters	Size	Activation	Options
1	Conv2D	8	(1, 54)		Padding
2	DepthwiseConv2D	2*8	(8, 1)		
	Activation			ELU	
	AveragePool2D		(1, 4) for EEG+Hbo /(1, 16) for EEG+Hbr		
	Dropout				$P = 0.2$
3	SeparableConv2D	2*8	(1, 8)		padding
	Activation			ELU	
	AveragePool2D		(1, 16)for EEG+Hbo /(1, 12)for EEG+Hbr		
	Dropout				$P = 0.2$

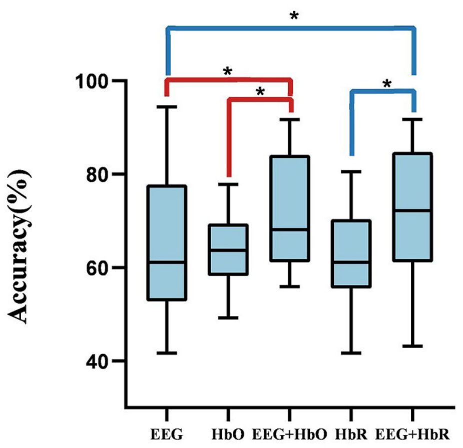


FIGURE 8
The comparison between unimodality and bimodality. A *P*-value of <0.05 and the performance of bimodal fusion is significantly superior to that of unimodality.

with the same hyperparameters related to HbO in bimodal process. When using HbR-fNIRS, the average accuracy was 62.43% and the standard deviation was 1.08% using ANN with the same hyperparameters related to HbR in a bimodal process. The average accuracy was 71.60% using EEG-HbO with E-N-0. The average accuracy was 71.21% using EEG-HbR with E-N-0. Statistical analysis was performed by using the Wilcoxon signed-rank test to compare the performance of unimodality with that of bimodality. As shown in Figure 8, it was found that *P*-values were below 0.05, and a *P*-value of below 0.05 was regarded as statistically significant. The results are summarized in Figure 8, which demonstrated a consistent and significant model performance improvement, with the introduction of the other modalities. Multimodal fusion can complement advantages of each modality and improve classification performance significantly.

3.5. Voting results

In this study, we divided each trial into 8 overlapping sub-trials and used the majority voting method to achieve the final result of each trial. We used training strategy B to train networks and used the same hyperparameter to perform 500 epochs. As shown in Figure 9, during the leave-one-out analysis, the average accuracy without the voting mechanism was 72.13% and the standard deviation was 0.1391, while the average accuracy with the voting mechanism was 76.21% and the standard deviation was 0.1611.

4. Discussion

Bimodal fusion methods demonstrated higher performance than that of unimodal data, whether using traditional machine learning methods with feature extraction classification schemes or deep learning methods with an end-to-end learning process. The heterogeneity between EEG and fNIRS data does exist; however, the heterogeneity is not as high as we thought based on the signal sources. In addition, incorporating special methods for bimodal fusion boosts BCI performance.

The classification results of the proposed Y-shape model are summarized in Figure 7, which contains conditions with different types of fNIRS data (Hbo vs. Hbr) and different kernel sizes on the first layers of the EEG branch. Figure 7A showed the fusion of EEG and Hbo, and Figure 7B showed the fusion of EEG and Hbr. According to the consistent performance of the two types of fNIRS information, models with an early fusion of EEG and fNIRS data have better classification accuracy than those of other stages, regardless of the size of the kernel and fNIRS data type. Comparing two types of fNIRS data, models with Hbo as input demonstrated higher resilience of model hyperparameter than the models using Hbr as input, although these two types of information were inherently correlated by the mechanism of blood supply in the human brain. Hbr data might be more

