Received 20 April 2024, accepted 11 May 2024, date of publication 20 May 2024, date of current version 28 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3402965

RESEARCH ARTICLE

Multimodal Medical Image Fusion Network Based on Target Information Enhancement

YUTING ZHOU[1], XUEMEI YANG[1], SHIQI LIU[1], AND JUNPING YIN[2]

[1]China Academy of Engineering Physics, Beijing, Sichuan 100193, China
[2]Institute of Applied Physics and Computational Mathematics, Beijing 100094, China

Corresponding author: Junping Yin (yin_junping@iapcm.ac.cn)

This work was supported in part by Beijing Natural Science Foundation under Grant Z210003, in part by the National Natural Science Foundation of China under Grant NSFC12026607 and Grant NSFC12031016, and in part by the Key Research and Development Program of the Scientific Research Department under Grant 2020YFA0712203 and Grant 2020YFA0712201.

ABSTRACT Glioma is a kind of brain disease with high incidence, high recurrence rate, high mortality, and low cure rate. To obtain accurate diagnosis results of brain glioma, doctors need to manually compare the imaging results of different modalities many times, which will increase the diagnosis time and reduce the diagnostic efficiency. Image fusion technology has been widely used in recent years to obtain information on multimodal medical images. This paper proposes a novel image fusion framework, target information enhanced image fusion network (TIEF), using cross-modal learning and information enhancement techniques. The framework consists of a multi-sequence feature extraction block, a feature selection block, and a fusion block. The multi-sequence feature extraction block consists of multiple sobel dense conv leaky ReLu block (SDCL-block). SDCL-block mainly realizes the extraction of edge features, shallow features, and deep features. The feature selection block identifies the feature channels with rich texture information and strong discrimination ability through the effective combination of global information entropy criterion and feature jump connection. The feature fusion block mainly comprises multi-head and spatial attention mechanisms, which can realize the fusion of intra-modality and inter-modality features. On this basis, considering the influence of tumor spatial location and structure information on the fusion results, a loss function is designed, which is a weighted combination of texture loss, structure loss, and saliency loss so that texture information from multimodal magnetic resonance imaging (MMRI) and saliency information from different anatomical structures of the brain can be fused at the same time to improve the expression ability of features. In this paper, the TIEF algorithm is trained and validated on the MMRI and (Single-Photon Emission Computed Tomography-MRI) SPECT-MRI datasets of glioma and generalized on the (Computed Tomography-MRI) CT-MRI dataset of meningioma to verify the performance of the TIEF algorithm. In the image fusion task, quantitative results showed that TIEF exhibited optimal or suboptimal performance in information entropy, spatial frequency, and average gradient metrics. Qualitative results indicate that the fused images can highlight tumor and edematous features. A downstream image segmentation task was used for evaluation to further verify TIEF's effectiveness. TIEF achieved the best results in both (Dice similarity coefficient) Dice and (Hausdorff distance 95%) HD95 segmentation metrics. In the generalization task, quantitative results indicated that TIEF obtained more information in the meningioma dataset. In conclusion, TIEF can effectively achieve cross-domain information acquisition and fusion and has robustness and generalization ability.

INDEX TERMS Medical image fusion, multimodal magnetic resonance imaging, transformer, feature selection.

I. INTRODUCTION

Brain tumors rank among the most common diseases globally. From 2019 to 2020, China recorded an average of 12,768 brain tumor patients annually on the (National

The associate editor coordinating the review of this manuscript and approving it for publication was Zhen Ren.

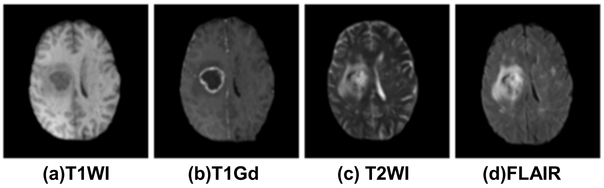


FIGURE 1. MMRI of brain tumors in the BraTs2019 dataset.

Brain Tumor Registry Research Platform) NBTRC platform, a figure nearly 10 times higher than the cases reported in the past decade [1]. Cancer arises from the mutation or change of cellular function [2], resulting in an inability of cells to undergo programmed death [3]. These tumors affect various organs and tissues [4], [5]. While brain tumors rarely spread to other parts of the body, they remain perilous. The growth of tumors can lead to the proliferation and harm of brain tissue in neighboring areas. Even benign tumors can exert significant pressure on brain tissue, causing high-impact complications [6], [7]. Brain tumors account for about 2.17% of all cancer-related deaths, and their 5-year survival rate is only 5.6% [8]. In diagnosing brain tumors, clinicians need to combine different sequences of multimodal MRI, or CT and MRI, to determine the condition of brain tumors and further determine whether the brain tumors are benign or malignant and what kind of treatment plan to use.

MMRI represents different sequences of MRI. Different MRI sequences offer distinct different details about brain tissues and anatomical features. Clinicians will combine multiple modes of MMRI to comprehensively judge the situation of tumors. In recent years, MMRI has emerged as an indispensable tool for precise and personalized medical care [9]. Unlike techniques involving ionizing radiation, MRI remains unaffected by sampling errors and internal variations. Fig. 1 shows a panel of brain tumor images obtained through various MRI sequences. Observation of Fig. 1 reveals unique characteristics across different sequences. T1 weighted imaging sequence (T1WI) presents clear anatomical structures but fails to distinctly depict lesions. In contrast, T1 weighted enhancement scanning imaging sequence (T1Gd) highlights areas with active blood flow, a crucial criterion for accentuating tumors. T2 weighted imaging sequence (T2WI) displays relatively straightforward images aiding in overall tumor assessment. Fluid attenuated inversion recovery image sequence (FLAIR) suppressing high signals in cerebrospinal fluid, delineates peritumoral edema areas. These diverse imaging patterns capture additional pathological information. Given the limitations of individual imaging modes, image fusion aims to merge multi-modal images into a unified output, amalgamating complementary information to facilitate enhanced human visual perception and automated tumor detection. Multimodal brain MRI image fusion contributes to more precise insights into lesion shapes, organizational structures, and relative space position [10], facilitating the design of more accurate individualized treatment plans.

Image fusion methods can be divided into two branches: traditional fusion framework and deep learning fusion framework. Conventional methods include multi-scale transform [11], sparse representation [12], spatial domain method [13], and hybrid method [14]. Traditional image fusion methods are limited by factors such as the complexity of source images, the complex design of artificial fusion rules, and the prolonged processing time [15]. Medical image fusion algorithms based on deep learning are divided into convolutional neural networks (CNN), generative adversarial networks (GAN) and Transformer. On the contrary, deep learning methods CNN [16] and GAN [17] have the advantages of detailed edge texture information, reduced computational cost, and elimination of explicit fusion rule design. However, due to the localized nature of convolution operations [18], they need help to capture comprehensive global knowledge. In addition, the Transformer model has been successful in various vision tasks [19] and applied to medical imaging. Still, the Transformer model mainly focuses on the global information within the domain and ignores the crucial cross-domain integration in the image fusion task. This approach faces challenges distinguishing between target volumes, such as enhanced tumors and background.

Although the results of existing multimodal image fusion algorithms are better than those of traditional image fusion methods, many things could still be improved. Due to the lack of ground truth for medical image fusion, most methods based on deep learning achieve image fusion by designing loss functions. At present, most of the design of the loss function is limited to the global pixel information level, which is not enough to form a fused image better than the source image and limits the quality of the fused image, thereby limiting the applicability of image fusion in medical applications. Although the fusion results of some methods also contain rich texture details, they have no significant contrast and cannot clearly distinguish the target from the background. The features containing rich texture information and edge information only exist in specific feature channels. Using all the shallow features for fusion reduces the fusion effect of the model.

To solve the above problems, this paper proposes a brain tumor MMRI fusion network based on feature selection and attention mechanism: TIEF. This network's fusion image of a brain tumor contains clear brain structure and anatomical information. More importantly, it integrates the description of the edema part, enhanced tumor, and necrotic tumor core of multiple modalities, and the discrimination degree of each area is evident, which can provide doctors with more precise and more accurate tumor information. This work contributes significantly to multiple aspects.

1) We proposed a feature information measurement block based on information entropy, a simple yet robust tool that measures feature information effectively. This block establishes efficient skip connections between encoding and

decoding stages, filtering high-detail and texture-rich feature channels to enhance feature reuse.

2) We designed a fusion block devised to extract and merge multi-modality deep features. Comprising a cross-modality-based token learner block, transformer block, token fusion block, and spatial attention block, this block dynamically identifies critical areas within the input multi-modality deep features, enabling spatial and cross-modal fusion.

3) We proposed a new loss function that incorporates modality and tissue weighting, utilizing the regional contrast index. This function controls the preservation degree of information from source images and focuses on regions of interest vital in various medical applications.

The remainder of this paper is organized as follows. Section II provides a brief review of existing methods in multimodal medical image fusion. Section III introduces an efficient method tailored to the task of multimodal MRI brain tumor fusion. Section IV delves into experimental Settings, Outlines implementation details, presents fusion experimental results, performs ablation studies and generalization studies, compares efficiencies and parameters, and discusses limitations and potential future directions. Section V draws conclusions based on the findings in this paper.

II. RELATED WORK

In this section, the focus is on reviewing pertinent research in image fusion and vision transformer techniques. These two techniques hold considerable relevance to the method adopted in this study, and we aim to provide an overview of their significant developments.

A. TRADITIONAL MEDICAL IMAGE FUSION METHOD

Traditional methods for medical image fusion can be categorized into spatial domain techniques, frequency domain-based fusion, and sparse representation approaches. Jiang et al. [20] introduced and applied the linked independent component analysis method in a multi-modal MRI study of Alzheimer's patients. The study revealed increased mean diffusivity, decreased gray matter volume, alterations in anisotropy fraction and diffusion tensor patterns in the corpus callosum and forceps, and increased anisotropy fraction and diffusion tensor pattern in the regions of the superior longitudinal fasciculus passing through the descending fibers, such as the internal capsule, corona radiata, and superior longitudinal fasciculus. Wang et al. [21] proposed a joint Laplacian pyramid method integrating multiple features to effectively transfer salient features from source images to a single fused image, improving indicators such as standard deviation (STD) by 10-15% compared to other traditional methods. Kang et al. [22] presented a novel approach utilizing group sparsity and graph positivity regularization in dictionary learning (DL-GSGR) for medical image denoising and fusion. This method demonstrated more effective feature extraction compared to standard sparse representation and multi-resolution analysis, enhancing indicators like mutual information (MI) and universal quality index (UIQI) by 5-15%. Additionally, Guo et al. [23] proposed a multi-modal image fusion framework based on two-scale image decomposition and sparse representation, overcoming the limitations of single traditional methods. This approach retained finer details and edge features, showing an average improvement of 30% in metrics like edge intensity (EN) compared to optimal strategies.

Traditional multimodal medical image fusion methods combine the target task to set the fusion rules and improve the image clarity by processing the complementary information between multiple images. However, although these traditional algorithms are relatively simple, they are only applicable to specific tasks or specific datasets, have limited generalization ability, and they require more demanding feature extraction and processing, leading to slower computation speeds. The image fusion algorithm based on deep learning provides a promising solution to solve the limitations of traditional methods by enhancing the image fusion effect.

B. DEEP LEARNING IMAGE FUSION METHOD

Deep learning-based image fusion methods encompass various techniques such as CNN, GAN, and Transformer. CNN excel at processing spatial and structural information within adjacent regions of input medical images. Typically composed of convolutional, pooling, and fully connected layers, CNN extract features from source images, mapping them to final outputs. These networks define image fusion as a classification problem, utilizing CNN-based algorithms to transform images, measure activity levels, and devise fusion rules. Medical image fusion based on CNN mainly includes pixel-level fusion and feature-level fusion. Pixel-level fusion is simply a weighted average of pixel values. The fusion of feature levels mostly involves joining or adding the channels of the feature map. For instance, Vaswani et al. [24] designed an early CNN-based fusion method, integrating traditional activity level measurements with CNN-based feature extraction to produce fused images via pixel-weighted averages or selected fusion strategies. Similarly, Li et al. [25] designed a multi-scale CNN framework, training the network to generate decision graphs for image fusion. Despite CNN's ability to learn from limited medical image datasets, the challenge of overfitting persists due to the scarcity of medical image samples. CNN learn hierarchical features, enhancing image content comprehension and analysis. However, because CNN only focuses on local information, the complexity and diversity of multimodal medical image fusion limits its ability to achieve optimal results.