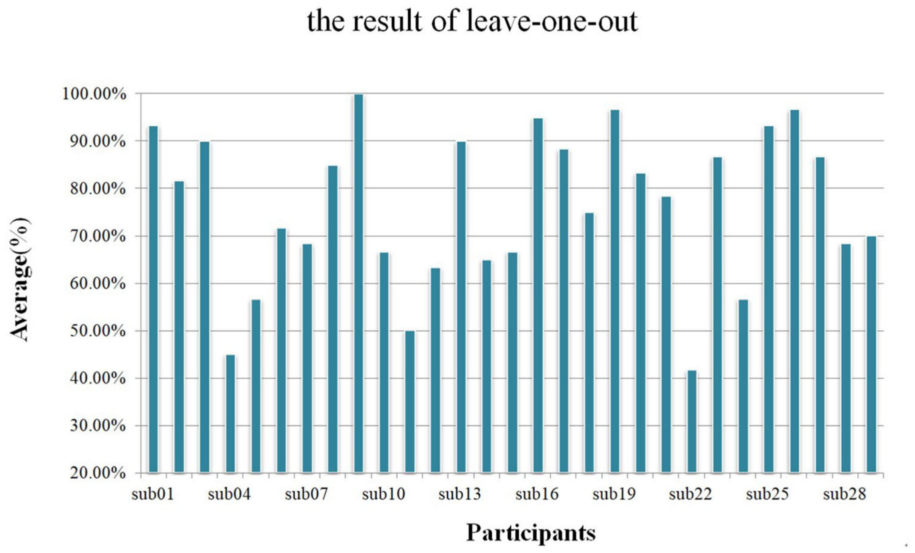


FIGURE 9
Using a leave-one-out analysis scheme, the average accuracy of each participant with the voting mechanism.

sensitive to subtle changes in brain activities, which introduced more irrelevant activities other than motor imagery.

The common analysis for fNIRS signals was limited to the temporal-domain features such as the mean value, slope, peak, and so on, due to the low sampling rate. However, these features might not be informative enough to reflect the overall and detailed characteristics of fNIRS signals. Thus, the resultant information loss deteriorated the classification performance. We noticed that the classification accuracy of each participant with temporal convolution was lower than that without temporal convolution in fNIRS models using deep learning methods, which is consistent with our prior knowledge of fNIRS signals. In addition, the classification accuracies using deep learning methods for fNIRS signals (HbO-only, 63.13%) were better than that using traditional machine learning with handcrafted features and a predefined learning model (HbO-only, 58.33%), which demonstrated the superiority of the deep learning methods in the field of BCI research.

The average accuracy for the EEG-only model was 66.09% using traditional machine learning techniques, and for the HbO-only model, the average accuracy was 54.31% without data augmentation. With data augmentation, the highest average accuracy for EEG-only was 69.25%, and for the HbO-only model, the highest average accuracy was 58.33%. With data augmentation combined with deep learning methods, the highest average accuracy for the EEG-only model was 65.00%, and the HbO-only model was 63.13%. Therefore, the size of the dataset had a great impact on the classification performance of left-vs.-right MI tasks. An effective data augmentation method was able to boost model performance and improve generalization. The data augmentation method we propose in this study is valid and effective, especially for long recording interval paradigms when integrating EEG and fNIRS data.

Based on the classification results from different networks, it was clear that the early fusion techniques demonstrated significant positive impacts on the bimodal MI classification. A slight decreasing trend was observed with early, middle, and late fusion methods, respectively (shown in Figure 7 and Table 2). Although all three of these networks were feature-level fusion, the difference in model performance might be a compound effect of the heterogeneity of data, the level of feature (high-level features vs. low-level features), and bimodal co-adapted learning. Early stage fusion of bimodal data might have added additional constraints on the learning process and subsequently regularized the two feature extraction branches in the Y-shaped network. It seemed that early fusion could mitigate the loss of information. In previous studies in computer vision, it was suggested that multimodal data with higher heterogeneity tend to have better performance in late-fusion models, while multimodal data with low heterogeneity tend to perform better in the early fusion in the field of medical image (Ramachandram and Taylor, 2017; Mogadala et al., 2021; Yan et al., 2021). The heterogeneity of EEG and fNIRS might not be as high as we expected since they were able to be fused in the temporal domain, although these two types of data were recorded from completely different signal sources. However, this phenomenon was preliminarily observed and validated with only one open dataset due to limited access to bimodal BCI datasets of EEG and fNIRS; further analysis with more datasets should be done in the future. In addition, it was interesting that no statistically significant difference was found between middle-stage fusion and late-stage fusion, which might be caused by insufficient complementary features. The temporal-spatial feature of EEG and the spatial feature of fNIRS were extremely important.

In the dataset investigated in this study (Shin et al., 2017), there were only 30 trials for each task of each participant. The major limitation in the amount of data severely limited the scale of the neural network as well as the final classification performance. Future studies should be done to validate the conclusions in this study with a large bimodal EEG-fNIRS dataset. In addition, compared to the classification results in the literature, the proposed framework showed the same level of performance compared to that of the state-of-the-art methods (ours at 76.21 vs. 78.59% in the literature) (Kwak et al., 2022). More advanced learning techniques should be investigated to further improve the performance of the proposed network.