The GAN algorithm differs significantly from CNN, employing a generator and discriminator for feature extraction and optimization. In GAN, the generator produces an image, while the discriminator discerns between real and generated images. GAN is trained using an adversarial loss function, where the generator and discriminator are

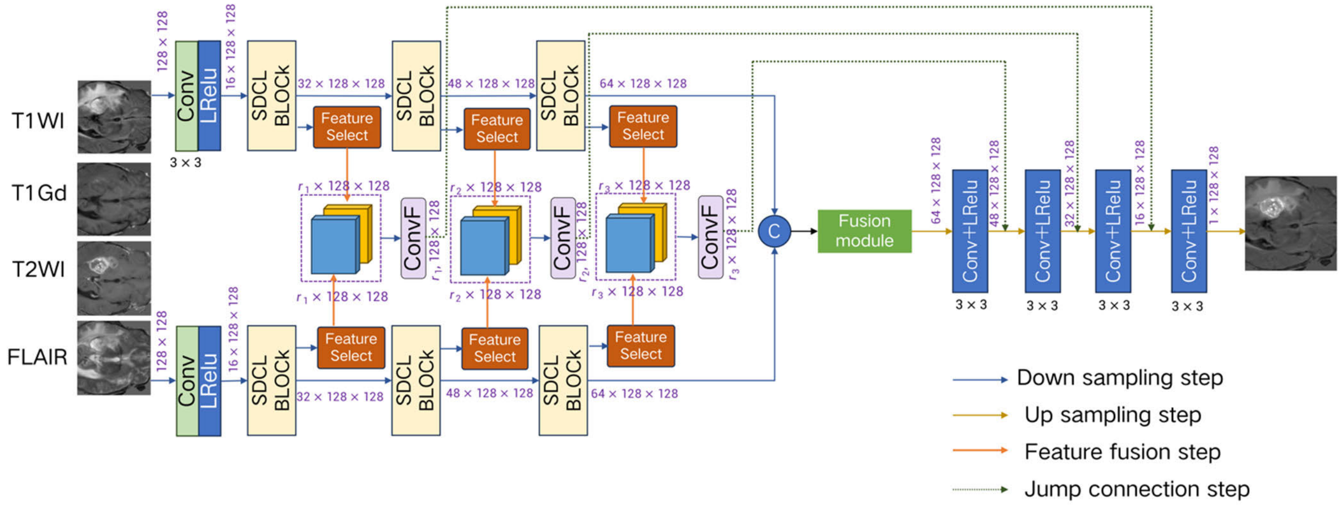


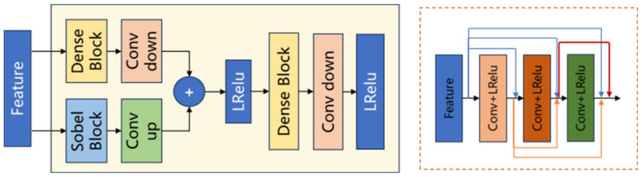
FIGURE 2. The overall architecture of the proposed TIEF method.

in a constant adversarial interplay, striving for equilibrium. The medical image fusion of GAN mainly generates the fused image through the generator and judges whether the generated image is realistic by discrimination. The confrontation between the generator and the discriminator optimizes the fusion effect. Liu et al. [26] proposed the fusion GAN algorithm, treating image fusion as an image generation task and utilizing the least squares GAN objective to stabilize training. While GAN-based fusion methods address some issues through adversarial confrontation, the simplicity of single-scale networks in generators may lead to information loss and excessive smoothing, causing distortions in the fused image. To address this challenge, Liang et al. [27] introduced the fusion network, incorporating lightweight transformer blocks and adversarial learning to emphasize global fusion. This model enables interaction between shallow CNN-extracted features and the transformer fusion block, refining spatial and cross-channel fusion relationships. GAN models excel in retaining the selected information from source images without requiring labeled data, delivering clear and minimally distorted images. However, due to the complexity of the GAN model, the gradient is prone to disappear. Although the generator and discriminator of GAN can realize cross-modal learning, they cannot adaptively learn complementary information and screen important information and channels.

Recently, Transformer-based algorithms has received a lot of attention in the image fusion community. There are many medical image fusion frames based on Transformer that have achieved impressive performance. To extract local and global information, Du et al. [28] used Patch Pyramid Transformer (PPT) to extract non-local information from the entire image [29], based on the AE-based fusion framework. In addition, Maqsood et al. designed spatio-Transformer as a multi-scale fusion strategy to capture both local and global contexts [30], based on CNN-based fusion framework. For better fusion result, Du et al. introduced parallel Transformer and CNN architecture into the AE-based fusion framework, (i.e., TransMEF [31]). Furthermore, Wang et al. also injected Transformer into GAN-based fusion framework to learn the global fusion relations [32]. To reduce computational costs, Guo et al. [33] proposed a hierarchical Transformer (i.e., Swin Transformer) by adopting shifted windows to compute the representation. In their method, Swin Transformer allowed cross-window connection and limited self-attention computation to non-overlapping local windows, which achieved greater efficiency and flexibility. Motivated by [34], residual Swin Transformer (RSTB) has been proposed to extract deep feature for image restoration [35]. The image fusion method based on the Transformer ignores the cross-domain information, fails to capture the local and correlation information between different modalities, and fails to highlight the lesion in the tumor tissue area emphasized by enhanced tumor and other multimodal images. This is the key to the problem of MMRI brain tumor image fusion.

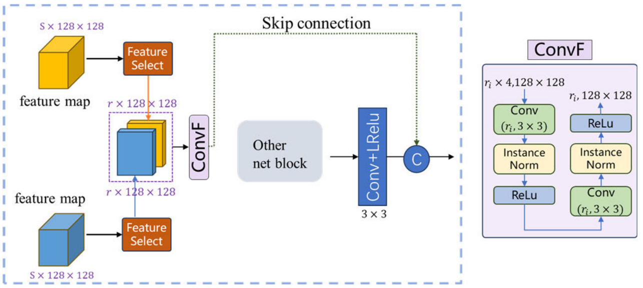
III. METHODS

In this section, we designed a TIEF network tailored for mining and fusing multi-modal MRI images, illustrated in Fig. 2. The architecture primarily comprises the SDCL-block, feature selection block, and fusion block. The SDCL-block operates as a detail-enhanced dual branch for deep feature extraction, while the feature selection block serves as a block for channel selection, focusing on information-rich channels. The fusion block is responsible for integrating intra-modality and inter-modality deep features obtained from the encoder. TIEF adopts a U-shaped framework, featuring four branches in the encoding section for individual extraction of deep features from four source images. Conversely, the decoding section consists of a single branch dedicated to reconstructing the fused image.



(a) SDCL block. (b) Dense block.

FIGURE 3. Illustration of the proposed SDCL block.



(a) Feature selection block. block. (b) ConvF

FIGURE 4. Illustration of the process of feature selection.

We denoted the source multi-modal MRI images as $X \in \mathbb{R}^{I \times H \times W}$, where I represents the number of modalities. For the MMRI fusion task, $I = 4$, cor-responding to T1WI, T1Gd, T2WI, and FLAIR modalities. In the case of SPECT -MRI fusion, $I = 3$, cor-responding to T1WI, T2WI, and SPECT modalities. The symbols of W and H denote the image's width and height.

To enhance feature extraction within the encoding stage, we designed a novel SDCL-block, depicted in Fig. 3, employing a double parallel structure. One branch comprised a dense block, optimizing the utilization of features extracted through various convolutional layers. The other branch employed gradient operations to calculate feature gradient magnitudes, focusing on texture information extraction. The Conv up and Conv down stages incorporated 1 × 1 convolutional layers to standardize channel counts within the double-branch structure features. Subsequently, an additional operation integrated the depth and detail features obtained from the dual branches. The latter part of the SDCL-block further accentuates feature integration by repeating the dense block structure, reinforcing the propagation strength of the extracted features.

A. FEATURE SELECTION

According to the pruning algorithm [36], the importance of neurons, filters, and channels can be measured using specific criteria. The less important branches can be pruned to reduce the model size and speed up the calculation without compromising the accuracy. To optimize feature utilization in decoding, we've devised a novel feature selection block. This block dynamically filters richer-detail features for more effective skip connections.

We opted for information entropy as the criterion of choice. Information theory supports entropy as an evaluation metric to quantify the information content within an image or feature. It effectively reflects the intensity distribution's spatial and aggregative characteristics. A higher entropy value signifies greater information content within an image. While one-dimensional entropy assesses the gray value aggregation, it does not capture spatial information. Contrastingly, two-dimensional entropy encapsulates spatial characteristics. In this study, the two-dimensional entropy enables the characterization of content abundance within each feature channel. For a feature map $F_d \in \mathbb{R}^{H \times W}$ in l^{th} layer, where $d = 1, \ldots, D$(with D being the number of channels). We used a 3 × 3 sliding window to traverse the whole map (stride is 1). The two-dimensional information entropy of F_d is defined as:

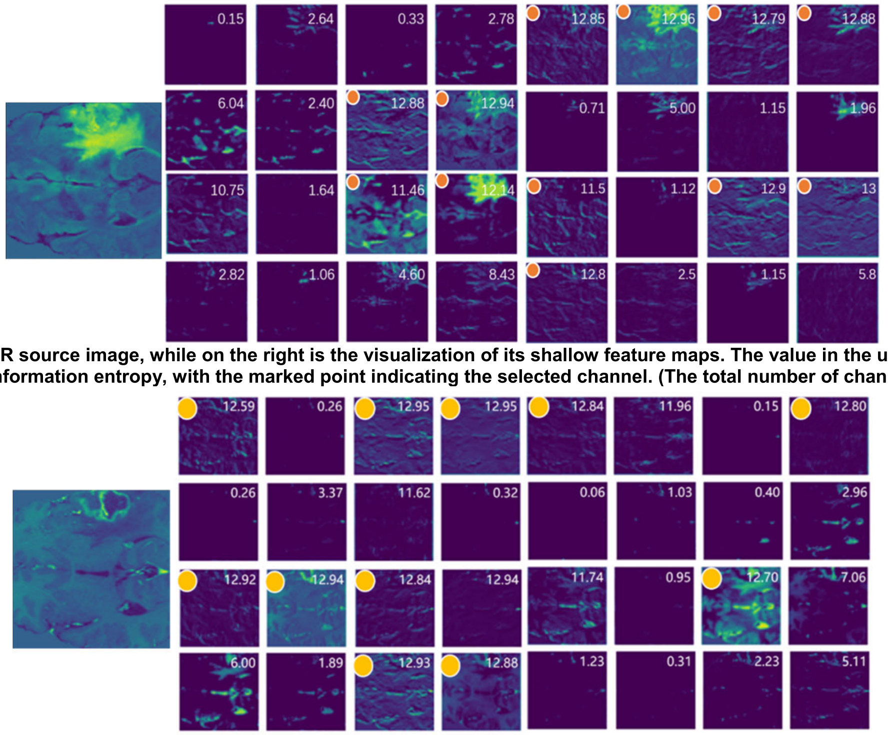
$$H(F_d) = -\sum\nolimits_{i=0}^{255} \sum\nolimits_{j=0}^{255} p_{ij} \log_2 p_{ij} \tag{1}$$

$$p_{ij} = \frac{f(i_n, j_n)}{WH} \tag{2}$$

where i_n denotes the gray value of the center pixel in the nth sliding window, j_n represents the mean gray value of the neighborhood of the rest pixels centered at i_n in the nth sliding window. Thus, we get a set $\{(i_n, j_n)\}_{n=1}^{HW}$ to reflect the comprehensive characteristics of the central pixel and its surrounding pixels. The occurrence probability of $f(i_n, j_n)$ in the image is defined as p_{ij}, where $f(i_n, j_n)$ is the occurrence number of (i_n, j_n), and W and H are the dimensions of the feature map.

Visualizing feature maps across different depths and channels and calculating their information entropy assists in discerning the relationship between feature information richness and feature map entropy. This process is particularly valuable in identifying feature maps that align more closely with human vision.

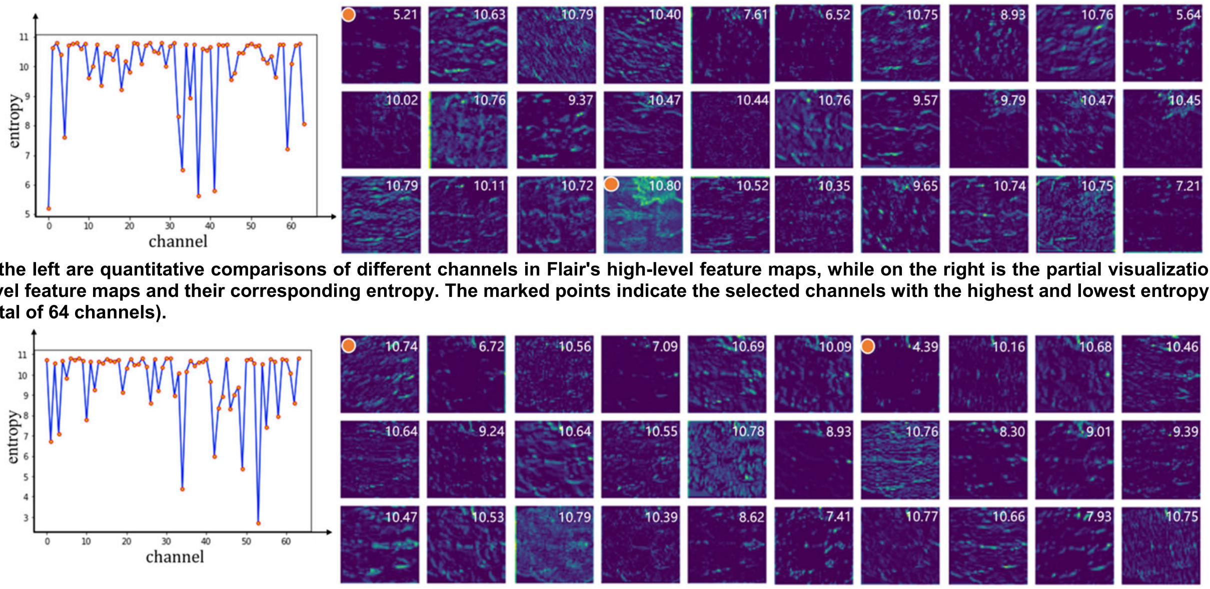
We calculated the entropy of Flair and T1ce feature maps across different layer depths (Fig. 5 and Fig. 6). Both figures demonstrate that each channel extracts distinct information. The value in the upper right corner of the figure is the information calculated by Eq. (1), and the circle in the upper left corner represents the channel whose information entropy is greater than the threshold. Channels within the same layer focus on varied details and different areas. Channels with higher entropy values, compared to those with lower entropy values, exhibit richer texture details and more salient pixels in tumor areas, aiding in visual tumor detection. To optimize fusion image reconstruction in the decoding phase, selecting feature maps with rich information–specifically, channels with high entropy–is essential. We calculated and ranked all channel features' entropy values $\{H(F_d)\}_{d=1}^{D}$ in lth layer, selecting the top r entropy channels. In this paper, the researcher adopted $r = 8$. Merely selecting features with rich spatial information might not accurately and comprehensively represent image content. Hence, a feature selection model is integrated as a network branch, addressing crucial features in decoding through skip connections (Fig. 4).



(a). On the left is the FLAIR source image, while on the right is the visualization of its shallow feature maps. The value in the upper corner of the feature map represents information entropy, with the marked point indicating the selected channel. (The total number of channels is 32).

(b). On the left is the T1ce source image, and on the right is the visualization of its shallow feature maps. The value in the upper corner of the feature map represents information entropy, with the marked point indicating the selected channel. (There are a total of 32 channels.)

FIGURE 5. Visualization and entropy of shallow features.



(a). On the left are quantitative comparisons of different channels in Flair's high-level feature maps, while on the right is the partial visualization of its high-level feature maps and their corresponding entropy. The marked points indicate the selected channels with the highest and lowest entropy (There are a total of 64 channels).

(b). On the left are quantitative comparisons of different channels in T1ce high-level feature maps, while on the right is the partial visualization of its high-level feature maps alongside their corresponding entropy. The marked points indicate the selected channels with the highest and lowest entropy. (There are a total of 64 channels.)

FIGURE 6. Visualization and entropy of deeper layer features.

B. FUSION BLOCK

In this study, we introduce a fusion block aimed at mining and integrating multi-modality deep features. This block comprises a cross-modality-based token learner block, transformer block, token fusion block, and spatial attention block. These components adaptively tokenize crucial regions within

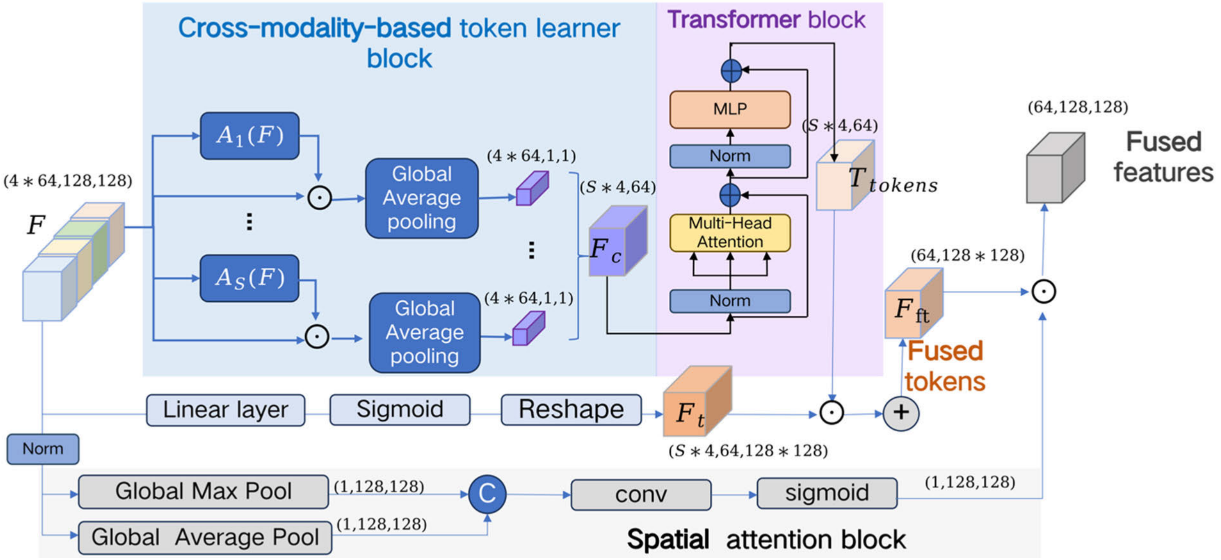


FIGURE 7. The architecture of Fusion Block: Cross-Modality-Based Token Learner Block, Transformer Block, Fused Tokens Block, and Spatial Attention Block.

the input multi-modality deep features, facilitating spatial and modality-based fusion. The network architecture of this proposed fusion block is depicted in Fig. 7. We represented the deep feature maps from each modality as $\{F_1, \ldots, F_I\}$, $F_i \in \mathbb{R}^{C\times H\times W}(i = 1, 2, \ldots, I)$, where C denotes the number of channels. Upon concatenating $\{F_1, \ldots, F_I\}$, we derived a tensor $F \in \mathbb{R}^{IC\times H\times W}$.

1) CROSS-MODALITY-BASED TOKEN LEARNER BLOCK

We learned to generate a series of tokenizer function $\{A_i\}_{i=1}^{S}$, where S represents the number of tokens, aiming to map the modality feature F to a token vector V_i.

$$V_i = Global\ average\ pool(F \odot A_i(F)) \tag{3}$$

where $V_i \in \mathbb{R}^{IC\times 1\times 1}$ and $\odot$ is the Hadamard product (i.e., element-wise multiplication). This approach enabled the tokens to dynamically adapt their spatial selections rather than being fixed splits of the input tensor. These varying tokens effectively mine intra-modality and inter-modality deep features, enabling the modeling of their relationships and interactions. The resulting tokens are aggregated to form the learned token tensor $V \in \mathbb{R}^{SI\times C}$. In this paper, we adopted $S = 8$. Subsequently, the learned token tensor is forwarded to the subsequent transformer block.

2) FUSED TOKENS BLOCK

Following the token generation by the cross-modality-based token learner block and subsequent processing by the transformer block, the fused tokens block is employed to further amalgamate information among the tokens. This functionality facilitates the model in capturing cross-modality 'patterns' formulated by these tokens. The synergy between the cross-modality-based token learner block and the fused tokens block aims to fuse intra- and inter-modality deep features effectively, ensuring robust integration of complementary information.

We started by applying a simple linear layer (denoted as f_{linear}, where $f_{linear} \in \mathbb{R}^{HW\times SI}$) independently across each channel of F, incorporating a sigmoid activation function and reshaping operation. This operation results in $F_t \in \mathbb{R}^{SI\times C\times HW}$. Subsequently, the token tensor, denoted as $T_{tokens} \in \mathbb{R}^{SI\times C}$, is generated by the transformer block. We then executed $F_s = F_t \odot T_{tokens}(resulting\ in\ F_s \in \mathbb{R}^{SI\times C\times HW})$ and executed token-wise addition on F_s along the token axis. Consequently, we obtained the modality-enhanced feature embedding $F_{fused\ tokens} \in \mathbb{R}^{C\times HW}$.

3) SPATIAL ATTENTION BLOCK

Spatial attention serves to identify crucial regions within an image by assigning significance scores to various spatial regions within the feature map. This mechanism accentuates important areas while dampening features in less relevant regions. The spatial attention block employs global max-pooling and global average-pooling operations along the channel axis, concatenating their outputs to generate an effective feature descriptor. The computation of the spatial attention mechanism unfolds as follows:

$$Matrix_{max} = Global\ \max\ pool\left(Layernorm\left(F^C\right)\right) \tag{4}$$

$$Matrix_{avg} = Global\ average\ pool\left(Layernorm\left(F^C\right)\right) \tag{5}$$

$$Weight_{spatial} = Sigmoid\left(Conv\left(Concat\left(Matrix1, Matrix2\right)\right)\right) \tag{6}$$

$$F_{fused\ features} = Weight_{spatial} \odot F_{fused\ tokens} \tag{7}$$

Subsequently, we acquired the fused feature maps that encompass a selection of pixels, spatial locations, and modalities, ensuring an adaptive and informative amalgamation across modalities and spatial aspects.

C. LOSS FUNCTION

To facilitate the reconstruction of multi-modal image fusion, we established a comprehensive loss function considering three perspectives: texture information, structural information, and salient target information.

1) TEXTURE LOSS

Different source images exhibit different features, such as independent units, signal-to-noise ratio, voxel count, spatial smoothness, and intensity distribution. Images of the same morphology but different regions share overall structural similarity but demonstrate different specific details and textures. The purpose of fused images is to bridge the detail gap caused by modal heterogeneity while preserving the complex texture details. Through feature visualization experiments, an optimal texture loss function is determined in this paper. This function preserves more texture information by fusing different modes of the image. At the same time, to reduce the loss of image details, the loss function introduces the Canny operator to depict the subtle differences in the texture. With these considerations in mind, texture loss was formulated to encourage fused images to contain richer texture information. Mathematically defined as:

$$L_{texture} = \frac{1}{HW}\,\||\nabla G| - \max(|\nabla I_i|)\|_1 \tag{8}$$

where ∇ denotes the canny operator and $\|.\|_1$ denotes the loss of L_1.