5. Conclusion

In this study, bimodal fusion methods of EEG and fNIRS were investigated with an open dataset. Compact Y-shaped ANN architectures are proposed and validated to investigate EEG-fNIRS fusion methods and strategies. The main framework of EEGNet is used in the proposed network. The results suggested that networks with EEG-fNIRS features integrated at an early stage demonstrated statistically higher accuracy compared to the other fusion methods in motor imagery classification tasks, which partially suggested that the heterogeneity of EEG and fNIRS might be relatively low despite the fact that these two types of signals were acquired from different sources. With the proposed framework, the final classification accuracy of the proposed method reached 76.21%, which was at the same level compared to the state-of-the-art on an EEG-fNIRS hybrid BCI open dataset in discriminating left and right motor imagery.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: 10.1109/TNSRE.2016.2628057.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

YL conceived and designed the experiments, analyzed the experimental data, and wrote the manuscript. XZ conceived the experiments, guided the experiments, and participated in this study in the process of manuscript drafting and revision. DM gave some valuable suggestions and participated in this study as a consultant in the process of manuscript revision. All authors contributed to the article and approved the submitted version.

Funding

This study was supported in parts by the National Key Research and Development Program of China under Grant 2022YFF1202900, the National Natural Science Foundation of China under Grant 82102174, and the China Postdoctoral Science Foundation under Grant 2021TQ0243.

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

Abtahi, M., Bahram Borgheai, S., Jafari, R., Constant, N., Diouf, R., Shahriari, Y., et al. (2020). Merging fNIRS-EEG brain monitoring and body motion capture to distinguish Parkinsons disease. *IEEE Transac. Neural Syst. Rehabil. Eng.* 28, 1246–1253. doi: 10.1109/TNSRE.2020.2987888

Al-Shargie, F., Tang, T. B., and Kiguchi, M. (2017). Assessment of mental stress effects on prefrontal cortical activities using canonical correlation analysis: an fNIRS-EEG study. *Biomed. Opt. Expr.* 8, 2583. doi: 10.1364/BOE.8.002583

Ang, K. K., Guan, C., Chua, K. S. G., Ang, B. T., Kuah, C., Wang, C., et al. (2010). "Clinical study of neurorehabilitation in stroke using EEG-based motor imagery brain-computer interface with robotic feedback," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10*, 5549–5552.

Asahi, S., Okamoto, Y., Okada, G., Yamawaki, S., and Yokota, N. (2004). Negative correlation between right prefrontal activity during response inhibition and impulsiveness: A fMRI study. *Eur. Arch. Psychiatry Clin. Neurosci.* 254, 245–251. doi: 10.1007/s00406-004-0488-z

Aygün, M., Sahin, Y. H., and Ünal, G. (2018). Multi Modal Convolutional Neural Networks for Brain Tumor Segmentation *arXiv preprint* 1–8. arXiv:1809.06191.

Baltrusaitis, T., Ahuja, C., and Morency, L. P. (2019). Multimodal machine learning: a survey and taxonomy. *IEEE Transac. Pattern Anal. Mach. Intell.* 41, 423–443. doi: 10.1109/TPAMI.2018.2798607

Buccino, A. P., Keles, H. O., and Omurtag, A. (2016). Hybrid EEG-fNIRS asynchronous brain-computer interface for multiple motor tasks. *PLoS ONE*, 11, 1–16. doi: 10.1371/journal.pone.0146610

Cramer, S. C., Orr, E. L. R., Cohen, M. J., and Lacourse, M. G. (2007). Effects of motor imagery training after chronic, complete spinal cord injury. *Exp. Brain Res.* 177, 233–242. doi: 10.1007/s00221-006-0662-9

Dagdevir, E., and Tokmakci, M. (2021). Optimization of preprocessing stage in EEG based BCI systems in terms of accuracy and timing cost. *Biomed. Signal Process. Control* 67, 102548. doi: 10.1016/j.bspc.2021.102548

Dong, X., Yu, Z., Cao, W., Shi, Y., and Ma, Q. (2020). A survey on ensemble learning. *Front. Comput. Sci.* 14, 241–258. doi: 10.1007/s11704-019-8208-z

Fazli, S., Mehnert, J., Steinbrink, J., Curio, G., Villringer, A., Müller, K. R., et al. (2012). Enhanced performance by a hybrid NIRS-EEG brain computer interface. *NeuroImage*, 59, 519–529. doi: 10.1016/j.neuroimage.2011.07.084

Hallez, H., Vanrumste, B., Grech, R., Muscat, J., De Clercq, W., Vergult, A., et al. (2007). Review on solving the forward problem in EEG source analysis. *J. NeuroEng. Rehabil.* 4, 46. doi: 10.1186/1743-0003-4-46

Herath, K., and Mel, W. (2021). Controlling an anatomical robot hand using the brain-computer interface based on motor imagery[J]. *Adv. Hum. Comput. Interact* 2021, 1–15. doi: 10.1155/2021/5515759

Hétu, S., Grégoire, M., Saimpont, A., Coll, M. P., Eugène, F., Michon, P. E., et al. (2013). The neural network of motor imagery: An ALE meta-analysis. *Neurosci. Biobehav. Rev.* 37, 930–949. doi: 10.1016/j.neubiorev.2013.03.017

Hong, K. S., Naseer, N., and Kim, Y. H. (2015). Classification of prefrontal and motor cortex signals for three-class fNIRS-BCI. *Neurosci. Lett.* 587, 87–92. doi: 10.1016/j.neulet.2014.12.029

Jeannerod, M. (1995). Mental imagery in the motor context. Special Issue: the neuropsychology of mental imagery. *Neuropsychologia*, 33, 1419–1432. doi: 10.1016/0028-3932(95)00073-C

Jeon, Y., Nam, C. S., Kim, Y. J., and Whang, M. C. (2011). Event-related (De)synchronization (ERD/ERS) during motor imagery tasks: implications for brain-computer interfaces. *Int. J. Indus. Ergon.* 41, 428–436. doi: 10.1016/j.ergon.2011.03.005

Kaiser, V., Kreilinger, A., Müller-Putz, G. R., and Neuper, C. (2011). First steps toward a motor imagery based stroke BCI: New strategy to set up a classifier. *Front. Neurosci*. 5, 86. doi: 10.3389/fnins.2011.00086

Kasemsumran, P., and Boonchieng, E. (2019). EEG-based motor imagery classification using novel adaptive threshold feature extraction and string grammar fuzzy k-nearest neighbor classification. *J. Comput*. 30, 27–40. doi: 10.3966/199115992019043002003

Kee, C. Y., Ponnambalam, S. G., and Loo, C. K. (2017). Binary and multi-class motor imagery using Renyi entropy for feature extraction. *Neural Comput. Appl.* 28, 2051–2062. doi: 10.1007/s00521-016-2178-y

Khan, M. J., and Hong, K. S. (2017). Hybrid EEG-FNIRS-based eight-command decoding for BCI: Application to quadcopter control. *Front. Neurorobot.* 11, 6. doi: 10.3389/fnbot.2017.00006

Kwak, Y., Song, W-. J., and Kim, S-. E. (2022). FGANet: fNIRS-Guided Attention Network for Hybrid EEG-fNIRS Brain-Computer Interfaces. *IEEE Transac. Neural Syst. Rehabil. Eng.* 30, 329–339. doi: 10.1109/TNSRE.2022.3149899

Lan, H., Jiang, D., Yang, C., and Gao, F. (2019). Y-Net: a hybrid deep learning reconstruction framework for photoacoustic imaging in vivo. *ArXiv Preprint ArXiv:*1908, 00975.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., Lance, B. J., et al. (2018). EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *J. Neural Eng.* 15, 1–30. doi: 10.1088/1741-2552/aace8c

Liu, Z., Shore, J., Wang, M., Yuan, F., Buss, A., Zhao, X., et al. (2021). A systematic review on hybrid EEG/fNIRS in brain-computer interface. *Biomed. Signal Process. Control* 68, 102595. doi: 10.1016/j.bspc.2021.102595

Lulé, D., Diekmann, V., Kassubek, J., Kurt, A., Birbaumer, N., Ludolph, A. C., et al. (2007). Cortical plasticity in amyotrophic lateral sclerosis: Motor imagery and function. *Neurorehabil. Neural Repair* 21, 518–526. doi: 10.1177/1545968307300698

Mogadala, A., Kalimuthu, M., and Klakow, D. (2021). Trends in integration of vision and language research: a survey of tasks, datasets, and methods. *J. Artif. Intell. Res.* 71, 1183–1317. doi: 10.1613/jair.1.11688

Naseer, N., and Hong, K. S. (2015). fNIRS-based brain-computer interfaces: A review. *Front. Hum. Neurosci.* 9, 1–15. doi: 10.3389/fnhum.2015.00003

Nicolas-Alonso, L. F., and Gomez-Gil, J. (2012). Brain computer interfaces, a review. *Sensors* 12, 1211–1279. doi: 10.3390/s120201211

Pfurtscheller, G. (2010). The hybrid BCI. *Front. Neurosci.* 4, 30. doi: 10.3389/fnpro.2010.00003

Prechelt, L. (2012). Early stopping—But when? *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7700 LECTU, 53–67. doi: 10.1007/978-3-642-35289-8_5