2) STRUCTURE LOSS

The Structural Similarity (SSIM) [37] metric is commonly employed to impose structural constraints, ensuring that the fusion results encompass adequate structural details. The SSIM applies a structural similarity index measurement to constrain the resemblance between the fusion image and the source images. Mathematically defined as:

$$SSIM\,(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{9}$$

where μ and σ are the mean and variance operations, σ_{xy} denotes the covariance, C_1 and C_2 are two constants.

Entropy is used to measure the richness of image information. From the information theory perspective, regions with richer texture details have higher information content and entropy. It is worth noting that although both entropy and gradient methods can evaluate texture richness, the entropy method is more advantageous than the gradient method. The imaging gradients showed varying degrees of response in different modes. For example, on T2WI, there is a strong gradient in the tumor region at the edge of the tumor, while the gradient in the other areas is sparse and more minor. T1WI showed a slight tumor response with a weak gradient. T1Gd showed marked intratumoral enhancement with a relatively uniform distribution of pixels and a mild gradient outside the tumor. The distribution of pixel intensity in FLAIR images was not uniform, and the gradient change was noticeable. However, relying solely on gradients as a measure may lead to misleading results that significantly affect marginal assessments.

The entropy calculation depends on the probability distribution of the individual gray levels within the image. Considering the overall pixel intensity distribution through the probability distribution, it is not easy to be affected by the sparse gradient. The higher the information entropy in the image, the richer the content contained in the image. The contribution of different modal photos to the final image fusion was calculated according to the information entropy of the image, and the corresponding weight was given.

$$w_{ij} = \frac{e^{\kappa H_{ij}}}{\sum_{i,j} e^{kH_{ij}}} \tag{10}$$

where κ is the adjustment coefficient, balancing the ratio between H_{ij}.

In similarity calculation, a mask is considered for the tumor region to ensure that critical information is covered within a small receptive field range. Therefore, this paper designs a mask region similarity loss function:

$$L_{M-SSIM} = 1 - \frac{1}{IJ}\sum\nolimits_{i=1}^{I}\sum\nolimits_{j=1}^{J} SSIM(w_{ij}I_{Mij}, G_j) \tag{11}$$

$$I_{Mij} = I_i \odot Mask_j \tag{12}$$

$$G_j = G \odot Mask_j \tag{13}$$

Here, $i = 1, 2, \ldots\ldots, I; j = 1, 2, \ldots\ldots, J; I = 4; J = 4.$ $\odot$ denotes element-wise multiplication, and $Mask_j$ represents masks of different tumor regions, with $Mask_1$ indicating the normal tissue area.

3) SALIENT LOSS

To better meet the human vision and realize the significant presentation of different tumor parts in the fusion image, this paper increased the contrast between different tissues of the tumor to make the tumor salient. Therefore, we introduce the Salient Loss term:

$$L_{salient} = 1 - \frac{1}{3}\sum\nolimits_{j=2}^{4}\left|\frac{R\left(G_j\right) - R(G_j')}{R(G)}\right| \tag{14}$$

$$R\left(G_j\right) = \frac{sum(G_j)}{sum(Mask_j)} \tag{15}$$

G_j' denotes the rest of the image region of G except G_j.

$$R\left(G_j'\right) = \frac{sum(G_j')}{sum(A - Mask_j)} \tag{16}$$

Here, A is an all-ones matrix of size $H \times W$.

4) TOTAL LOSS

Total loss is calculated by the following formulae:

$$L_{total} = \alpha L_{texture} + \beta L_{M-SSIM} + \eta L_{salient} \tag{17}$$

where α, β, η is the balance coefficient, which is used to control the proportion of the influence of texture information, structure information, and saliency information on the fusion result.

IV. RESULTS AND DISCUSSION

This section presents a comparative analysis of TIEF and several state-of-the-art methods using multimodal medical images. These experiments involve qualitative and quantitative comparisons using publicly available datasets. In addition, we performed ablation and generalization studies to delve into the performance and components of the method.

A. DATASETS

The fusion effect of TIEF was verified by using several different multimodal medical datasets, and the generalization effect of TIEF was verified by using one multimodal medical dataset of various diseases. The BraTs2019 dataset [38] consists of 335 cases, each containing four MRI sequences (FLAIR, T1WI, T1Gd, T2WI) and tag sequences that outline the tumor core, post-enhancement tumor, edema, and the entire tumor region. These labels helped we make masks. A fusion experiment was performed on RGB multimodal medical images from the neoplastic disease (brain tumor) dataset in AANLIB [39] to verify the fusion performance. This dataset included SPECT-T1WI, GAD, and T2WI images. Notably, the BraTs2019 dataset consists of grayscale images, while the SPECT-T1WI images in the AANLIB dataset are in RGB format. Medical image fusion of CT and MRI was performed using meningioma data from the ANNLIB dataset to verify the model's generalization.

B. COMPARSION METHODS AND EVALUATION INDICATORS

We compared the proposed TIEF with a comprehensive set of established image fusion methods used in the field. This comparison included traditional methods such as CBF (2015) [40] and MGFF (2019) [41], alongside contemporary techniques like U2Fusion (2020) [42], EMFusion (2021) [43], and SeAFusion (2022) [44], which are CNN-based fusion approaches. Additionally, we evaluate the proposed technique against GAN-based methodologies such as FusionGAN (2019) [45], DDcGAN (2020) [46], and GANMcC (2020) [47]. Furthermore, the performance of the proposed TIEF was assessed against recent Transformer-based fusion methods, including SwinFusion (2022) [48], MRSCFusion (2023) [49] and DesTrans (2024) [50].

Further, for quantitative comparison, we utilized eight metrics to assess fusion performance across all models presented in this study. These metrics included average gradient (AG) [51], spatial frequency (SF) [52], entropy (EN) [53], mutual information (MI) [54], peak signal-to-noise ratio (PSNR) [55], structural similarity index measure (SSIM) [56], gradient-based fusion performance ($Q_{AB/F}$) [39], and contrast index (CI) [57]. SSIM evaluates structural similarities between source and fused images in terms of correlation, luminance, and contrast distortion. Higher SSIM values indicate lower structural loss and distortion. PSNR represents the ratio of peak value power to noise power in the fused image, where higher PSNR values signify closer proximity to the source images. EN quantifies image information, where greater information entropy signifies richer knowledge in the fused image. AG measures grayscale changes across image boundaries, indicating image sharpness and detail contrast. Higher AG values indicate better fusion performance. SF gauges row and column frequency in the fused image, reflecting image texture and edge detail richness. CI signifies contrast between foreground and background, aiding in differentiating diseased and normal tissue areas visually. Higher CI values improve tumor visibility. MI assesses image intensity similarity between source and fused images, while $Q_{AB/F}$ measures edge information similarity. Greater MI and $Q_{AB/F}$ values denote superior fusion performance.

C. EXPERIMENTAL DETAILS

The epoch count was set at 320, employing an initial learning rate of 0.0005 with exponential decay and using the Adam optimizer. Each batch size is set to 32. BraTs2019 dataset cases, comprising FLAIR, T1WI, T1Gd, and T2WI multimodal MRI images, were aligned to FLAIR modality and resized to 128 × 128 × 32. AANLIB dataset cases, which encompassed T2WI, GAD, and SPECT-T1WI multi-modal MRI images, were aligned to GAD modality and resized to 128 × 128 × 32. For hyperparameters in Eq. (10), κ was set to 0.25. In Eq. (17), α, β, and η were set to 0.3, 0.4, and 0.3. All experiments were conducted using PyTorch on a Windows workstation equipped with an Intel®Core™i9-10900X CPU and an NVIDIA Geforce GTX Titan A100 GPU.

D. RESULT

1) BRATS2019 MULTI-MODAL MRI FUSION

The fusion results on the BraTs2019 dataset are shown in Fig 8, showcasing outcomes from three experiments, each involving four distinct MRI image modalities: FLAIR, T1 WI, T1Gd, and T2WI. We selected three specific images that highlight variations in modalities regarding information richness and imaging quality, notably the image quality of T1WI. Each mode encompasses distinct details about the tumor, resulting in noticeable differences.

The fusion results on the BraTs2019 dataset are shown in Fig. 8, and each experiment involved four different MRI image modalities: FLAIR, T1WI, T1Gd, and T2WI. We selected three images that showed differences in information richness and imaging quality. Combined with Fig. 8, it can be found that the fusion results of CBF, FusionGAN, DDcGAN, GANMcC, and MGFF lose more details of the source map, which reduces the identifiability of structural information and makes the image unclear. In contrast, TIEF preserves the critical information, modal

TABLE 1. Quantitative comparison of different methods for 8 evaluation items indicators in the Brats2019 dataset (Red: Optimal, Blue: Suboptimal).

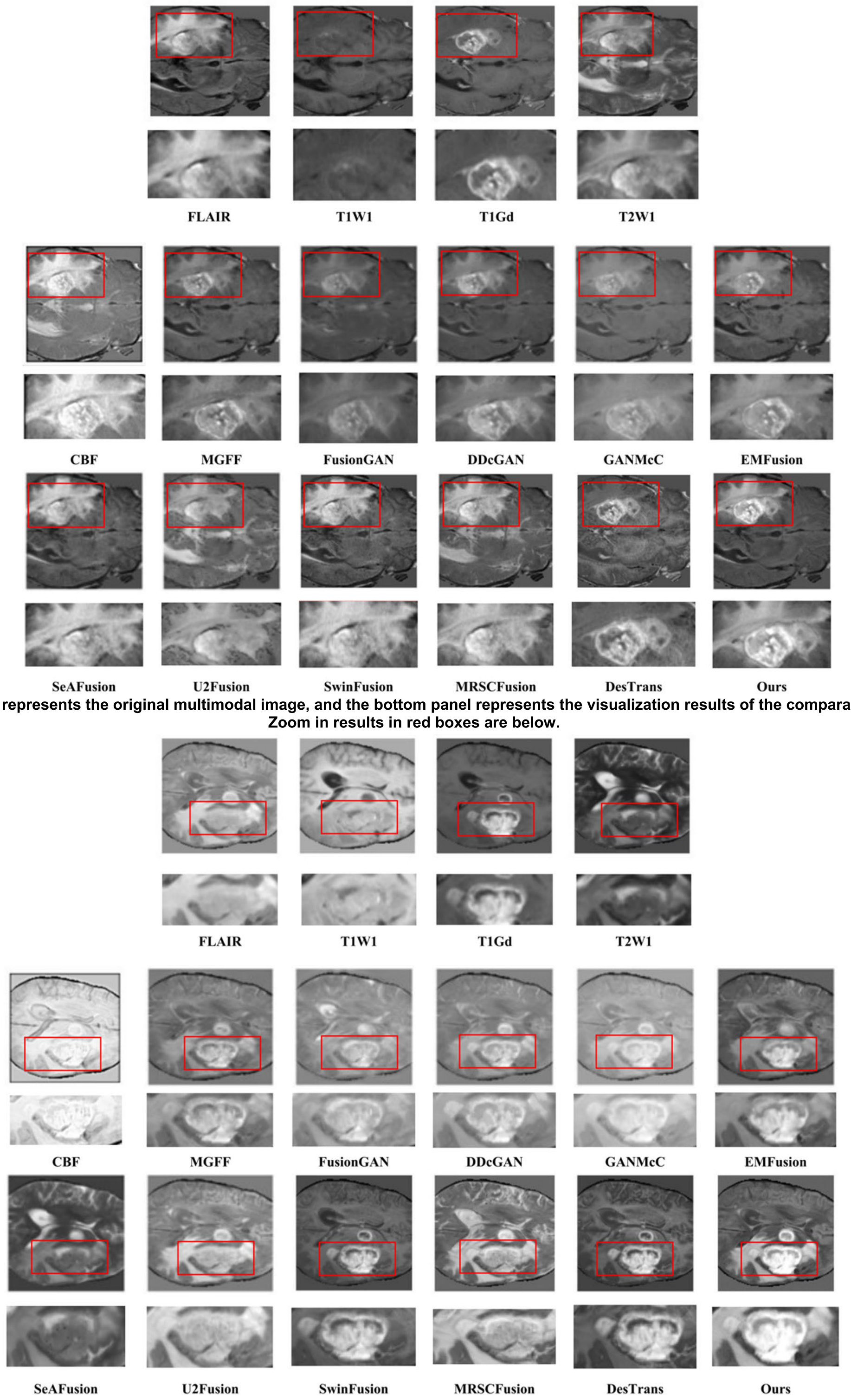
Categories	Methods	$Q_{\frac{AB}{F}}$	EN	AG	SF	CI	MI	PSNR	SSIM
traditional methods	CBF	0.408	11.623	6.802	27.990	1.191	4.289	44.891	0.416
	MGFF	0.358	11.564	5.739	29.885	1.305	4.281	46.478	0.378
CNN	EMFusion	0.407	11.279	9.448	39.375	1.843	5.282	47.871	0.553
	SeAFusion	0.428	12.128	9.481	**39.387**	2.055	5.169	47.441	0.550
	U2Fusion	0.359	7.845	**11.243**	37.700	**2.363**	**7.790**	**49.129**	0.500
GAN	FusionGAN	0.364	11.400	8.210	35.165	0.468	4.143	46.685	0.436
	DDcGAN	0.372	10.753	9.077	36.190	1.745	4.432	48.269	**0.587**
	GANMcC	0.349	12.598	8.424	38.230	1.146	4.882	45.878	0.520
Transformer	SwinFusion	0.417	12.778	9.485	38.136	2.275	5.707	47.417	0.514
	MRSCFusion	**0.441**	**12.815**	10.515	36.564	2.236	4.720	48.021	0.531
	DesTrans	0.405	11.879	9.798	37.712	2.137	5.574	47.474	0.598
Our	TIEF	**0.495**	**13.226**	**11.339**	**40.700**	**2.450**	**5.870**	**48.472**	**0.676**

structural details, texture details for each modality, and pixel intensities in the fusion results, thus enhancing clarity, structure, and texture. This advantage is because the adopted method enhances the extraction and transfer of structural information, ensuring that the extraction and retention of source image knowledge is more comprehensive than other techniques. TIEF shares information in the fusion block and the loss function part. Therefore, when a particular mode image quality is low and the texture is unclear, the fusion result minimizes the interference of the low-quality mode, strengthens the information of other modes, and prevents the loss of edge texture cues. On the contrary, in the third line of the experiment, the fused images of other methods were significantly affected by the T1WI mode, resulting in a decrease in image quality. In terms of tumor details, while EMFusion, SeAFusion, and U2Fusion retain information from a variety of patterns representing different tumor tissues, like SwinFusion, MRSCFusion, and DesTrans, in their fusion results, The boundaries of the tumor core, post-enhancement tumor, and edema were not apparent. In contrast, TIEF delineates the pixels of different tumor tissues in the loss function, which ensures a more obvious distinction between tumor core, enhanced tumor, and edema in the final fusion result.

The qualitative fusion results are shown in Table 1. The comparative analysis with the other 11 methods showed that TIEF had better EN, SF, AG, and SSIM scores, indicating that the fusion image quality was higher, and the multimodal image feature information was better preserved. In addition, TIEF obtained the best $Q_{AB/F}$ and CI scores, indicating reduced distortion, improved visual quality, and good agreement with human visual perception. The PSNR index of TIEF is suboptimal. The reason is that PSNR is the most common and widely used objective image evaluation index, but it is different from human visual characteristics. The human eye has a high sensitivity to luminance contrast differences, and the perception result of a region will be affected by the brightness of its neighboring areas. Just as TIEF expects the brightness of different brain lesion tissues to be significantly different in fused images. Therefore, PSNR is inconsistent with human subjective feeling, and PSNR is suboptimal in several algorithms. The MI metric of TIEF is suboptimal because the information is converted from the source map to the fusion map, and theoretically, the amount of information should remain the same. However, the fusion images obtained by TIEF emphasize the advantages of different modes, and other lesions are clearly distinguished in the fusion images. The fused image enhanced the individual and cooperative information, reducing the interference of redundant information. So, although the overall mutual information decreases from a macro perspective, the synergy of information increases. So, there is a specific reduction in MI.

2) EXTENSION TO SPECT AND MRI FUSION

TIEF was experimentally performed on SPECT and MRI images within the AANLIB dataset to further demonstrate the generality of the proposed method. Seventy pairs of multi-model MRI/SPECT images were used for training, resampled to 128 × 128, of which 5 pairs were used for testing. Since the SPECT images are RGB, the investigators converted them to YUV and extracted the Y-channel for fusion with the TIEF single-channel grayscale MRI images. The output y-component is the basis for the fused image, which is then converted back to the RGB of the final image. T2WI images have rich texture details, and in GAD and SPECT-T1WI images, there is a clear contrast between the pixel values of normal tissue and the diseased area. Therefore, the adopted fusion evaluation criteria prioritize preserving precise texture details, structural information, and pixel contrast within the lesion area. The evaluation results are shown in Table 2 and Fig. 9.

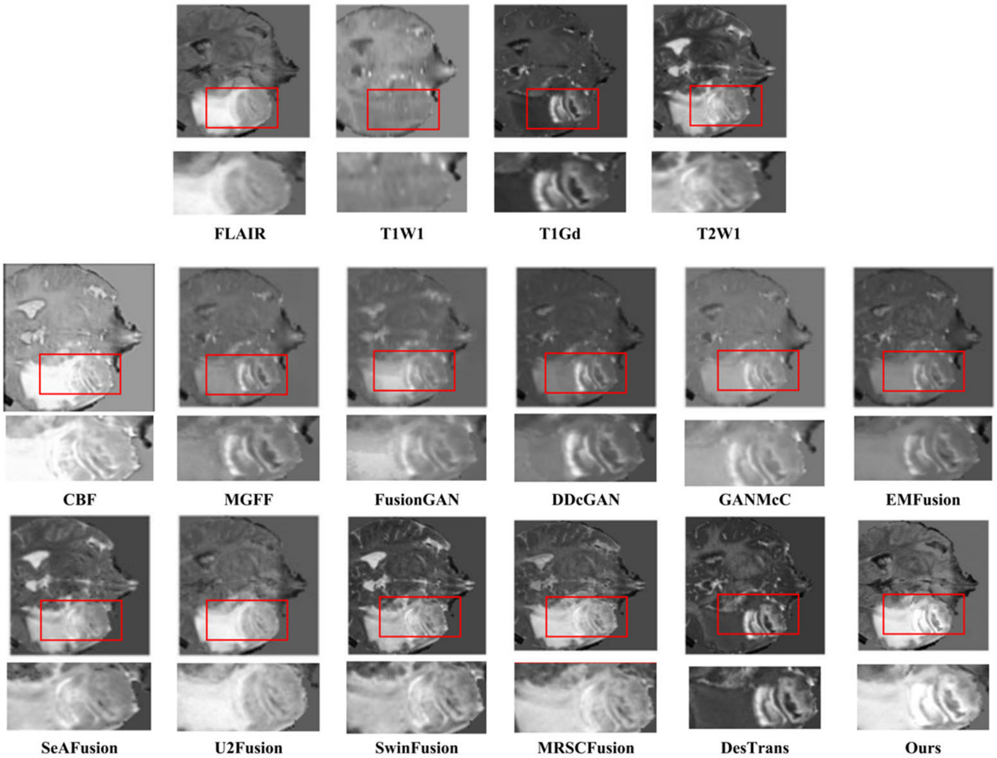


(a) The top panel represents the original multimodal image, and the bottom panel represents the visualization results of the comparative experiment. Zoom in results in red boxes are below.

(b) The top panel represents the original multimodal image, and the bottom panel represents the visualization results of the comparative experiment. Zoom in results in red boxes are below.

FIGURE 8. On the three typical image pairs of the Brats2019 dataset. a, b, and c correspond to three typical images. The fusion results obtained by CBF, MGFF, FusionGAN, DDcGAN, GANMcC, EMFusion, SeAFusion, U2Fusion, SwinFusion, MRSCFusion, DesTran and TIEF are shown in order. The enlarged section in the bottom corner provides a more detailed comparison.

The results showed that CBF introduced noise and reduced image quality. MGFF, FusionGAN, GANMcC, SwinFusion, and U2Fusion images are blurred, lack texture detail, and have low pixel intensity and contrast in critical areas.



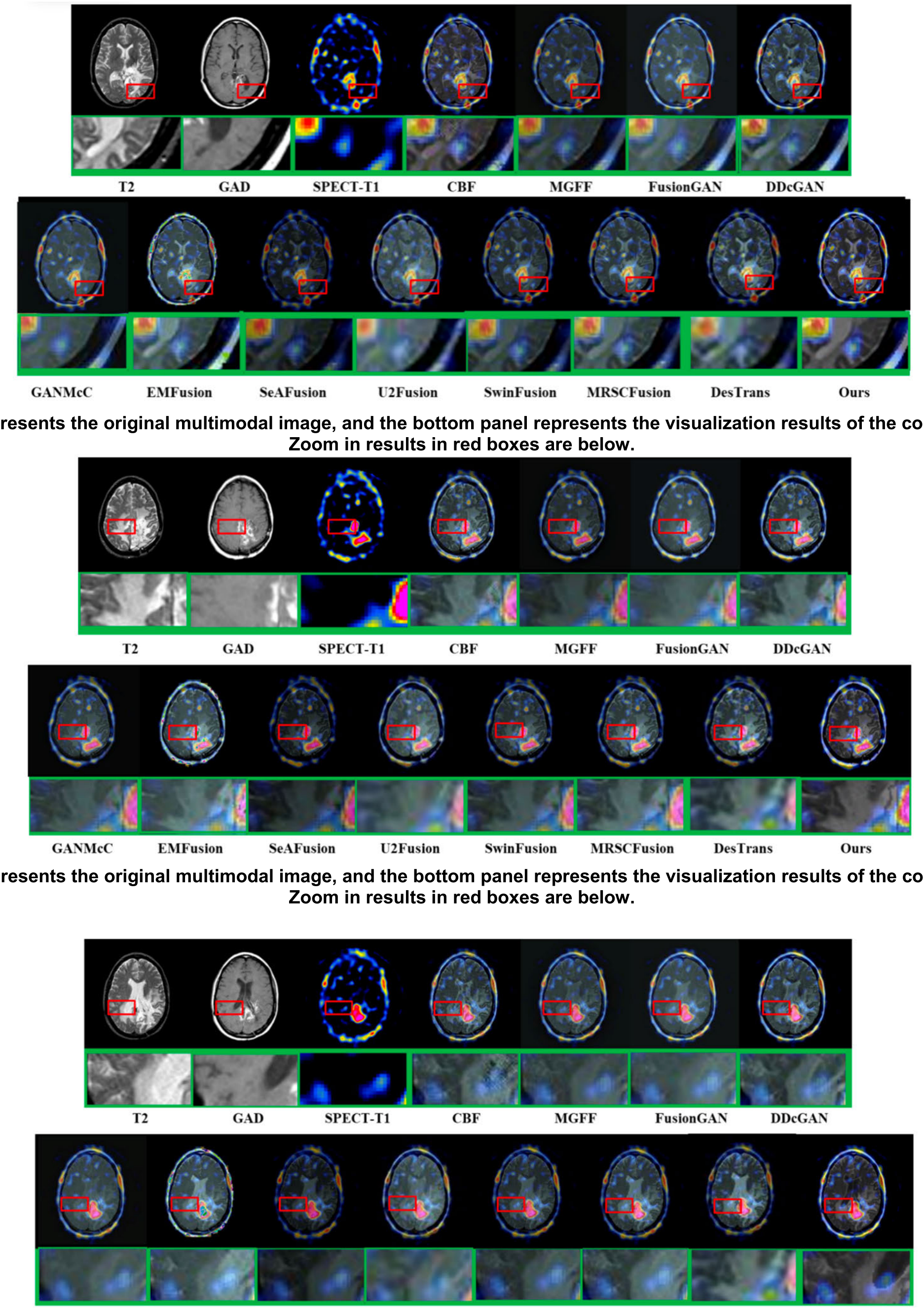
(c) The top panel represents the original multimodal image, and the bottom panel represents the visualization results of the comparative experiment. Zoom in results in red boxes are below.

FIGURE 8. *(Continued.)* On the three typical image pairs of the Brats2019 dataset. a, b, and c correspond to three typical images. The fusion results obtained by CBF, MGFF, FusionGAN, DDcGAN, GANMcC, EMFusion, SeAFusion, U2Fusion, SwinFusion, MRSCFusion, DesTran and TIEF are shown in order. The enlarged section in the bottom corner provides a more detailed comparison.

TABLE 2. Quantitative comparison of different methods for 8 evaluation items indicators in the AANLIB dataset (Red: Optimal, Blue: Suboptimal).

Categories	Methods	$Q_{\frac{AB}{F}}$	EN	AG	SF	CI	MI	PSNR	SSIM
traditional methods	CBF	0.424	13.039	6.614	43.000	17.570	4.021	46.003	0.578
	MGFF	0.532	10.473	6.856	31.100	11.344	3.640	42.039	0.368
CNN	EMFusion	0.572	11.522	9.660	41.250	19.153	5.114	45.673	0.522
	SeAFusion	0.480	11.895	11.623	38.290	19.304	5.219	46.159	0.641
	U2Fusion	0.507	10.093	12.538	44.000	22.070	6.997	46.275	0.542
GAN	FusionGAN	0.377	10.696	10.148	36.990	18.455	3.932	48.116	0.310
	DDcGAN	0.466	9.726	7.930	35.700	13.350	4.318	46.364	0.592
	GANMcC	0.549	11.647	8.067	37.670	11.690	4.757	47.079	0.538
Transformer	SwinFusion	0.555	13.035	11.573	45.842	19.417	4.656	48.285	0.612
	MRSCFusion	0.476	13.344	10.230	35.812	16.021	4.504	47.230	0.548
	DesTran	0.564	12.214	11.725	43.148	20.215	5.415	48.291	0.625
Our	TIEF	0.734	14.204	13.898	46.000	22.461	5.423	48.329	0.629

SeAFusion retained relatively rich texture and structural information, but the overall image was dark, probably due to the ubiquitous black background in SPECT-T1WI, which affected the pixel intensity of the final image. MRSCFusion, DDcGAN, EMFusion, and DesTrans retain more comprehensive source image information but require enhanced detail contrast. Compared with other methods, TIEF effectively preserves the intricate texture of the source image and the color information of RGB, which is more suitable for human visual perception. Based on the quantitative analysis, TIEF obtained the optimal EN, SF, $Q_{AB/F}$, CI, AG, and the suboptimal MI, PSNR, and SSIM in the



(a) The top panel represents the original multimodal image, and the bottom panel represents the visualization results of the comparative experiment. Zoom in results in red boxes are below.

(b) The top panel represents the original multimodal image, and the bottom panel represents the visualization results of the comparative experiment. Zoom in results in red boxes are below.

(c) The top panel represents the original multimodal image, and the bottom panel represents the visualization results of the comparative experiment. Zoom in results in red boxes are below.

FIGURE 9. On the three typical image pairs of the ANNLIB dataset. a, b, and c correspond to three typical images. The fusion results obtained by CBF, MGFF, FusionGAN, DDcGAN, GANMcC, EMFusion, SeAFusion, U2Fusion, SwinFusion, MRSCFusion, DesTran and TIEF are shown in order. The enlarged section in the bottom corner provides a more detailed comparison.

test images. The main reasons for suboptimal MI and PSNR are the same as the BraTs2019 dataset, but the main reason for suboptimal SSIM is that the SeAFusion network takes fine-grained details into account when it is constructed, and the SeAFusion network does not use any downsampling, which indicates that SEAFusion keeps more similar information. This is why the SSIM of SeAFusion is higher than that of TIEF 0.012. However, from the visual point of view, the gradient change inside the tumor of TIEF is more prominent, while there is no gradient change inside the tumor of SeAFusion, which will have a particular impact on the localization of the cancer, and this result is further verified in the downstream segmentation task.

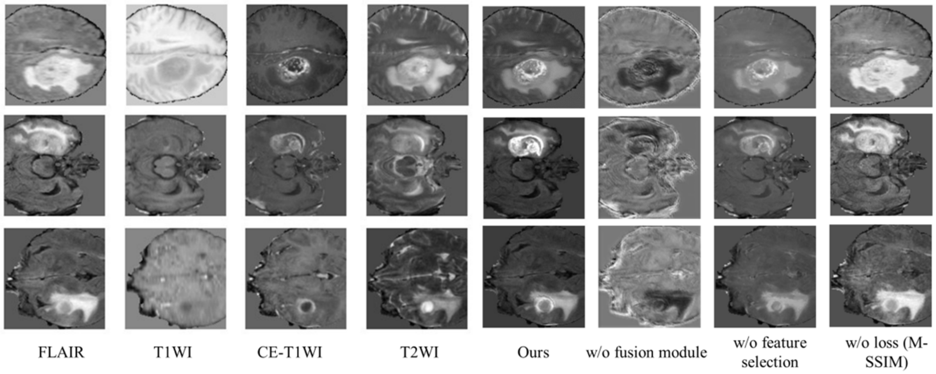


FIGURE 10. Qualitative comparison of three typical image pairs in the BraTs2019 dataset to validate the effect of different blocks. From left to right, the sequence comprises FLAIR, T1WI, T1Gd, and T2WI images, followed by the fusion results of our method, fusion results without the fusion block, fusion results without the feature selection block, and fusion results without the feature loss (M-SSIM) block.

TABLE 3. Quantitative results obtained with a combination of different blocks in the Brats2019 dataset (Red: Optimal, Blue: Suboptimal).

Combination of different blocks	$Q_{\frac{AB}{F}}$	EN	AG	SF	CI	MI	PSNR	SSIM
Baseline block	0.276	8.173	6.228	21.000	0.639	3.872	30.140	0.323
Baseline+L_{M-SSIM}	0.198	8.396	6.688	20.000	0.773	3.953	32.010	0.318
Baseline+feature selection	**0.394**	8.353	6.701	24.000	0.923	4.222	36.179	0.322
Baseline+fusion block	0.250	8.426	9.220	37.000	1.229	**4.814**	40.513	0.408
Baseline+L_{M-SSIM}+ feature selection	0.317	**9.347**	9.319	25.000	1.715	4.253	37.231	0.347
Baseline+ feature selection + fusion block	0.322	9.222	**10.066**	**39.000**	**1.788**	4.323	**46.014**	**0.539**
Baseline+ L_{M-SSIM}+ fusion block	**0.344**	**9.595**	**9.595**	**43.000**	**1.748**	**4.822**	**48.522**	**0.426**
ALL	**0.495**	**13.226**	**11.239**	**45.700**	**2.450**	**5.870**	**49.472**	**0.676**

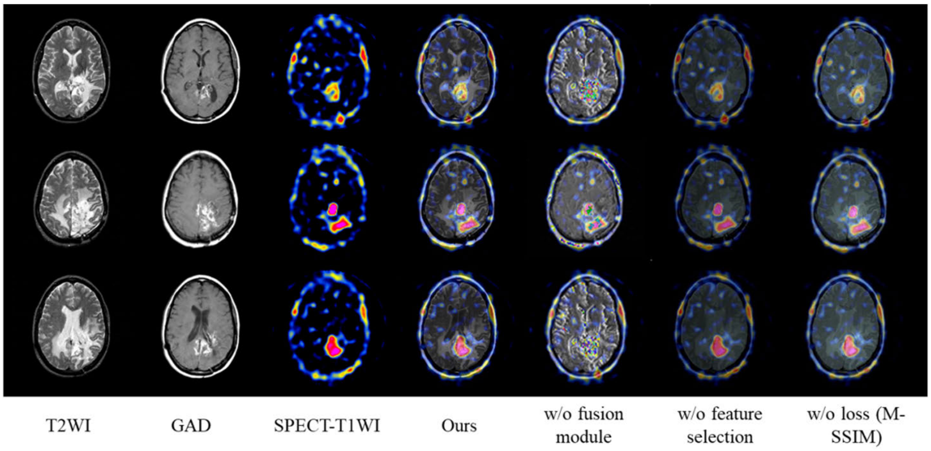


FIGURE 11. Qualitative comparison of three typical image pairs in the AANLIB dataset to validate the effect of different blocks. The sequence from left to right includes T2WI, GAD, and SPECT-T1WI images, followed by the fusion results of our method, fusion results without the fusion block, fusion results without the feature selection block, and fusion results without the feature loss (M-SSIM) block.

3) ABLATION STUDY

To assess the impact of various blocks on the model's efficacy, we conducted experiments on both the BraTs2019 and AANLIB test datasets. Three representative images were selected for these experimental evaluations. These images embody diverse modalities, each showcasing distinct tumor-related information. The primary aim was to ensure that the fusion results maintained the intricate details from the source maps while accentuating discrepancies among tumor regions. The adopted ablation experiments encompassed different

TABLE 4. Quantitative results obtained with a combination of different blocks in the AANLIB dataset (Red: Optimal, Blue: Suboptimal).

Combination of different blocks	$Q_{\frac{AB}{F}}$	EN	AG	SF	CI	MI	PSNR	SSIM
Baseline block	0.276	8.173	9.228	21.000	0.639	3.872	30.140	0.323
Baseline+L_{M-SSIM}	0.198	8.396	6.688	20.000	0.773	3.953	32.010	0.318
Baseline+feature selection	**0.394**	8.353	8.701	24.000	0.923	4.222	36.179	0.322
Baseline+fusion block	0.250	7.426	9.220	37.000	1.229	**4.814**	**40.513**	0.408
Baseline+L_{M-SSIM}+ feature selection	0.317	**9.347**	9.319	25.000	1.715	4.253	35.231	0.347
Baseline+ feature selection + fusion block	0.322	9.222	**10.066**	**39.000**	**1.788**	4.323	36.014	**0.539**
Baseline+ L_{M-SSIM}+ fusion block	**0.344**	**9.595**	**9.595**	**43.000**	**1.798**	**4.822**	**38.522**	**0.426**
ALL	**0.734**	**14.204**	**13.898**	**46.000**	**22.461**	**5.423**	**48.329**	**0.629**

TABLE 5. Segmentation task results.

Categories	Methods	Dice			HD95		
		ET	TC	WT	ET	TC	WT
traditional methods	CBF	0.542	0.513	0.441	145.738	144.024	195.406
	MGFF	0.580	0.603	0.649	108.573	132.906	170.759
GAN	FusionGAN	0.758	0.819	0.793	57.077	36.427	60.952
	DDcGAN	0.763	0.821	0.771	43.386	27.455	56.390
	GANMcC	0.784	0.848	0.783	39.242	18.874	58.570
CNN	EMFusion	0.782	0.806	0.685	41.921	35.990	92.324
	SeAFusion	0.776	0.813	0.844	55.623	33.772	34.245
	U2Fusion	0.795	0.850	0.844	44.800	20.484	37.065
Transformer	SwinFusion	0.775	0.799	0.741	47.156	28.621	70.157
	MRSCFusion	0.838	0.855	0.901	14.935	13.867	10.919
	DesTran	0.820	0.822	0.863	29.207	19.525	28.434
Our	TIEF	**0.866**	**0.885**	**0.907**	**14.159**	**11.256**	**10.918**

combinations of the loss function, feature selection, and fusion blocks. The "Baseline" scenario denotes training the network without any additional blocks. Observing the outcomes (Fig. 10), in the absence of the fusion block, although the fused image retains part of the texture and structural information and the tumor area is also apparent, there is a significant deviation from the actual image. This bias leads to substantial information loss, contrary to human visual perception. Comparing the results across experiments (Table. 3), the addition of the fusion block enhances various performance metrics such as $Q_{AB/F}$, EN, SF, MI, and PSNR, while showing suboptimal performance in AG, CI, and SSIM metrics.

4) ABLATION STUDY

To assess the impact of various blocks on the model's efficacy, we conducted experiments on both the BraTs2019 and AANLIB test datasets. Three representative images were selected for these experimental evaluations. These images embody diverse modalities, each showcasing distinct tumor-related information. The primary aim was to ensure that the fusion results maintained the intricate details from the source maps while accentuating discrepancies among tumor regions.

The adopted ablation experiments encompassed different combinations of the loss function, feature selection, and fusion blocks. The "Baseline" scenario denotes training the network without any additional blocks. Observing the outcomes (Fig. 10), in the absence of the fusion block, although the fused image retains part of the texture and structural information and the tumor area is also apparent, there is a significant deviation from the actual image. This bias leads to substantial information loss, contrary to human visual perception. Comparing the results across experiments (Table. 3), the addition of the fusion block enhances various performance metrics such as $Q_{AB/F}$, EN, SF, MI, and PSNR, while showing suboptimal performance in AG, CI, and SSIM metrics.

5) ABLATION STUDY

To assess the impact of various blocks on the model's efficacy, we conducted experiments on both the BraTs2019 and AANLIB test datasets. Three representative images were

TABLE 6. Quantitative comparison of different methods for 8 evaluation items indicators in the AANLIB meningioma dataset.

Categories	Methods	$Q_{\frac{AB}{F}}$	EN	AG	SF	CI	MI	PSNR	SSIM
traditional methods	CBF	0.355	5.638	5.102	15.311	13.733	3.239	10.545	0.300
	MGFF	0.298	5.771	9.630	35.941	20.765	4.499	17.204	0.498
CNN	EMFusion	0.447	5.374	7.765	24.913	20.771	4.804	15.107	0.584
	SeAFusion	0.546	5.485	9.753	35.454	17.751	4.471	16.916	0.529
	U2Fusion	0.537	5.807	7.571	23.671	15.727	4.008	15.755	0.555
GAN	FusionGAN	0.317	6.921	10.599	28.994	16.645	2.942	10.094	0.268
	DDcGAN	0.409	6.205	11.103	31.197	4.590	3.703	11.727	0.489
	GANMcC	0.353	5.943	9.821	30.329	17.696	3.881	18.388	0.558
Transformer	SwinFusion	0.476	5.877	8.418	31.848	17.753	4.618	16.460	0.407
	MRSCFusion	0.538	5.268	8.619	33.245	19.785	5.401	15.697	0.581
	DesTran	0.447	5.849	7.875	26.719	15.752	4.139	11.967	0.337
Our	TIEF	**0.580**	**7.100**	**12.911**	**38.003**	**21.456**	**5.417**	**20.334**	**0.597**

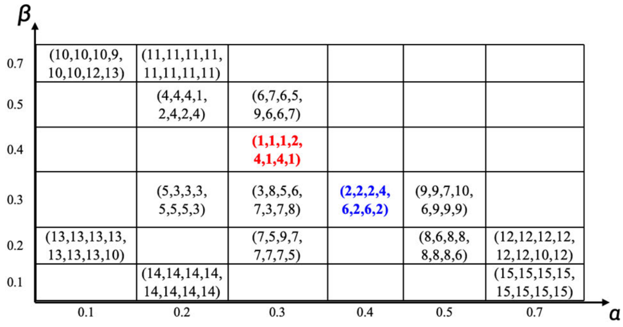
selected for these experimental evaluations. These images embody diverse modalities, each showcasing distinct tumor-related information. The primary aim was to ensure that the fusion results maintained the intricate details from the source maps while accentuating discrepancies among tumor regions. The adopted ablation experiments encompassed different combinations of the loss function, feature selection, and fusion blocks. The "Baseline" scenario denotes training the network without any additional blocks. Observing the outcomes (Fig. 10), in the absence of the fusion block, although the fused image retains part of the texture and structural information and the tumor area is also apparent, there is a significant deviation from the actual image. This bias leads to substantial information loss, contrary to human visual perception. Comparing the results across experiments (Table. 3), the addition of the fusion block enhances various performance metrics such as $Q_{AB/F}$, EN, SF, MI, and PSNR, while showing suboptimal performance in AG, CI, and SSIM metrics.

The fusion block better integrated the original image information and fused imaging. This underscores its critical role in producing high-quality fusion results. In this study, the baseline approach of feature selection blocks to enhance fusion results was compared with TIEF, eliminating the effect of feature selection blocks. For example, although the image's edge texture information from the first row of T2WI is very prominent, this detail must be accurately represented in the fusion results. The advantages of using feature selection blocks become more apparent through our comparison. The baseline method with a feature selection block significantly improved the evaluation indicators compared with the baseline method alone. From the comparative analysis in the table, the lack of feature selection reduces the information richness of the source map in the fusion results. The experimental results show that the feature selection block dramatically improves the network's fusion effect. In addition, the impact of adding loss function blocks to the experiment is investigated. The addition of this block significantly improved EN, AG, CI, MI, PSNR, and other evaluation indicators. Comparing the TIEF method with or without the addition of the loss function block confirmed its importance in enhancing the contrast between different tumor tissues. The figure shows that the paired comparison of the fusion results could be better without the loss function, highlighting the superiority of TIEF.

The same test was conducted on the AANLIB dataset. Fig. 11 shows that the experimental results without the fusion block deviate from the image's source and distort the color part. The absence of the feature selection block in the experiments leads to a noticeable lack of detailed information, resulting in blurred images. Similarly, when the experiments lacked a loss function block, the contrast in the front and back scenes was notably diminished. Table. 4 also illustrates the different effects of the fusion, feature selection, and loss function blocks from different perspectives.

6) DOWNSTREAM TASK VALIDATION

To verify the model's effectiveness further, we connected the segmentation network (no-new-Net) nn-Unet after 11 comparison methods and TIEF to further verify the model's effectiveness through the effect of segmentation. The obtained segmentation results are shown in Table 5. Dice represents the similarity of the two samples, ranging from 0 to 1, and the closer to 1, the better the segmentation effect. HD95 indicates the degree of overlap of the boundaries, and smaller values represent better segmentation. ET represents the enhancing tumor, TC represents the tumor core, and WT represents the whole tumor. Through the segmentation results of enhancing tumor, tumor core, and the whole tumor, it was found that TIEF had the largest Dice and the lowest HD95,



β \ α	0.1	0.2	0.3	0.4	0.5	0.7
0.7	(10,10,10,9, 10,10,12,13)	(11,11,11,11, 11,11,11,11)				
0.5		(4,4,4,1, 2,4,2,4)	(6,7,6,5, 9,6,6,7)			
0.4			**(1,1,1,2, 4,1,4,1)**			
0.3		(5,3,3,3, 5,5,5,3)	(3,8,5,6, 7,3,7,8)	**(2,2,2,4, 6,2,6,2)**	(9,9,7,10, 6,9,9,9)	
0.2	(13,13,13,13, 13,13,13,10)		(7,5,9,7, 7,7,7,5)		(8,6,8,8, 8,8,8,6)	(12,12,12,12, 12,12,10,12)
0.1		(14,14,14,14, 14,14,14,14)				(15,15,15,15, 15,15,15,15)

FIGURE 12. The parameters are selected as grid plots. Values represent index ranking.

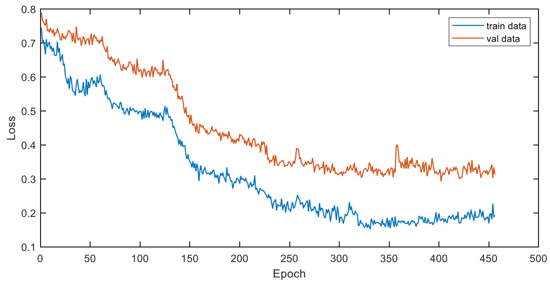


FIGURE 13. Loss function curve.

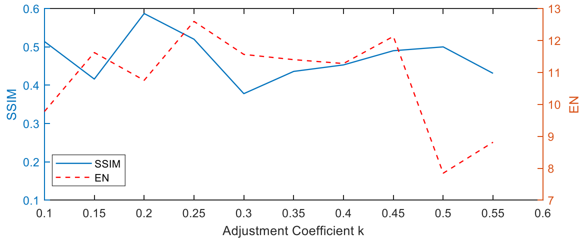


FIGURE 14. Curves for changes in age-adjusted coefficients of EN and SSIM.

indicating that the segmentation effect of TIEF-based image fusion was the best, indicating that the fused image of TIEF was of high quality.

7) GENERALIZATION STUDY

To further verify the generalization of the model. We used meningioma CT and MRI data from the ANNLIB data set. Meningioma is a primary intracranial tumor, primarily benign and asymptomatic in the early stage. However, with the compression of the tumor, headache and epilepsy may occur, and the loss of vision, hearing, and smell may occur in severe cases. Meningiomas grow between the human skull and brain tissue, which differs from the growth location of glioma. Early screening and diagnosis of meningioma can prolong the survival of patients. TIEF performed a fusion of CT and MRI, and the results are shown in Table 6. By calculating the generalization results, it was found that the indexes reached optimal, indicating that the generalization effect of TIEF was good.

E. HYPER PARAMETERS COMPARISON

In Eq. (17), α, β, η is the balance coefficient ($\eta = 1-\alpha-\beta, \alpha \neq 0, \beta \neq 0, \eta \neq 0$). In this paper, we set the value of $\alpha = \{0.1,0.2,0.3,0.4,0.5,0.7\}$ and the value of $\beta = \{0.1,0.2,0.3,0.4,0.5,0.7\}$ for experimentation. When $\alpha = 0.6$ or $\beta = 0.6$, the loss function fails to converge. Each combination was assessed based on various evaluation indices, and the values were recorded and ranked in descending order. The best result is when $\alpha = 0.3, \beta = 0.4$, $\eta = 1-\alpha-\beta = 0.3$. Observing the loss function of the pre-experiment, it is found that TIEF begins to decline smoothly at an epoch equal to 200 and reaches convergence around an epoch equal to 320, so epoch 320 is selected for the formal experiment in this paper. In this paper, κ is the adjustment coefficient, which is used to control the proportional change of information entropy of different modes in weight calculation, to ensure the best fusion effect of the model. κ mainly plays a key role in Eq. (10) and Eq. (11), which correspond to the two evaluation indexes EN and SSIM respectively. The curve of EN and SSIM with the change of κ value showed that when κ was 0.25, both EN and SSIM reached the maximum.

V. CONCLUSION

This paper proposed a multi-modal MRI image fusion method, emphasizing the description of the edema part, enhanced tumor, and necrotic tumor core in different modalities to generate a fusion image with rich texture information and clear structure. In the coding region, based on denseness, we adopted the parallel double-branch design of deep feature extraction and structural feature extraction to maintain the balance between structural information and functional information to better extract and transmit the knowledge of the source image. The feature information measurement method based on information entropy was introduced, and the feature channels containing rich texture information and complex structure information were selected to fuse with the deep features to enhance the richness of the fused image information. The content richness of the source image of different modes was used as the weight and combined with the regional structural similarity index and regional contrast, the loss function was constructed to enhance the difference between tumor tissues. The transformer block with an attention mechanism replaced the manually designed fusion strategy, and the cross-modal image features were fused. Experiments demonstrate that the linked images generated by the TIEF model produce satisfactory results in multi-modal brain tumor image fusion tasks for both MMRI, SPECT-MRI and CT-MRI images. Qualitative and quantitative analysis verified the validity and generalization of TIEF.

REFERENCES

[1] D. Xiao, C. Yan, D. Li, T. Xi, X. Liu, D. Zhu, G. Huang, J. Xu, Z. He, A. Wu, C. Ma, J. Long, and K. Shu, "National brain tumour registry of China (NBTRC) statistical report of primary brain tumours diagnosed in China in years 2019–2020," *Lancet Regional Health-Western Pacific*, vol. 34, May 2023, Art. no. 100715.

[2] T. M. Mack and M. Cockburn, *Cancers in the Urban Environment*. Amsterdam, The Netherlands: Elsevier, 2004, pp. 5–8.

[3] D. Ray et al., “Apoptosis reference block in biomedical sciences,” *Biomed. Sci.*, vol. 10, pp. 10–16, Dec. 2014.

[4] J. R. Foster et al., “Introduction to neoplasia,” in *Comprehensive Toxicology*, vol. 14, Feb. 2017, pp. 1–10.

[5] J. Yokota, “Tumor progression and metastasis,” *Carcinogenesis*, vol. 21, no. 3, pp. 497–503, 2000.

[6] S. J. Moon et al., “Tumors of the brain central nervous system cancer rehabilitation,” *Tumors Brain*, vol. 10, pp. 19–39, Feb. 2019.

[7] H. Sontheimer, “Infectious diseases of the nervous system,” 2015.

[8] N. Reynoso-Noverón et al., “Epidemiology of brain tumors,” in *Principles of Neuro-Oncology*, vol. 7, Dec. 2020, pp. 15–25.

[9] S. Li, B. Yang, and J. Hu, “Performance comparison of different multi-resolution transforms for image fusion,” *Inf. Fusion*, vol. 12, no. 2, pp. 74–84, Apr. 2011.

[10] Y. Zheng, E. Blasch, and Z. Liu, *Multispectral Image Fusion and Colorization*. Bellingham, WA, USA: SPIE Press, Mar. 2018.

[11] J. Jose, N. Gautam, M. Tiwari, T. Tiwari, A. Suresh, V. Sundararaj, and R. Mr, “An image quality enhancement scheme employing adolescent identity search algorithm in the NSST domain for multimodal medical image fusion,” *Biomed. Signal Process. Control*, vol. 66, Apr. 2021, Art. no. 102480.

[12] K. Zhang, Y. Huang, and C. Zhao, “Remote sensing image fusion via RPCA and adaptive PCNN in NSST domain,” *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 16, no. 5, Sep. 2018, Art. no. 1850037.

[13] J. Ma, C. Chen, C. Li, and J. Huang, “Infrared and visible image fusion via gradient transfer and total variation minimization,” *Inf. Fusion*, vol. 31, pp. 100–109, Sep. 2016.

[14] W. Xu, Y.-L. Fu, H. Xu, and K. K. L. Wong, “Medical image fusion using enhanced cross-visual cortex model based on artificial selection and impulse-coupled neural network,” *Comput. Methods Programs Biomed.*, vol. 229, Feb. 2023, Art. no. 107304.

[15] A. R. Groves, C. F. Beckmann, S. M. Smith, and M. W. Woolrich, “Linked independent component analysis for multimodal data fusion,” *NeuroImage*, vol. 54, no. 3, pp. 2198–2217, Feb. 2011.

[16] W. Huang, H. Zhang, H. Guo, W. Li, X. Quan, and Y. Zhang, “ADDNS: An asymmetric dual deep network with sharing mechanism for medical image fusion of CT and MR-T2,” *Comput. Biol. Med.*, vol. 166, Nov. 2023, Art. no. 107531.

[17] B. Zhan, D. Li, X. Wu, J. Zhou, and Y. Wang, “Multi-modal MRI image synthesis via GAN with multi-scale gate mergence,” *IEEE J. Biomed. Health Informat.*, vol. 26, no. 1, pp. 17–26, Jan. 2022.

[18] J. Huang, Z. Le, Y. Ma, F. Fan, H. Zhang, and L. Yang, “MGMDc-GAN: Medical image fusion using multi-generator multi-discriminator conditional generative adversarial network,” *IEEE Access*, vol. 8, pp. 55145–55157, 2020.

[19] T. Zhou, Q. Li, H. Lu, Q. Cheng, and X. Zhang, “GAN review: Models and medical image fusion applications,” *Inf. Fusion*, vol. 91, pp. 134–148, Mar. 2023.

[20] M. Jiang, M. Zhi, L. Wei, X. Yang, J. Zhang, Y. Li, P. Wang, J. Huang, and G. Yang, “FA-GAN: Fused attentive generative adversarial networks for MRI image super-resolution,” *Computerized Med. Imag. Graph.*, vol. 92, Sep. 2021, Art. no. 101969.

[21] L. Wang, C. Chang, B. Hao, and C. Liu, “Multi-modal medical image fusion based on GAN and the shift-invariant shearlet transform,” in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2020, pp. 2538–2543.

[22] J. Kang, W. Lu, and W. Zhang, “Fusion of brain PET and MRI images using tissue-aware conditional generative adversarial network with joint loss,” *IEEE Access*, vol. 8, pp. 6368–6378, 2020.

[23] K. Guo, X. Hu, and X. Li, “MMFGAN: A novel multimodal brain medical image fusion based on the improvement of generative adversarial network,” *Multimedia Tools Appl.*, vol. 81, no. 4, pp. 5889–5927, Feb. 2022.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, and L. Jones, “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–13.

[25] W. Li, Y. Zhang, G. Wang, Y. Huang, and R. Li, “DFENet: A dual-branch feature enhanced network integrating transformers and convolutional feature learning for multimodal medical image fusion,” *Biomed. Signal Process. Control*, vol. 80, Feb. 2023, Art. no. 104402.

[26] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[27] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “SwinIR: Image restoration using Swin transformer,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[28] J. Du, W. Li, B. Xiao, and Q. Nawaz, “Union Laplacian pyramid with multiple features for medical image fusion,” *Neurocomputing*, vol. 194, pp. 326–339, Jun. 2016.

[29] S. Li, H. Yin, and L. Fang, “Group-sparse representation with dictionary learning for medical image denoising and fusion,” *IEEE Trans. Biomed. Eng.*, vol. 59, no. 12, pp. 3450–3459, Dec. 2012.

[30] S. Maqsood and U. Javed, “Multi-modal medical image fusion based on two-scale image decomposition and sparse representation,” *Biomed. Signal Process. Control*, vol. 57, Mar. 2020, Art. no. 101810.

[31] C. Du and S. Gao, “Image segmentation-based multi-focus image fusion through multi-scale convolutional neural network,” *IEEE Access*, vol. 5, pp. 15750–15761, 2017.

[32] Z. Wang, X. Li, H. Duan, X. Zhang, and H. Wang, “Multifocus image fusion using convolutional neural networks in the discrete wavelet transform domain,” *Multimedia Tools Appl.*, vol. 78, no. 24, pp. 34483–34512, Aug. 2019.

[33] X. Guo, R. Nie, J. Cao, D. Zhou, L. Mei, and K. He, “FuseGAN: Learning to fuse multi-focus image via conditional generative adversarial network,” *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 1982–1996, Aug. 2019.

[34] D. Rao, T. Xu, and X.-J. Wu, “TGFuse: An infrared and visible image fusion approach based on transformer and generative adversarial network,” *IEEE Trans. Image Process.*, early access, May 10, 2023, doi: 10.1109/TIP.2023.3273451.

[35] Y. Fu, T. Xu, X. Wu, and J. Kittler, “PPT fusion: Pyramid patch transformerfor a case study in image fusion,” 2021, *arXiv:2107.13967*.

[36] V. Vs, J. M. Jose Valanarasu, P. Oza, and V. M. Patel, “Image fusion transformer,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 3566–3570.

[37] L. Qu, S. Liu, M. Wang, and Z. Song, “Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning,” in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 2126–2134.

[38] D. P. Bavirisetti, G. Xiao, J. Zhao, R. Dhuli, and G. Liu, “Multi-scale guided image and video fusion: A fast and efficient approach,” *Circuits, Syst., Signal Process.*, vol. 38, no. 12, pp. 5576–5605, May 2019.

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[40] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, and L. M. Prevedello, “The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification,” 2021, *arXiv:2107.02314*.

[41] D. Summers, “Harvard whole brain atlas,” *J. Neurol. Neurosurgery Psychiatry*, vol. 73, no. 3, p. 288, 2003.

[42] B. K. S. Kumar, “Image fusion based on pixel significance using cross bilateral filter,” *Signal, Image Video Process.*, vol. 9, no. 5, pp. 1193–1204, Jul. 2015.

[43] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, “U2Fusion: A unified unsupervised image fusion network,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.

[44] H. Xu and J. Ma, “EMFusion: An unsupervised enhanced medical image fusion network,” *Inf. Fusion*, vol. 76, pp. 177–186, Dec. 2021.

[45] L. Tang, J. Yuan, and J. Ma, “Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network,” *Inf. Fusion*, vol. 82, pp. 28–42, Jun. 2022.

[46] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, “FusionGAN: A generative adversarial network for infrared and visible image fusion,” *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.

[47] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, “DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion,” *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.

[48] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu, "GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.

[49] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via Swin transformer," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 7, pp. 1200–1217, Jul. 2022.

[50] X. Xie, X. Zhang, S. Ye, D. Xiong, L. Ouyang, B. Yang, H. Zhou, and Y. Wan, "MRSCFusion: Joint residual Swin transformer and multiscale CNN for unsupervised multimodal medical image fusion," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–17, 2023.

[51] Y. Song, Y. Dai, W. Liu, Y. Liu, X. Liu, Q. Yu, X. Liu, N. Que, and M. Li, "DesTrans: A medical image fusion method based on transformer and improved DenseNet," *Comput. Biol. Med.*, vol. 174, May 2024, Art. no. 108463.

[52] G. Cui, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition," *Opt. Commun.*, vol. 341, pp. 199–209, Apr. 2015.

[53] N. Yu, T. Qiu, F. Bi, and A. Wang, "Image features extraction and fusion based on joint sparse representation," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 5, pp. 1074–1082, Sep. 2011.

[54] J. Van Aardt, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *J. Appl. Remote Sens.*, vol. 2, no. 1, May 2008, Art. no. 023522.

[55] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," *Electron. Lett.*, vol. 38, no. 7, p. 313, 2002.

[56] A. R. Alankrita, A. Shrivastava, and V. Bhateja, "Contrast improvement of cerebral MRI features using combination of non-linear enhancement operator and morphological filter," *IEEE J. Sel. Topics Signal Process.*, vol. 4, pp. 182–187, 2011.

[57] C. S. Xydeas and V. Petrovic, "Objective image fusion performance measure," *Electron. Lett.*, vol. 36, no. 4, pp. 308–309, 2000.

YUTING ZHOU received the bachelor's degree in applied statistics from Southern Medical University, in 2019, and the master's degree in applied statistics from Northeast Normal University, in 2021. She is currently pursuing the Ph.D. degree in computational mathematics with China Academy of Engineering Physics. Her current research interests include image processing, computer vision, and information fusion.

XUEMEI YANG received the Bachelor of Science degree in mathematics from Minzu University of China, in 2015, and the Master of Science degree in statistics from North China Electric Power University, in 2019. She is currently pursuing the degree in computational mathematics with China Academy of Engineering Physics. Her research interest includes multimodal information fusion.

SHIQI LIU received the Bachelor of Science degree in mathematics and applied mathematics from Beihang University, in 2021. She is currently pursuing the Ph.D. degree in computational mathematics with China Academy of Engineering Physics. Her research interests include biomedical signal processing, time series analysis, and deep learning.

JUNPING YIN received the Bachelor of Science and the Master of Science degrees from the School of Mathematics and Statistics, Northeast Normal University, in 2002 and 2005, respectively, and the Doctor of Science degree from the School of Mathematics, Xiamen University, in 2008. He is currently a Researcher with the Institute of Applied Physics and Computational Mathematics, Beijing, and the President of Shanghai Zhangjiang Institute of Mathematics. He has been engaged in data science and applied mathematics research for a long time and presided over more than 20 major projects of the National Natural Science Foundation. He has published more than 30 papers, more than ten patents and soft books, and one monograph.