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. *arXiv [Preprint]*. Available online at: http://arxiv.org/abs/2103.00020

Ramachandram, D., and Taylor, G. W. (2017). Deep multimodal learning. *IEEE SIgnal Process. Magaz.* 34, 96–108. doi: 10.1109/MSP.2017.2738401

Sadiq, M. T., Yu, X., Yuan, Z., Zeming, F., Rehman, A. U., Ullah, I., et al. (2019). Motor imagery EEG signals decoding by multivariate empirical wavelet transform-based framework for robust brain-computer interfaces. *IEEE Access*, 7, 171431–171451. doi: 10.1109/ACCESS.2019.2956018

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* 38, 5391–5420. doi: 10.1002/hbm.23730

Selim, S., Tantawi, M. M., Shedeed, H. A., and Badr, A. (2018). A CSP/AM-BA-SVM Approach for Motor Imagery BCI System. *IEEE Access*, 6, 49192–49208. doi: 10.1109/ACCESS.2018.2868178

Shin, J., Von Luhmann, A., Blankertz, B., Kim, D. W., Jeong, J., Hwang, H. J., et al. (2017). Open Access Dataset for EEG+NIRS Single-Trial Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25, 1735–1745. doi: 10.1109/TNSRE.2016.2628057

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1). doi: 10.1186/s40537-019-0197-0

Sun, Z., Huang, Z., Duan, F., and Liu, Y. (2020). A Novel Multimodal Approach for Hybrid Brain-Computer Interface. *IEEE Access*, 8, 89909–89918. doi: 10.1109/ACCESS.2020.2994226

Wozniak, M., Graña, M., and Corchado, E. (2014). A survey of multiple classifier systems as hybrid systems. *Information Fusion*, 16, 3–17. doi: 10.1016/j.inffus.2013.04.006

Yan, X., Hu, S., Mao, Y., Ye, Y., and Yu, H. (2021). Deep multi-view learning methods: A review. *Neurocomputing*, 448, 106–129. doi: 10.1016/j.neucom.2021.03.090

Zhang, X., Yao, L., Sheng, Q. Z., Kanhere, S. S., Gu, T., Zhang, D., et al. (2018). Converting Your Thoughts to Texts: Enabling Brain Typing via Deep Feature Learning of EEG Signals. *2018 IEEE International Conference on Pervasive Computing and Communications, PerCom* 2018. doi: 10.1109/PERCOM.2018.8444575

Zhang, X., Yao, L., Wang, X., Monaghan, J., Mcalpine, D., Zhang, Y., et al. (2021). A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers. *Journal of Neural Engineering*, 18(3). doi: 10.1088/1741-2552/abc902

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms (1st ed.)*. Chapman and Hall/CRC. doi: 10.1201/b12207

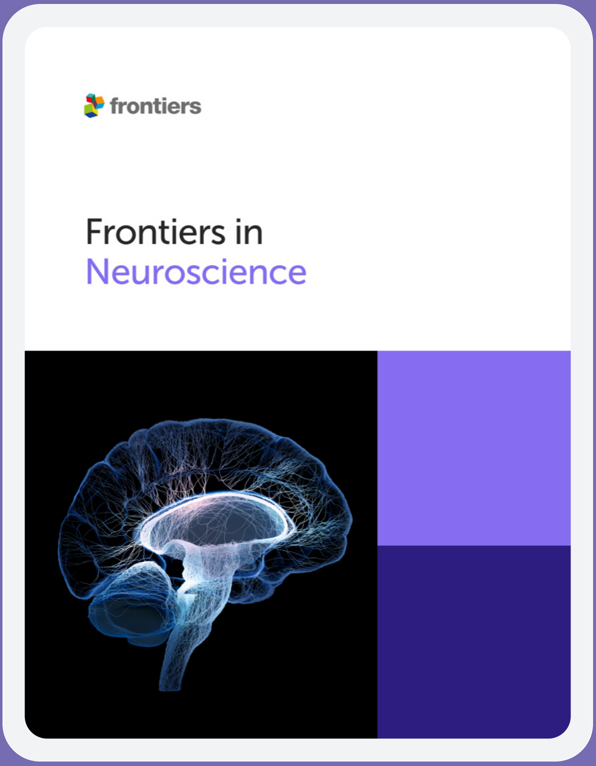
Frontiers in Neuroscience

Provides a holistic understanding of brain function from genes to behavior

Part of the most cited neuroscience journal series which explores the brain - from the new eras of causation and anatomical neurosciences to neuroeconomics and neuroenergetics.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Neuroscience

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact