MATR: Multimodal Medical Image Fusion via Multiscale Adaptive Transformer

Wei Tang, Fazhi He, *Member, IEEE*, Yu Liu, *Member, IEEE*, and Yansong Duan

***Abstract*— Owing to the limitations of imaging sensors, it is challenging to obtain a medical image that simultaneously contains functional metabolic information and structural tissue details. Multimodal medical image fusion, an effective way to merge the complementary information in different modalities, has become a significant technique to facilitate clinical diagnosis and surgical navigation. With powerful feature representation ability, deep learning (DL)-based methods have improved such fusion results but still have not achieved satisfactory performance. Specifically, existing DL-based methods generally depend on convolutional operations, which can well extract local patterns but have limited capability in preserving global context information. To compensate for this defect and achieve accurate fusion, we propose a novel unsupervised method to fuse multimodal medical images via a multiscale adaptive Transformer termed MATR. In the proposed method, instead of directly employing vanilla convolution, we introduce an adaptive convolution for adaptively modulating the convolutional kernel based on the global complementary context. To further model long-range dependencies, an adaptive Transformer is employed to enhance the global semantic extraction capability. Our network architecture is designed in a multiscale fashion so that useful multimodal information can be adequately acquired from the perspective of different scales. Moreover, an objective function composed of a structural loss and a region mutual information loss is devised to construct constraints for information preservation at both the structural-level and the feature-level. Extensive experiments on a mainstream database demonstrate that the proposed method outperforms other representative and state-of-the-art methods in terms of both visual quality and quantitative evaluation. We also extend the proposed method to address other biomedical image fusion issues, and the pleasing fusion results illustrate that MATR has good generalization capability. The code of the proposed method is available at https://github.com/tthinking/MATR.**

***Index Terms*— Image fusion, biomedical image, transformer, adaptive convolution, deep learning.**

Manuscript received 20 December 2021; revised 31 March 2022, 19 May 2022, and 6 July 2022; accepted 14 July 2022. Date of publication 28 July 2022; date of current version 3 August 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62072348 and Grant 62176081, in part by the Science and Technology Major Project of Hubei Province (Next-Generation Artificial Intelligence (AI) Technologies) under Grant 2019AEA170, and in part by the National Key Research and Development Program of China under Grant 2019YFC1509604. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ming Shao. *(Corresponding author: Fazhi He.)*

Wei Tang and Fazhi He are with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: weitang2021@whu.edu.cn; fzhe@whu.edu.cn).

Yu Liu is with the Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: yuliu@hfut.edu.cn).

Yansong Duan is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: ysduan@whu.edu.cn).

Digital Object Identifier 10.1109/TIP.2022.3193288

I. Introduction

IMAGE fusion aims to combine information from two or more source images and generate an informative composite image [1], which can promote the performance of subsequent computer vision tasks [2]. In this domain, multimodal medical image fusion has drawn increasing attention due to its significant clinical applications including tumor segmentation [3], cell classification [4], neurological research [5], and treatment strategies for recurrent high-grade gliomas [6]. In this paper, we focus on functional (i.e., single-photon emission computed tomography (SPECT)) and structural (e.g., magnetic resonance imaging (MRI)) medical image fusion. SPECT images reflect the metabolic information of the organism and can facilitate tumor detection and vascular disease diagnosis. However, the resolution is generally quite low. On the other hand, MRI images contain abundant anatomical information with high resolution, from which soft tissues can be clearly recognized. To conduct accurate diagnosis, physicians usually analyze SPECT and MRI images sequentially but this ordinal analysis method is inconvenient in most cases. To alleviate this limitation, the fusion of SPECT and MRI images acts as an effective method [7]. It aims to obtain a fused image that can simultaneously preserve the functional metabolic information of the SPECT image and the structural soft tissue information of the MRI image.

During the past decades, many multimodal medical image fusion methods have been proposed. These methods can be roughly divided into following categories: transform domain-based methods [9], sparse representative (SR)-based methods [10], hybrid methods [11], and other methods [12]. These methods generally require the manual design of feature extraction mechanisms and fusion strategies. To achieve better fusion performance, sophisticated schemes have made the fusion methods increasingly complicated and time-consuming. With powerful feature representation ability, deep learning (DL) has been widely applied to tackle computer vision (CV) tasks [13]–[15]. DL-based multimodal medical image fusion methods have also been proposed to treat the activity-level measurement and fusion rules as a whole to avoid hand-design. Although DL-based methods have achieved good fusion results, there still remain some defects: (i) Existing DL-based methods are generally based on convolutional operations, which can capture the local information from source images well but have limited ability in modeling long-range context dependencies. However, neuroscience studies illustrate that the global context capability of neurons is

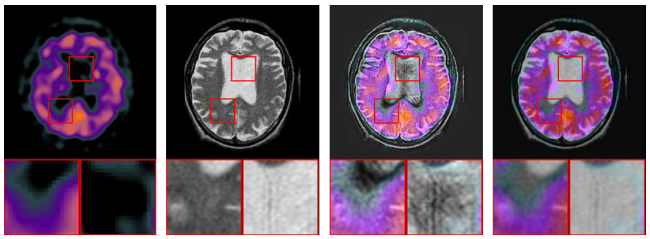

Fig. 1. Schematic illustration of multimodal medical image fusion. From left to right: SPECT image, MRI image, the fused images of the DDcGAN-based method [8] and the proposed method. Clearly, the proposed MATR can simultaneously maintain the functional metabolic information in the SPECT image and the abundant anatomical details in the MRI image.

the key component to effectively process intricate perceptual issues [16], [17]. (ii) Existing DL-based methods usually feed the source images into a single network without taking the interscale information into consideration. This may lead to the loss of some important information. (iii) The loss functions adopted in the existing DL-based methods are often based on pixel-level, which has low tolerance for noise. This can result in undesirable artifacts to some degree.

To overcome the above drawbacks, this paper proposes a novel multimodal medical image fusion method based on a multiscale adaptive Transformer named MATR. In the proposed method, adaptive convolution (AC) is introduced to implement convolutional operations instead of vanilla convolution. As a result, global context information can be extracted by adaptively modulating the convolution kernel. To further capture the global complementary information from the source images, an adaptive Transformer is employed to model the long-range dependencies. In order to fully utilize the interscale features, the network architecture is designed in a multiscale manner. An objective function consisting of a structural loss and a region mutual information (RMI) loss is designed to constrain information preservation at the structural-level and the feature-level, respectively. To demonstrate the effectiveness of the proposed method, Fig. 1 shows an example of a pair of SPECT and MRI images and the fusion results generated from the DDcGAN-based method [8] and the proposed MATR. For better comparison, two local regions are enlarged as close-ups in each image. It can be observed that both DDcGAN and MATR can achieve good fusion results, while MATR has more accurate fusion performance. Specifically, the proposed MATR can adequately preserve the functional information from the SPECT image and structural tissue details from the MRI image without introducing artifacts. The main contributions of this paper can be summarized as followings.

- To capture and exploit the global context information from source images effectively, we introduce adaptive convolution and adaptive Transformer to address the multimodal medical image fusion issue. Furthermore, the network structure is designed in a multiscale fashion to fully extract interscale complementary semantic information.
- To efficiently train the MATR, we design an objective function from the perspectives of both the structural-level and the feature-level. Specifically, a structural loss and an RMI loss are devised to train the proposed method in an unsupervised manner.
- Extensive experiments on a mainstream dataset demonstrate that the proposed MATR can simultaneously acquire functional metabolism information from SPECT images and abundant soft tissue details from MRI images. It outperforms other representative and state-of-the-art methods in qualitative and quantitative evaluations.
- We also extend the proposed method to deal with other biomedical functional and structural image fusion tasks. The resulting satisfactory performance reveals that the MATR has good generalization capability.

The remainder of this paper is arranged as follows. Section II introduces related works and motivations, including a short review of existing multimodal medical image fusion methods, a brief introduction of vision Transformers, and the motivations for this work. In Section III, the proposed method is described in detail. In Section IV, experimental implementation details, fusion results, ablation study results, and generalization to fuse other biomedical images are provided. Finally, this paper is concluded in Section V.

II. Related Studies and Motivations

We first provide a brief review of the existing multimodal medical image fusion methods in Section II-A, followed by the introduction of vision Transformers in Section II-B. Finally, the motivations of this work are clearly presented in Section II-C.

A. Multimodal Medical Image Fusion Methods

During the past decades, a plethora of algorithms have been proposed for multimodal medical image fusion. These methods can be broadly classified into two categories: conventional methods and DL-based methods. For conventional methods, multiscale transform (MST)-based methods play a representative role, e.g., pyramid transform [25], wavelet transform [26], curvelet transform [27], contourlet transform [28], shearlet transform [29], and other transform methods [30], [31]. These MST-based methods follow the "Decomposition-Fusion-Reconstruction" rule and therefore generally have two main drawbacks. First, different input modalities undergo the same transformation without considering their own specific features. In addition, the manually designed fusion strategies have limited capability in preserving all the complementary information from source images.

With the successful application of deep learning in the field of CV [32]–[34], DL-based studies have also emerged as an active topic in the field of multimodal medical image fusion. Liu *et al.* [18] first introduced a convolutional neural network (CNN) to deal with multimodal medical image fusion task. They employed a Siamese convolutional network to generate a weight map with which the pixel activity information can be integrated from two source images. Zhang *et al.* [19] proposed a fast unified image fusion network (PMGI) to solve several image fusion problems, including multimodal medical image fusion. In their method, a gradient path and an intensity

TABLE I
COMPARISON WITH STATE-OF-THE-ART IMAGE FUSION METHODS

Method	CNN [18]	PMGI [19]	U2Fusion [20]	DDcGAN [8]	EMFusion [21]	PPT Fusion [22]	TGFuse [23]	SwinFuse [24]	MATR
End-to-end	×	✓	✓	✓	✓	×	✓	×	✓
Convolutional Operation	✓	✓	✓	✓	✓	×	✓	×	✓
Transformer	×	×	×	×	×	✓	✓	✓	✓
Multiscale Architecture	×	×	×	×	×	×	×	×	✓
Structural-Level Loss	×	×	×	×	✓	×	✓	✓	✓
Feature-Level Loss	×	×	×	×	×	×	×	×	✓
Unsupervised	×	✓	✓	✓	✓	×	✓	×	✓
SPECT-MRI	✓	×	×	×	✓	×	×	×	✓
Generalization Ability	×	×	×	×	×	×	×	✓	✓

path are devised for information extraction to realize proportional maintenance of the gradient and intensity. Xu *et al.* [20] also presented a unified image fusion network (U2Fusion) to realize different fusion tasks, which can adaptively estimate the information preservation degree from source images. In [8], Ma *et al.* put forward a dual-discriminator conditional generative adversarial network (DDcGAN) to address the multiresolution image fusion issue. They established an adversarial game between a generator and two discriminators. Xu and Ma [21] proposed an enhanced medical image fusion network (EMFusion) in which surface-level and deep-level constraints are enforced to enhance information preservation. Although existing DL-based multimodal medical image fusion methods have achieved good performance, several limitations still exist that affect the fusion results. First, existing DL-based methods are generally based on convolutional operations, which have a powerful ability to model local patterns but fail to extract global semantic information, resulting in inaccurate and noised fusion results to some extent. In addition, existing DL-based methods usually employ a singlescale network to capture information without taking interscale features into consideration. However, the human visual system (HVS) perceives information in a multiresolution manner [35]. Therefore, useful information may be lost with the singlescale architecture. Moreover, existing DL-based algorithms normally construct loss functions at the pixel-level, which causes information distortion and hampers the fusion performance.

To solve the challenges mentioned above, we propose a novel multimodal medical image fusion method with a multiscale adaptive Transformer called MATR. In the proposed method, an unsupervised multiscale image fusion network is designed to fuse multiple inputs, which cannot only achieve the fusion process in an end-to-end fashion without elaborate decomposition approaches and fusion schemes but also adequately preserve interscale features from source images. To fully extract global context information, adaptive convolution and adaptive Transformer are introduced to model the long-range dependencies so that more accurate fusion results can be generated. Furthermore, we devise an objective function based on the structural-level and feature-level aspects to realize comprehensive information preservation constraints.

B. Vision Transformers

Inspired by the successful application of Transformer architectures to the field of natural language processing (NLP) [36], researchers have applied Transformer to address CV tasks [37]. Dosovitskiy *et al.* [38] proposed a vision Transformer model called ViT. They adopted a Transformer encoder for image classification by splitting images into patches with a fixed size. Carion *et al.* [39] presented DERT to treat object detection as a direct set prediction problem in which the relations of the objects and the global image context are reasoned and the final set of predictions are obtained directly. Chen *et al.* [40] introduced TransUNet for medical image segmentation. In their method, Transformer is used to encode the global context by arranging the image features as sequences and a U-shaped CNN is utilized to extract low-level features. In [41], Chen *et al.* proposed an image processing Transformer (IPT), which is composed of multiple heads, multiple tails, and a shared Transformer body, to address low-level computer vision tasks, e.g., denoising, super-resolution, and deraining.

More recently, Transformer-based algorithms have also been proposed for image fusion task. Fu *et al.* [22] proposed a patch Pyramid Transformer (PPT) to transform the image into a sequence of patches with a patch Transformer and utilized a Pyramid Transformer to extract the nonlocal information from the entire image. In 2022, Rao *et al.* [23] presented an infrared and visible image fusion approach by combining Transformer and GAN. In their method, Transformer is used to learn the global fusion relations, and two discriminators are adopted in the training process to improve the output discrimination.

To reduce computational costs, Liu *et al.* [42] presented a hierarchical Transformer (i.e., Swin Transformer) in which shifted windows are adopted to compute the representation. By limiting self-attention computation to non-overlapping local windows and allowing the cross-window connection, the Swin Transformer achieves greater efficiency and flexibility. Motivated by [42], Liang *et al.* [43] put forward a SwinIR for image restoration based on the Swin Transformer, in which several residual Swin Transformer blocks (RSTBs) are designed as the deep feature extraction module, and each RSTB has several Swin Transformer layers together with a residual connection. Following [43], Wang *et al.* [24] developed a pure Transformer network by employing RSTB to fuse infrared and visible images called SwinFuse. With the L_1-norm feature fusion strategy devised in [24], SwinFuse cannot be trained in an end-to-end fashion.

In this paper, we also employ an adaptive Transformer with a shifted window mechanism for long-term relationship construction to adequately extract global complementary

information from the source images. Inspired by [44], we propose an unsupervised and end-to-end fusion model through combining the advantages of both CNN and Transformer to fully extract the local and global complementary features.

C. Motivations

The aim of SPECT and MRI image fusion is to generate a composite image that simultaneously exhibits the functional information reflecting metabolism and provides abundant soft tissue structural details. To achieve this goal and to obtain fusion results directly without using intricately designed fusion strategies, we design an end-to-end fusion model that combines the merits of both CNN and Transformer to adequately extract the local and global complementary information. Specifically, to adaptively represent features, we devise the adaptive convolution (AC) that can adaptively modulate the convolution kernel. To further improve the receptive field, we employ an adaptive Transformer module (ATM) for long-range dependency modeling. In addition, to preserve significant features through different scales, the proposed network is designed in a multiscale fashion with different depths. Furthermore, to better train the proposed MATR in an unsupervised manner, a structural-level loss and a feature-level loss are designed.

A comprehensive comparison with the state-of-the-art image fusion methods from different aspects is listed in Table I. a) End-to-end. One of the main differences between DL-based image fusion methods and conventional methods is that DL-based methods can be designed in an end-to-end model so that fusion results can be directly generated without elaborately designed fusion strategies. b) Convolutional operation. As the convolutional operation has promising local feature representation ability, it has been widely used in image fusion algorithms. c) Transformer. Since the CNN has limited global feature representation capability, the Transformer equipped model can remedy this deficiency. d) Multiscale architecture. A single network can produce good fusion results, but useful information that exists at different depths may be lost during the training process. Therefore, multiscale structures with different depths have better feature extraction ability. e) Structural-level loss. Pixel-level loss may introduce undesirable noise to fusion results. The structural-level loss can avoid this defect and can guide the network to obtain more adequate structural details. f) Feature-level loss. With only structural-level loss, the fused image may fall into inaccurate performance. To avoid this, the feature-level loss can be used to further ensure that the fusion result has accurate information preservation. g) Unsupervised. Due to the absence of ground-truth for multi-modal medical image fusion, some fusion approaches utilize pre-trained networks to extract image features. In these approaches, the extracted features cannot well meet the requirements of the image fusion issue. Hence, an unsupervised training strategy is more suitable for medical image fusion. h) SPECT-MRI. The most recently presented Transformer-based image fusion methods are not designed for SPECT-MRI image fusion task. This is however an important field and can facilitate subsequent medical research and application [3]–[6]. i) Generalization ability. The generalization ability to other image fusion tasks without fine-tuning plays a crucial role in practical applications. However, most of the existing image fusion methods are either designed for specific image fusion tasks or require some parameters or fusion rules to be adjusted to fit other types of image fusion issues. Both of these kinds of models lack generalization capability.

Taking the above analysis and comparison into consideration, in this paper, we design a SPECT and MRI image fusion method via a multiscale adaptive Transformer termed MATR. Specifically, we merge the advantages of both CNN and Transformer by designing the adaptive convolution with accommodative convolution kernel and adaptive Transformer for global feature extraction. In addition, the proposed model is devised in a multiscale manner to fully capture the complementary information from diverse scales with various depths. Furthermore, a structural-level loss and a feature-level loss are designed to better guide unsupervised training. To demonstrate the generalization capability of the MATR, extension experiments to other functional and structural image fusion tasks are implemented without fine-tuning.

III. Proposed Method

In this section, the proposed MATR is introduced in detail. The overall framework is first described in Section III-A. Then, the novel network structure is illustrated in Section III-B. Finally, we formulate the proposed objective function in Section III-C.

A. Framework Overview

The network architecture of the proposed multimodal medical image fusion method is exhibited in Fig. 2. Considering that the SPECT image I_{SPE} is the RGB color image while the MRI image I_{MRI} is the single-channel greyscale image, RGB-to-YUV color conversion is utilized to solve the channel mismatching problem. It is worth mentioning that RGB-to-YUV color conversion is a widely used method to address the three-channel RGB functional and single-channel gray structural biomedical image fusion issue [45]–[48]. Specifically, the SPECT image I_{SPE} is first converted into YUV color space to obtain the Y, U, and V components: I^Y_{SPE}, I^U_{SPE}, and I^V_{SPE}. Afterward, I^Y_{SPE} and I_{MRI} are concatenated in the channel dimension and fed into the proposed multiscale adaptive Transformer network. As the proposed MATR is an end-to-end model, the design of fusion strategies can be avoided [21], [49], [50]. The output of the network is the fused Y component I^Y_F and the final fusion result I_F is generated by implementing YUV-to-RGB color conversion among I^Y_F, I^U_{SPE} and I^V_{SPE}.

B. Network Architecture

As shown in Fig. 2, the concatenated I^Y_{SPE} and I_{MRI} are fed into a basic module (BM), which is composed of an AC, a batch normalization (BN) layer, and a rectified linear unit (ReLU), which can be formulated as

$$F^{Out}_{BM} = ReLU\left(BN\left(AC\left(F^{In}_{BM}\right)\right)\right), \tag{1}$$

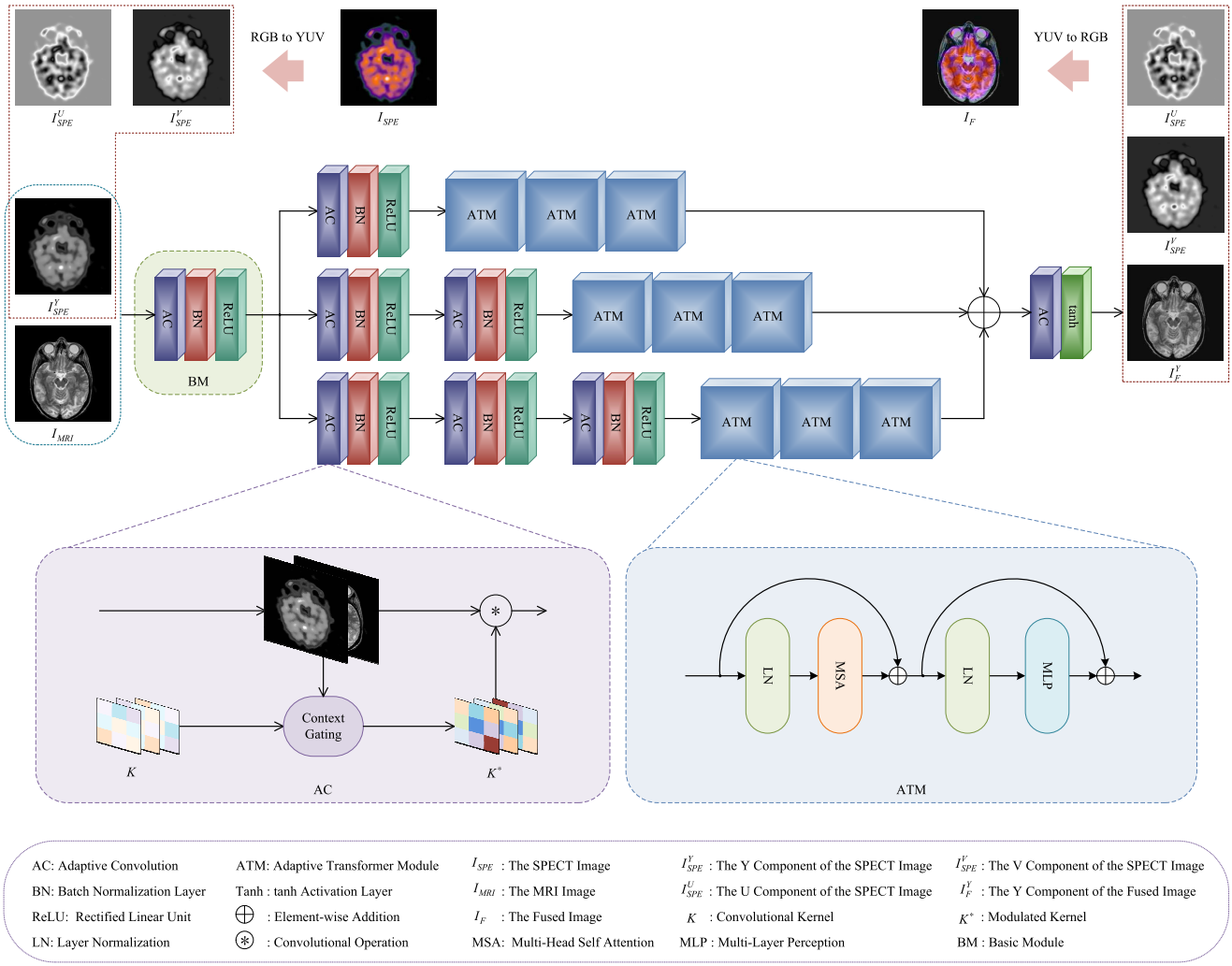

Fig. 2. The framework of the proposed MATR. The SPECT image I_{SPE} is converted to YUV color space by RGB-to-YUV color conversion firstly, and then the Y component I^{Y}_{SPE} and the MRI image I_{MRI} are concatenated in the channel dimension and fed into the multi-scale adaptive Transformer network to obtain the fused Y component I^{Y}_{F}. Finally, the fusion result I_F is generated through YUV-to-RGB conversion among I^{U}_{SPE}, I^{V}_{SPE}, and I^{Y}_{F}.

where F^{In}_{BM} and F^{Out}_{BM} represent the input and output features of the BM, respectively. It is worth mentioning that the aim of multimodal medical image fusion is to obtain a composite image that can simultaneously contain the functional metabolic information in the SPECT image and the structural soft tissue details in the MRI image. Therefore, the essential element that affects fusion performance is whether the significant complementary information can be adequately preserved from the source images. As neuroscience studies have reported [16], [17], global context information plays a critical role in interpreting visual scenes. Based on this consideration and motivated by [51], AC, instead of vanilla convolution, is introduced to extract features. Unlike former feature map-based modification, the AC introduced in this paper modulates the convolutional kernel. As shown in Fig. 3, compared with vanilla convolution, the modified AC can directly modulate the convolution kernel to adaptively represent features under the guidance of the global context information.

As the HVS processes information in a multiresolution manner, to generate high-quality fusion results, the proposed model is designed in a multiscale manner for multimodal interscale complementary feature extraction. Specifically, three branches are arranged in a multiscale fashion to handle the output of the BM for better preservation of the interscale information. The top branch consists of a BM and three adaptive Transformer modules (ATMs) to represent latent features. The middle branch has two BMs and three ATMs and the bottom branch owns three BMs and three ATMs. The branch with more BMs can extract deeper features, and therefore, possesses more powerful information extraction capability. The multiscale structure with different depths facilitates the generation of more informative fusion results.

As shown in the blue rounded rectangle in Fig.2, the ATM has two addition operations. The first addition operation can be formulated as

$$F^{Out}_{ATM1} = MSA\left(LN\left(F^{In}_{ATM}\right)\right) + F^{In}_{ATM}, \tag{2}$$

where F^{In}_{ATM} and F^{Out}_{ATM1} denote the input of the ATM and the first addition operation result, respectively. LN means

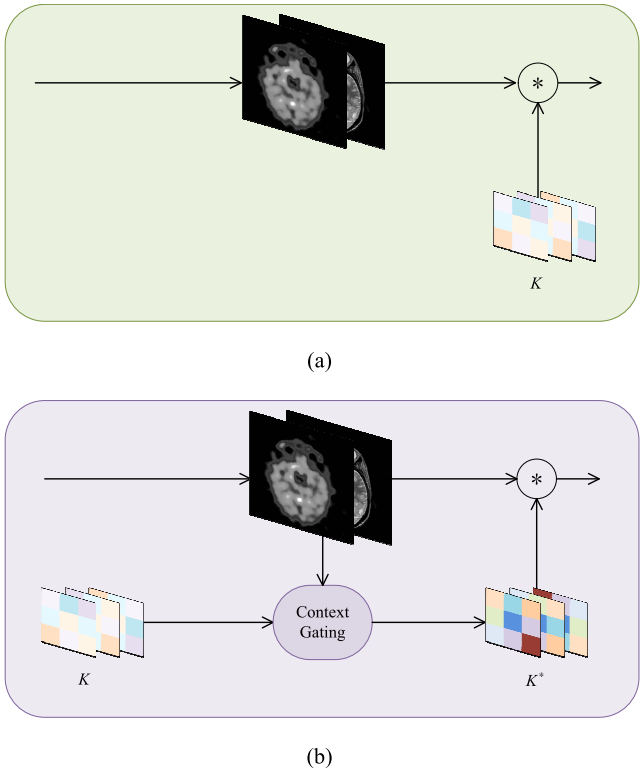

Fig. 3. Comparison of the vanilla convolution with our modified adaptive convolution. (a) Vanilla Convolution. (b) Adaptive Convolution. ⊛ represents convolutional operation.

layer normalization and MSA is multihead self-attention. The second addition operation can be represented as

$$F_{ATM}^{Out} = MLP\left(LN\left(F_{ATM1}^{Out}\right)\right) + F_{ATM1}^{Out}, \tag{3}$$

where F_{ATM}^{Out} stands for the output of the ATM and MLP indicates multilayer perception.

The outputs of these three branches are added together to aggregate the features extracted by individual pathways and then fed into an AC followed by a tanh activation function to obtain I_F^Y.

C. Loss Function

The absence of ground-truth for multimodal medical image fusion presents a challenge for the training process. Since the MATR is proposed to generate an informative fused medical image that contains abundant structural information from the MRI image and accurate functional information from the SPECT image, the loss function is devised from both the structural-level and region-level standpoints to train the proposed network. The total loss function of our MATR is defined as follows:

$$L_{\text{total}} = L_{SSIM} + L_{RMI}, \tag{4}$$

where L_{SSIM} and L_{RMI} are the structural-level loss and the region-level loss, respectively.

L_{SSIM} is introduced for the structural-level constraint to ensure that the fusion result has adequate structural details. Structural similarity index measurement (SSIM) [52] is utilized to constrain the similarity between the fused image and the source images, which is defined as:

$$L_{SSIM} = 1 - SSIM\left(I_F^Y, I_{SPE}^Y\right) + \alpha \cdot \left(1 - SSIM\left(I_F^Y, I_{MRI}\right)\right), \tag{5}$$

where α is a trade-off factor that controls the balance between two terms. Let S represent the source image, i.e., I_{SPE}^Y or I_{MRI}. Then, $SSIM$ can be expressed as:

$$SSIM(S, I_F^Y) = \frac{\left(2\mu_S\mu_{I_F^Y} + C_1\right)\left(2\sigma_{SI_F^Y} + C_2\right)}{\left(\mu_S^2 + \mu_{I_F^Y}^2 + C_1\right)\left(\sigma_S^2 + \sigma_{I_F^Y}^2 + C_2\right)}, \tag{6}$$

where μ_S and $\mu_{I_F^Y}$ represent the mean intensities of S and I_F^Y, receptively. C_1 and C_2 are constants designed to avoid instability when $\mu_S^2 + \mu_{I_F^Y}^2$ or $\sigma_S^2 + \sigma_{I_F^Y}^2$ is very close to zero. σ_S^2 and $\sigma_{I_F^Y}^2$ stand for the variance of S and I_F^Y, respectively. $\sigma_{SI_F^Y}$ denotes the covariance of S and I_F^Y.

L_{RMI} is presented to further ensure that the fusion result has accurate information preservation. It is designed from the perspective of region-level restriction in which region mutual information (RMI) [53] is employed to restrict the information from the source images to be transferred to the fused image from the perspective of not pixel-to-pixel but region-to-region to avoid undesirable artifacts. L_{RMI} is formulated as:

$$L_{RMI} = \beta \cdot RMI\left(I_F^Y, I_{SPE}^Y\right) + \gamma \cdot RMI\left(I_F^Y, I_{MRI}\right), \tag{7}$$

where β and γ are two hyperparameters controlling the trade-off of these two items, and RMI is defined as follows:

$$RMI(S, I_F^Y) = \lambda\mathcal{L}_{ce}(S, I_F^Y) + (1-\lambda)\frac{1}{B}\sum_{b=1}^{B}\left(-I_l^b(S; I_F^Y)\right) \tag{8}$$

where $\lambda \in [0, 1]$ is a weight parameter. $\mathcal{L}_{ce}(S, I_F^Y)$ represents the normal cross entropy loss between S and I_F^Y. B represents batch size, and $I_l(S; I_F^Y)$ stands for the lower bound of the mutual information.

IV. Experiments

In this section, the dataset employed in this work and training details are introduced in Section IV-A. Then, in Section IV-B, the algorithms employed for the comparison and objective assessment indexes are presented. To demonstrate the effectiveness of the proposed MATR, qualitative and quantitative comparisons with representative and state-of-the-art methods and the corresponding analysis are given in Section IV-C. In Section IV-D, ablation studies on the network structure, objective function, and trade-off parameters in the loss function are described. Finally, to verify the generalization capability of the proposed method, it is extended to address PET and MRI image fusion task in Section IV-E, and GFP and PC image fusion task in Section IV-F.

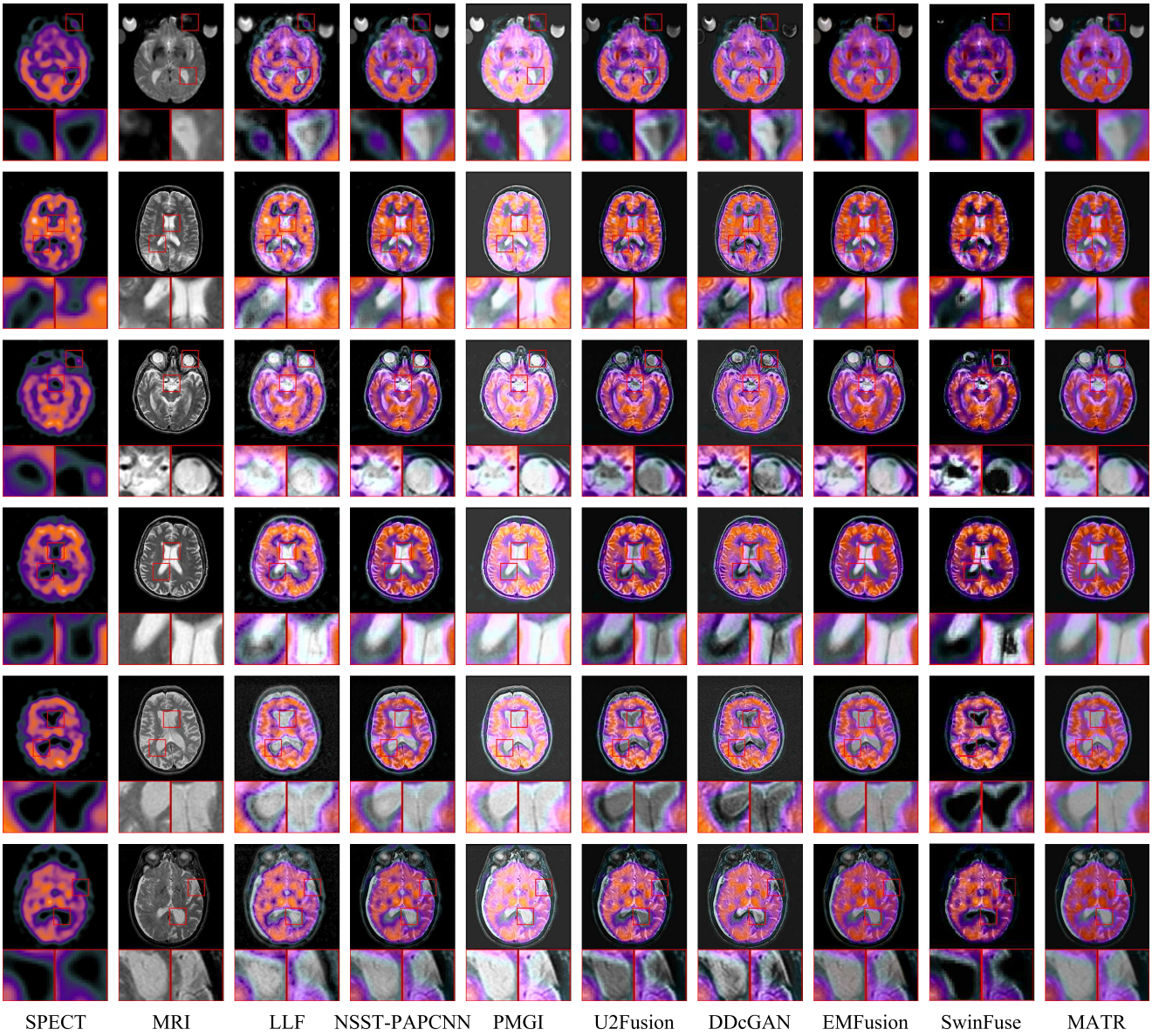

Fig. 4. Qualitative comparison of the proposed MATR with 7 typical and state-of-the-art methods on 6 representative SPECT and MRI image pairs. From left to right: SPECT image, MRI image, fusion results of LLF [7], NSST-PAPCNN [45], PMGI [19], U2Fusion [20], DDcGAN [8], EMFusion [21], SwinFuse [24], and the proposed MATR. For better comparison, two local regions are enlarged as close-ups in each image.

A. Dataset and Training Details

In this work, 354 pairs of SPECT and MRI images with 256×256 pixels are downloaded from the mainstream medical image database, namely, Harvard.[1] It is worth mentioning that Harvard is a widely used dataset for multi-modal medical image fusion [7], [8], [20]. It is not only suitable for fair comparison but it can also enhance the engineering application value of this work. As registration is a prerequisite for image fusion, all image pairs are assumed to be accurately aligned. The downloaded image pairs are randomly divided into a training set, a validation set, and a testing set. Specifically, 319 image pairs are selected as training samples. To obtain sufficient training data, the overlapping cropping strategy is employed for data augmentation, which is a widely used method in the field of image fusion [8], [20], [54]. All the images in the training set are cropped into patch pairs of size 120×120 with a crop size of 20. By doing so, 15631 pairs of patches are obtained for network training, and all of these samples are normalized to $[0, 1]$. Fifteen and twenty pairs of SPECT and MRI images are chosen as the validation set and testing set, respectively. As the cropping strategy is adopted for data augmentation, it is not employed for validation or testing, so that the entire image can be directly fed into the trained model to generate fusion results.

The experiments are implemented via the PyTorch framework on an NVIDIA GeForce RTX 3090 GPU. During the training process, the Adam optimizer is utilized and the learning rate is specified as 0.001. The batch size is 64, and the number of epochs is fixed as 10. The hyperparameters in the loss function are set as $\alpha = 1$, $\beta = 1$, and $\gamma = 2.5$, which

[1] http://www.med.harvard.edu/AANLIB/home.html

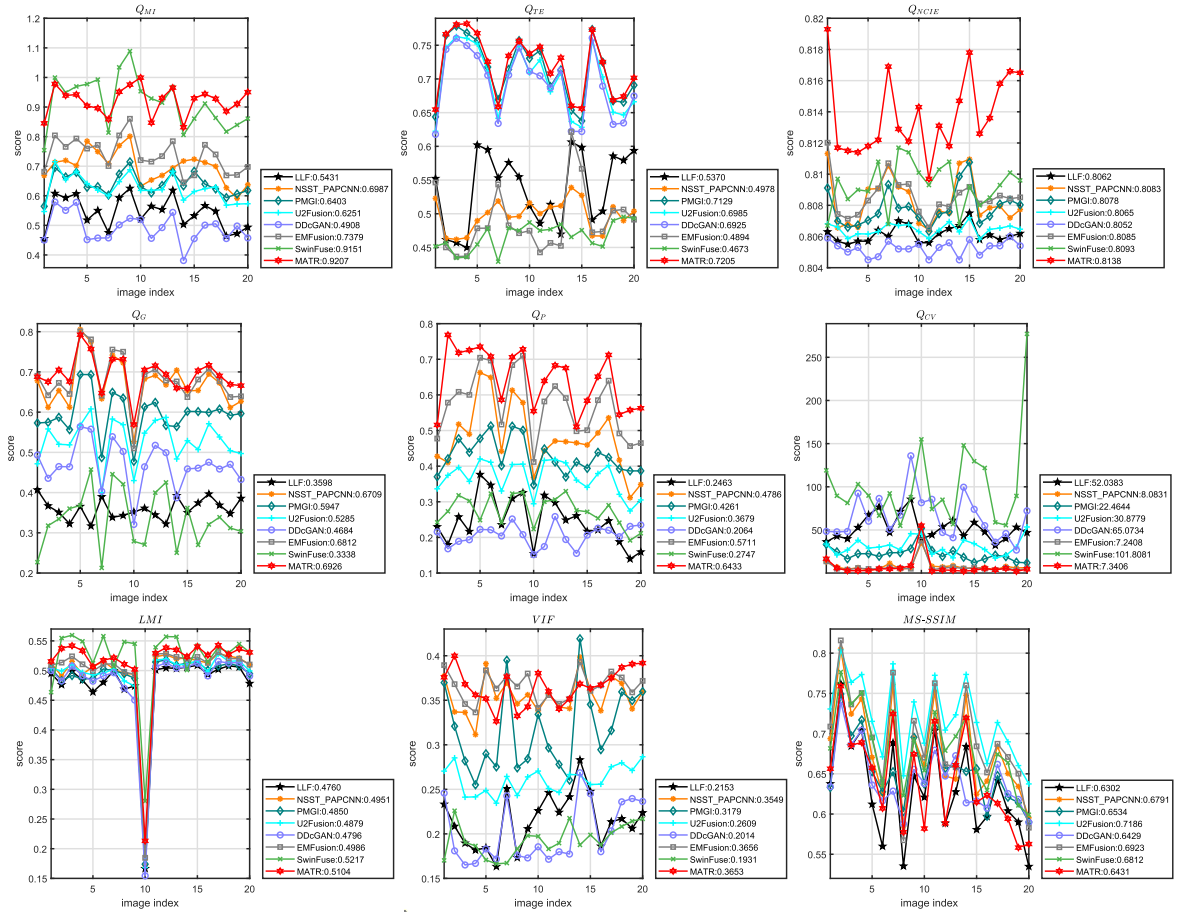

Fig. 5. Quantitative comparison of the proposed MATR for SPECT and MRI image fusion with 7 representative and state-of-the-art methods on nine objective evaluation metrics. Average scores for different methods are illustrated in the legends.

will be discussed thoroughly in Section IV-D. λ is fixed as 0.5, as suggested in [53]. The kernel sizes are all set as 3×3 with the padding operation except the last one with the kernel size of 1×1.

B. Comparison Methods and Evaluation Metrics

Seven representative and state-of-the-art methods are adopted for qualitative and quantitative comparisons: the local Laplacian filtering (LLF)-based method [7], the parameter-adaptive pulse coupled neural network in non-subsampled shearlet transform domain (NSST-PAPCNN)-based method [45], the PMGI-based method [19], the U2Fusion-based method [20], the DDcGAN-based method [8], the EMFusion-based method [21], and the SwinFuse-based method [24]. Among them, LLF and NSST-PAPCNN are two representative multimodal medical image fusion methods. PMGI, U2Fusion, DDcGAN, and EMFusion are four recently proposed DL-based methods. SwinFuse is the latest proposed Transformer-based image fusion approach. The codes of these seven methods are all publicly available or are provided by their authors. All the parameters are fixed as recommended in the corresponding publications to ensure unbiased comparison.

Nine widely used evaluation metrics are employed to implement a comprehensive objective assessment, including normalized mutual information Q_{MI} [55], Tsallis entropy Q_{TE} [56], nonlinear correlation information entropy Q_{NCIE} [57], gradient-based metric Q_G [58], image feature-based metric using phase congruency Q_P [59], Chen-Varshney metric Q_{CV} [60], localized mutual information (LMI) [61], visual information fidelity (VIF) [62], and multi-scale structural similarity index ($MS\text{-}SSIM$) [63]. Q_{MI} estimates the mutual information that the fused image extracted from the source images. Q_{TE} measures the dependence degree between the input and the fused images, which is a divergence measurement. Q_{NCIE} measures the nonlinear correlation information entropy preserved in the fused image. Q_G calculates the gradient information that is transformed from the source image to the fusion result. Q_P reveals the fusion quality based on the feature map measurement of the local correlation between the fused and input images. Q_{CV} is an image quality index inspired by human perception. The LMI computes the joint information between the source and the fused images. VIF quantifies the information fidelity between the input images and the fusion result. $MS\text{-}SSIM$ assesses the structural similarity between the source images and the fusion image. For Q_{CV}, a higher score indicates worse fusion performance, while for the other eight evaluation metrics, the higher the value is, the better the fusion result.

C. Results and Discussion

Fig. 4 reveals six typical SPECT and MRI image pairs and their corresponding fusion results obtained by different

methods. To better report an intuitive comparison, two local areas are enlarged as close shots. All eight methods can achieve relatively satisfactory fusion performance. An interesting observation is that when compared with the conventional methods, the DL-based methods tend to be slightly oversmoothed. The reason for this may be that the traditional methods suffer from checkerboard artifacts to some degree. In contrast, the fusion results of the DL-based approaches seem to be oversmoothed. On the whole, when compared with the proposed MATR, the other seven algorithms have several drawbacks. Specifically, the LLF-based method over-preserves the functional information from the SPECT image resulting in artifacts in the fused images. The NSST-PAPCNN-based method has better performance in terms of alleviating artifacts but still suffers from it to some extent. The PMGI-based method can well extract the structural information from the MRI image while having limited ability in capturing the metabolism information from the SPECT image, which leads to some color distortion. The U2Fusion-based and DDcGAN-based methods have good preservation capability in capturing the functional information from the SPECT image while introducing undesirable noises (e.g., the white matter is blurred). The EMFuion-based method illustrates better soft tissue maintaining capability while undergoing the problem of slight color distortion. The SwinFuse-based method can well exploit the functional information of the SPECT image, but some important structural details of the MRI image are lost. Overall, the proposed MATR reveals a better fusion property in retaining the complementary information from the source images.

Quantitative comparisons on the testing set with the nine widely used evaluation metrics are shown in Fig. 5. For each metric, the assessment scores of 20 testing samples are connected in a broken line and the average value for each method is labeled in the legend. It is clear that for Q_{NCIE} and Q_P, the MATR obtains the top marks on all 20 images. For Q_{MI}, Q_{TE}, Q_G, and VIF, higher scores are generally realized than the competitors with the highest average values. For LMI and $MS\text{-}SSIM$, the proposed method has a relatively lower score when compared with the most recently proposed algorithms, but the difference with other algorithms is not huge. For Q_{CV}, the proposed method can always reach the bottom goal, which indicates the best fusion performance. Based on the above analysis, our proposed MATR reports better objective performance than other representative and state-of-the-art methods.

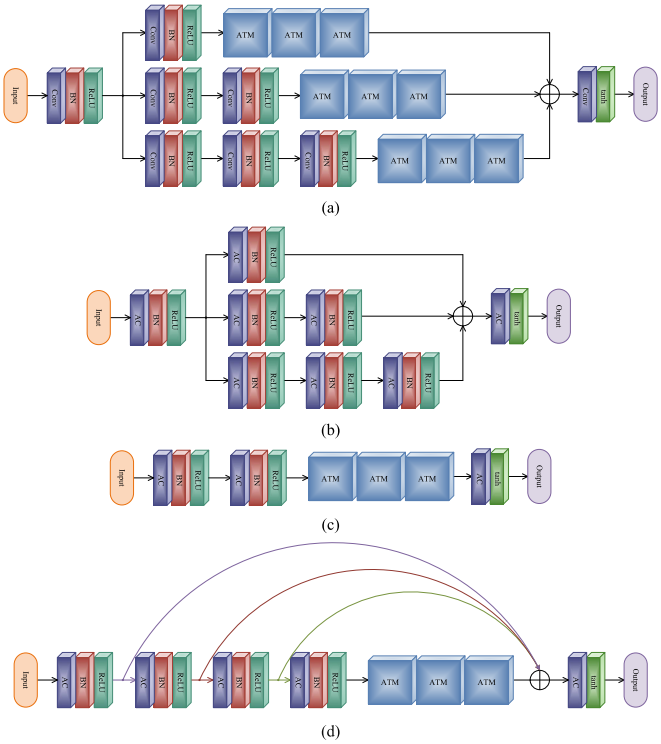

Fig. 6. Four different network structures of the proposed method for ablation study. (a) Vanilla Conv (i.e., all the adaptive convolution are replaced by the vanilla convolution). (b) Without ATM (viz., the ATM is removed from the proposed network architecture). (c) Single Scale (i.e., only one path is maintained from the proposed network structure). (d) Multi-Stage (viz., extracting features through different stages in the longest path).

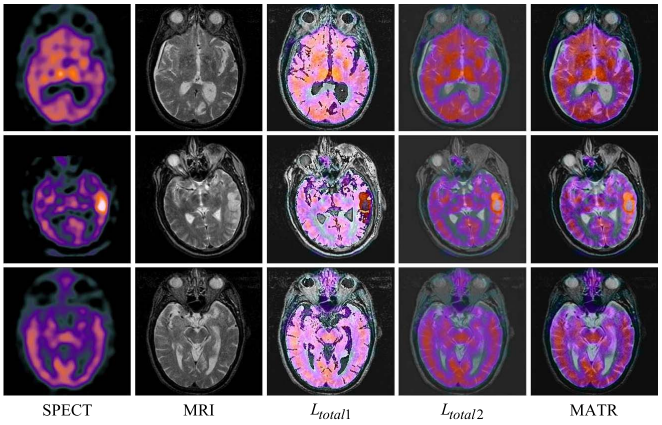

Fig. 7. Three sets of source images and their corresponding fusion results of the proposed method with different loss functions. From left to right: SPECT image, MRI image, fusion results of L_{total1} (i.e., the loss function is composed of region-level loss), L_{total2} (e.g., the loss function is constituted of structural-level loss), and the proposed MATR.

D. Ablation Study

1) Analysis of Network Structure: To verify the effectiveness of each component in the proposed method, namely, AC, ATM, and multiscale structure, an ablation study on the validation set is conducted. As shown in Fig. 6 (a), we replace all the AC with vanilla convolution to demonstrate the superiority of the adaptive convolution operation employed in our method. In Fig. 6 (b), ATM is deleted from the full architecture to illustrate its global semantic representation capability. In Fig.6 (c), only a single scale is employed for network training to investigate the availability of the multi-scale framework. To further illustrate the proposed multiscale structure, especially three parallel paths with different depths, as exhibited in Fig. 6 (d), we extract features through different stages in the longest path. The quantitative evaluation results of the MATR with different configurations are provided in Table II. For each metric, the method with the best result is labeled in bold. It is obvious that the intact model largely outperforms the other degraded models. When compared with

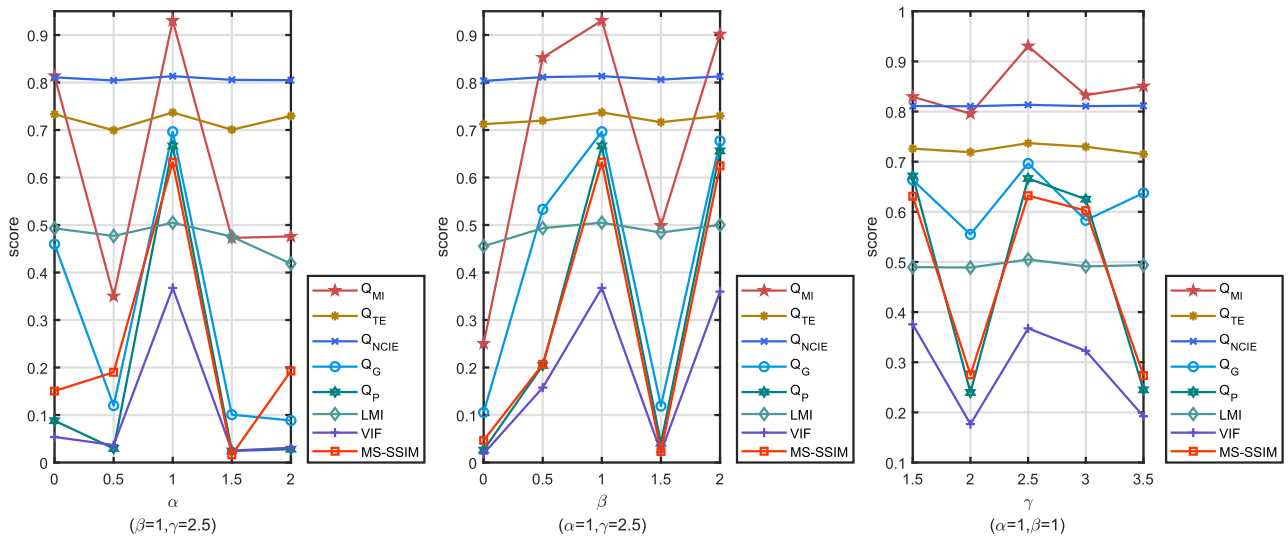

Fig. 8. Impacts on the trade-off parameters α, β and γ on quantitative performance.

Vanilla Conv, the higher scores on MATR demonstrate that the adaptive convolution has better feature extraction capability. The higher values in comparison to Without ATM illustrate that our ATM possesses good global semantic information preservation ability. The better performance than the Single Scale and Multi-Stage reveals that our proposed multiscale architecture can well exploit multimodal interscale complementary information, and facilitate the generation of more informative fusion results.

2) Analysis of the Loss Function: Since the proposed MATR is an unsupervised model, the loss function plays an important role in determining the fusion result. Taking this into consideration, we design a structural-level loss and a region-level loss to better guide the training process. To investigate the significance of these two losses, ablation experiments on the validation set are performed. Specifically, we cut off the structural-level loss to demonstrate its necessity. To this end, the loss function is reformulated as

$$L_{\text{total1}} = L_{RMI}. \tag{9}$$

Similarly, region-level loss is removed from the loss function to verify its significance, and the new loss function is defined as

$$L_{\text{total2}} = L_{SSIM}. \tag{10}$$

Qualitative comparisons of the proposed method with different loss functions is shown in Fig. 7. It is obvious that both the results of L_{total1} and L_{total2} suffer from color distortions severely. To be more specific, L_{total1} tends to lose structural details resulting in unreal fusion results. L_{total2} exhibits lower contrast and blurred fused images. In comparison, equipped with the two losses, our MATR has more accurate and natural fusion performance. Therefore, both structural-level loss and region-level loss are crucial for training the proposed model.

3) Analysis of Parameter Settings: To inquire into the trade-off parameters in the loss function (i.e., α, β and γ), we implement extensive experiments on the validation set and finally fix them as $\alpha = 1$, $\beta = 1$ and $\gamma = 2.5$. Since there exist too many combinations of these three factors, it is not appropriate and necessary to exhibit all of them. Therefore, one group of results is provided and the "control variate method" is employed. Specifically, to demonstrate the impact of one specific weight parameter, the other two factors are specialized as the default values. The quantitative results are shown in Fig. 8. For each quantitative assessment index, the average score of all the samples in the validation set is reported. It is worth mentioning that the numerical transformation range of Q_{CV} is much larger than those of the other eight indicators (similar observation can be illustrated in Fig. 5), and the overall trend is the same as that of the other eight indicators. Therefore, for a better comparison, Q_{CV} is not shown in Fig. 8. Visibly, when $\alpha = 1$, $\beta = 1$, and $\gamma = 2.5$, the best fusion performance is achieved on the validation set. Therefore, we set the default values of α, β, and γ as 1, 1, and 2.5, respectively.

E. Extension to PET and MRI Image Fusion

To investigate the generalization ability of the proposed method on other types of multi-modal medical images, we extend MATR to address positron emission tomography (PET) and MRI image fusion task without fine-tuning. PET images reflect metabolic information and highlight tumor areas with low spatial resolution. In contrast, MRI images provide high-resolution soft tissue information. Thus, the fusion of PET and MRI images can generate a composite image with abundant structural details and salient nidus, which is conducive to clinical analysis and treatment planning [5], [6]. We download 10 pairs of PET and MRI images from the mainstream medical database[2] as testing samples for generalization experiments. Six state-of-the-art algorithms are adopted for comparison, including the LLF-based method [7], the NSST-PAPCNN-based method [45], the

[2]http://www.med.harvard.edu/AANLIB/home.html

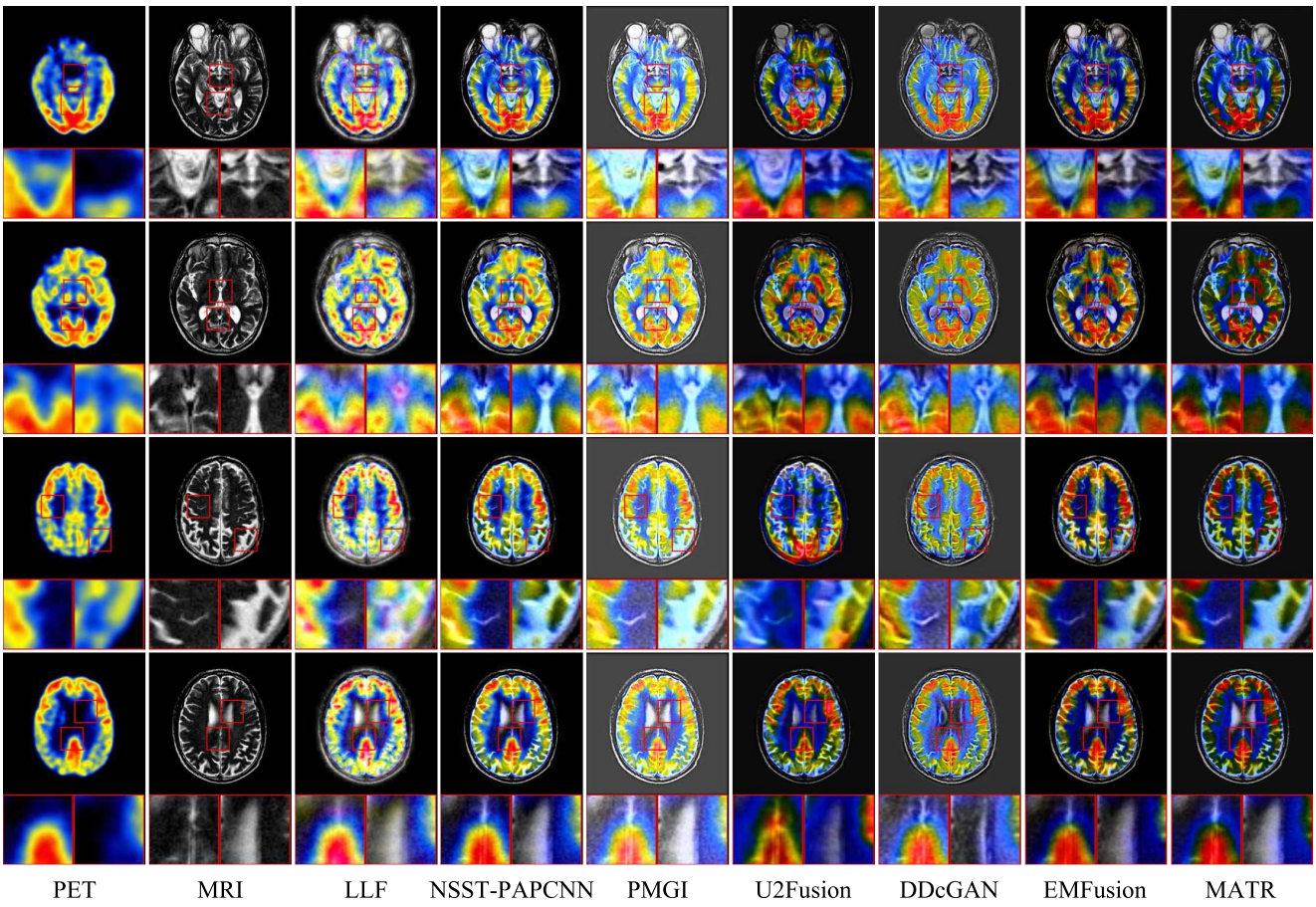

Fig. 9. Qualitative comparison of the proposed MATR with 6 typical and state-of-the-art methods on 4 representative PET and MRI image pairs. From left to right: PET image, MRI image, fusion results of LLF [7], NSST-PAPCNN [45], PMGI [19], U2Fusion [20], DDcGAN [8], EMFusion [21], and the proposed MATR. For better comparison, two local regions are enlarged as close-ups in each image.

TABLE II
QUANTITATIVE ASSESSMENT RESULTS OF THE PROPOSED METHOD WITH DIFFERENT NETWORK STRUCTURES

Metric	Vanilla Conv	Without ATM	Single Scale	Multi-Stage	MATR
Q_{MI}	0.4007	0.8277	0.5719	0.7675	**0.9304**
Q_{TE}	0.6691	0.7181	0.6907	0.7243	**0.7369**
Q_{NCIE}	0.8042	0.8102	0.8065	0.8105	**0.8134**
Q_G	0.1005	0.5210	0.3746	0.5645	**0.6963**
Q_P	0.0203	0.1664	0.0679	0.2056	**0.6664**
Q_{CV}	444.8358	524.7514	450.8177	432.1988	**6.3854**
LMI	0.3963	0.4942	0.4855	0.4849	**0.5049**
VIF	0.0157	0.1350	0.0810	0.1591	**0.3675**
$MS\text{-}SSIM$	0.0782	0.1611	0.1129	0.1803	**0.6322**

PMGI-based method [19], the U2Fusion-based method [20], the DDcGAN-based method [8], and the EMFusion-based method [21]. The codes of these six approaches are all publicly available, and we set the parameters by referring to their original papers.

Four sets of representative PET and MRI image pairs and their corresponding fused images obtained by different approaches are illustrated in Fig. 9. All seven methods exhibit good fusion performance. However, when compared with the proposed MATR, the six competitors suffer from some defects to a certain degree. Specifically, the LLF-based method cannot preserve the complementary features adequately resulting in blurred fusion results. The NSST-PAPCNN-based method can extract the functional information from the PET image well, but some important structural details are lost. The PMGI-based and DDcGAN-based methods have lower contrast, leading to gray backgrounds. The U2Fusion-based method achieves relatively better fusion performance but color distortion occurs (see the last row in Fig. 9). The EMFusion-based method alleviates this problem, but the structural information of MRI is lost to some extent. Overall, our proposed MATR illustrates the best fusion performance in terms of simultaneously exploiting functional information from the PET image and tissue details of the MRI image.

Fig. 10 shows the objective comparison of the proposed method for PET and MRI image fusion with six state-of-

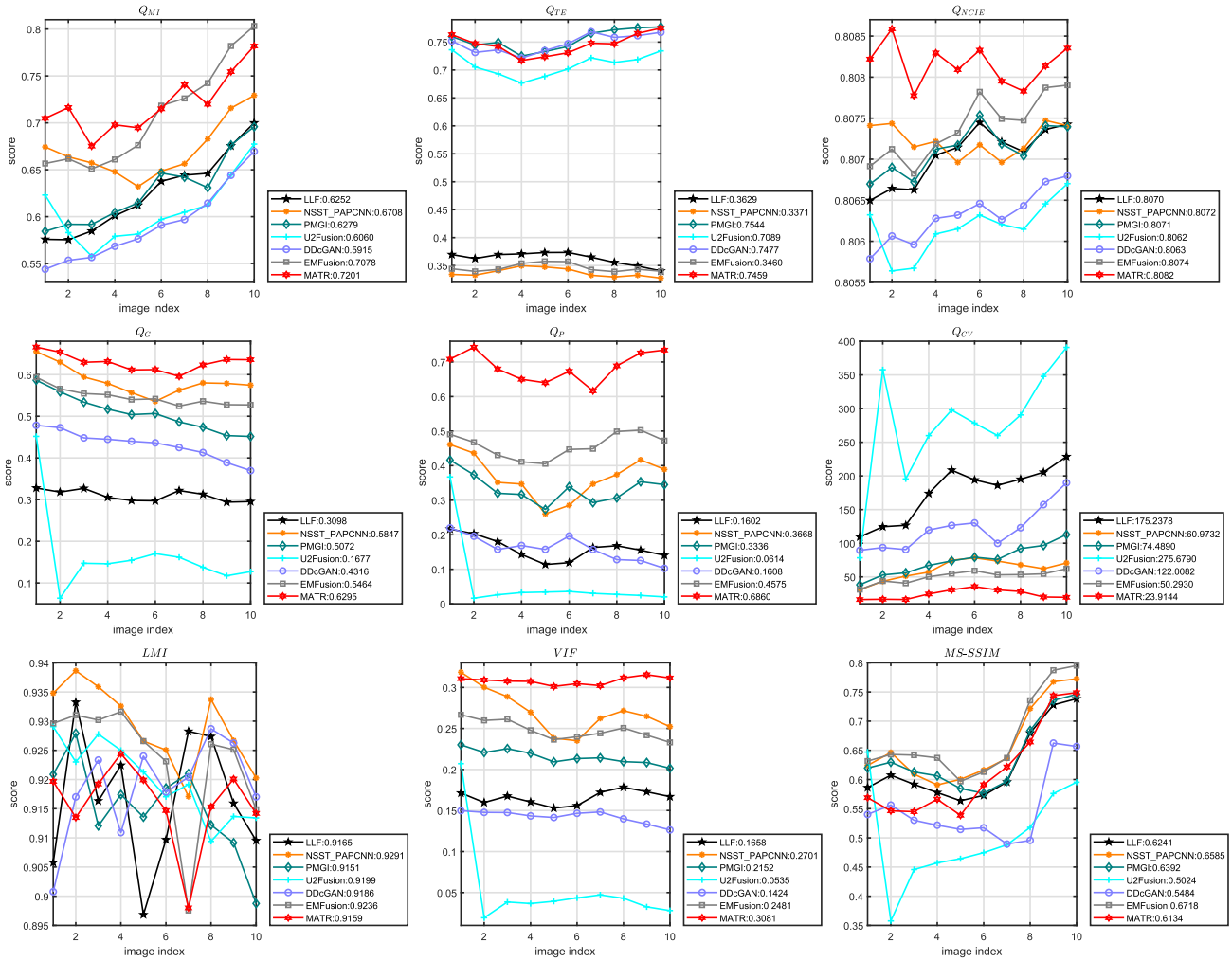

Fig. 10. Quantitative comparison of the proposed MATR for PET and MRI image fusion with 6 representative and state-of-the-art methods. Average scores for different methods are illustrated in the legends.

the-art methods. Obviously, Q_{NCIE}, Q_G, and Q_P reach the optimum scores on all ten testing samples. For Q_{MI} and VIF, higher evaluation values are generally obtained than other methods with the highest average numbers. For Q_{TE}, the proposed method reports higher scores on all testing samples than the LLF-based, NSST-PAPCNN-based, U2Fusion-based, and EMFusion-based methods, while it has relatively lower average values than the PMGI-based and DDcGAN-based methods. LMI and $MS\text{-}SSIM$ have comparatively lower quantitative assessments when compared with the most recently proposed algorithms, but the difference in number is not significant. Q_{CV} exhibits the bottom goals on all the testing images revealing the best fusion results. Based on the comparison of visual quality and objective evaluation, the proposed method always performs better than the other competitors, which illustrates that our MATR possesses good generalization capability.

F. Extension to GFP and PC Image Fusion

To further demonstrate the generalization capability of the proposed method, we extend the MATR to address other functional and structural image fusion issues (e.g., green fluorescent protein (GFP) and phase contrast (PC) images) without fine-tuning. The GFP image provides functional information associated with the distribution of proteins while the PC image contains abundant structural details of cells, including the nucleus and mitochondria. Therefore, the fusion of GFP and PC images can facilitate biological research, such as gene expression and protein functional analysis [67], [68]. Eighteen testing image pairs are downloaded from the mainstream database [3] to implement generalization experiments. Six representative and state-of-the-art methods are employed as comparison algorithms: DTCWT [64], NSST [65], NSST-PC [46], GAN [66], DPCN [47], and DARN [48]. The DTCWT-based, NSST-based, and NSST-PC-based methods are representative MST-based methods while the GAN-based, DPCN-based, and DARN-based methods are three recently proposed DL-based GFP and PC image fusion methods. All the parameters are fixed as reported in their publications for unbiased comparisons.

Fig. 11 exhibits six typical GFP and PC image pairs and their corresponding fusion results generated through different methods. It is clear that the DTCWT-based and NSST-based

[3]http://data.jic.ac.uk/Gfp /

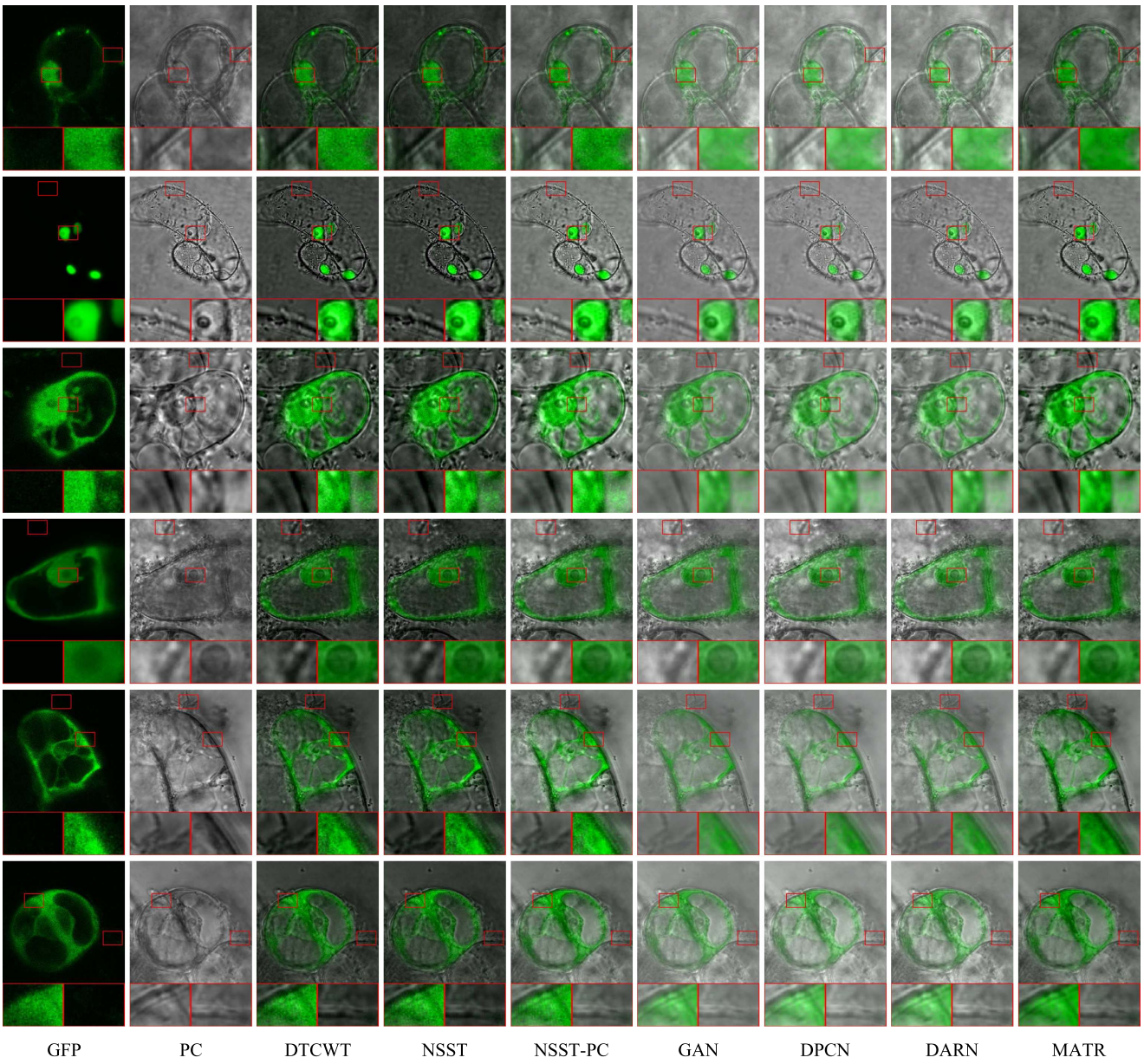

Fig. 11. Qualitative comparison of the proposed MATR with 6 representative and state-of-the-art methods on 6 GFP and PC image pairs. From left to right: GFP image, PC image, fusion results of DTCWT [64], NSST [65], NSST-PC [46], GAN [66], DPCN [47], DARN [48], and the proposed MATR. For better comparison, two local regions are enlarged as close-ups in each image.

methods have inaccurate preservation of the functional information of the GFP image, resulting in low contrast and blurred results. The NSST-PC-based method largely alleviates the over-extraction problem but still suffers from this drawback to some degree. The GAN-based, DPCN-based, and DARN-based methods can retain the complementary information from the source images well. However, compared with the proposed MATR, these three methods suffer slightly from inadequate capturing of the structural details from the PC image. On the whole, the proposed method achieves better performance in terms of adequately and accurately maintaining significant information from the input images.

The quantitative comparisons of the proposed method for GFP and PC image fusion with 6 representative and state-of-the-art methods are shown in Fig. 12. It can be seen that for Q_{MI}, Q_{TE}, Q_{NCIE}, and Q_G, the proposed MATR reaches the highest scores for almost all 18 testing samples and ranges the optimum average values. For Q_P, our method reveals clear advantages over the DTCWT-based, NSST-based, NSST-PC-based, GAN-based, and DARN-based methods, while it has a slightly lower mean value than the DPCN-based method. The proposed method reports moderate performance on LMI and VIF, and outperforms the most recently presented algorithms on Q_{CV} and $MS\text{-}SSIM$.

In general, for GFP and PC image fusion task, our method exhibits advantages over other contenders in terms of both subjective and objective comparisons.This demonstrates that the proposed MATR has satisfactory generalization ability.

Fig. 12. Quantitative comparison of the proposed MATR for GFP and PC image fusion with 6 representative and state-of-the-art methods on nine objective evaluation metrics. Average scores for different methods are illustrated in the legends.

V. Conclusion

In this paper, we propose a novel deep learning-based multimodal medical image fusion method via a multiscale adaptive Transformer called MATR. In our method, adaptive convolution and adaptive Transformer are introduced to extract global complementary context information for accurate fusion. To capture useful interscale information, the proposed model is designed in a multiscale manner. In adddition, an objective function is devised from both the structural-level and feature-level standpoints to train the proposed MATR in an unsupervised fashion. Extensive experiments on a publicly available medical database demonstrate that our method outperforms other representative and state-of-the-art methods in terms of both visual quality and objective evaluation. Furthermore, the proposed method is extended to address other biomedical functional and structural image fusion issues, and the satisfactory results illustrate its good generalization ability. Thus, our method has practical engineering application value to facilitate subsequent tasks, including diagnosis, treatment planning, and surgical navigation. We believe that the proposed MATR model is capable of dealing with more types of image fusion problems and for investigating more new image fusion approaches.

References

[1] R. Liu, J. Liu, Z. Jiang, X. Fan, and Z. Luo, "A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion," *IEEE Trans. Image Process.*, vol. 30, pp. 1261–1274, 2021.

[2] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and P. Bekaert, "Color balance and fusion for underwater image enhancement," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 379–393, Jan. 2017.

[3] A. Kumar, M. Fulham, D. Feng, and J. Kim, "Co-learning feature fusion maps from PET-CT images of lung cancer," *IEEE Trans. Med. Imag.*, vol. 39, pp. 204–217, 2020.

[4] P. Guo *et al.*, "Nuclei-based features for uterine cervical cancer histology image analysis with fusion-based classification," *IEEE J. Biomed. Health Inform.*, vol. 20, no. 6, pp. 1595–1607, Oct. 2016.

[5] C. Ciprian, D. Alexander, H. Wolf-Dieter, and R. R. Bruce, "PET/MRI for neurologic applications," *J. Nucl. Med.*, vol. 53, no. 12, pp. 1916–1925, 2012.

[6] A. Drzezga *et al.*, "Reirradiation of recurrent high-grade gliomas using amino acid PET (SPECT)/CT/MRI image fusion to determine gross tumor volume for stereotactic fractionated radiotherapy," *J. Nucl. Med.*, vol. 53, no. 6, pp. 845–855, 2012.

[7] J. Du, W. Li, and B. Xiao, "Anatomical-functional image fusion by information of interest in local Laplacian filtering domain," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5855–5866, Dec. 2017.

[8] 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.

[9] P. Hill, M. E. Al-Mualla, and D. Bull, "Perceptual image fusion using wavelets," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1076–1088, Mar. 2017.

[10] 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.

[11] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, pp. 147–164, Jul. 2015.

[12] J. Sun, H. Zhu, Z. Xu, and C. Han, "Poisson image fusion based on Markov random field fusion model," *Inf. Fusion*, vol. 14, no. 3, pp. 241–254, 2013.

[13] A. Wang, W. Wang, J. Liu, and N. Gu, "AIPNe: Image-to-image single image dehazing with atmospheric illumination prior," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 381–393, Jan. 2019.

[14] W. Xu and G. Wang, "A domain gap aware generative adversarial network for multi-domain image translation," *IEEE Trans. Image Process.*, vol. 31, pp. 72–84, 2022.

[15] T. Si, F. He, Z. Zhang, and Y. Duan, "Hybrid contrastive learning for unsupervised person re-identification," *IEEE Trans. Multimedia*, early access, May 11, 2022, doi: 10.1109/TMM.2022.3174414.

[16] W. Li, V. Pïëch, and C. D. Gilbert, "Perceptual learning and top-down influences in primary visual cortex," *Nature Neurosci.*, vol. 7, no. 6, pp. 651–657, Jun. 2004.

[17] C. D. Gilbert and W. Li, "Top-down influences on visual processing," *Nature Rev. Neurosci.*, vol. 14, no. 5, pp. 350–363, 2013.

[18] Y. Liu, X. Chen, J. Cheng, and H. Peng, "A medical image fusion method based on convolutional neural networks," in *Proc. 20th Int. Conf. Inf. Fusion (Fusion)*, Jul. 2017, pp. 1–7.

[19] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12797–12804.

[20] 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. 2020.

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

[22] Y. Fu, T. Xu, X. Wu, and J. Kittler, "PPT fusion: Pyramid patch transformer for a case study in image fusion," 2021, *arXiv:2107.13967*.

[23] D. Rao, X.-J. Wu, and T. Xu, "TGFuse: An infrared and visible image fusion approach based on transformer and generative adversarial network," 2022, *arXiv:2201.10147*.

[24] Z. Wang, Y. Chen, W. Shao, H. Li, and L. Zhang, "SwinFuse: A residual swin transformer fusion network for infrared and visible images," 2022, *arXiv:2204.11436*.

[25] Q. Zhang, L. Wang, H. Li, and Z. Ma, "Similarity-based multimodality image fusion with shiftable complex directional pyramid," *Pattern Recognit. Lett.*, vol. 32, no. 13, pp. 1544–1553, Oct. 2011.

[26] H. Wang, "A new multiwavelet-based approach to image fusion," *J. Math. Imag. Vis.*, vol. 21, no. 2, pp. 177–192, 2004.

[27] F. E. Ali, I. M. El-Dokany, A. A. Saad, and F. A. El-Samie, "A curvelet transform approach for the fusion of MR and CT images," *J. Mod. Opt.*, vol. 57, no. 4, pp. 273–286, Jan. 2010.

[28] W. W. Kong, Y. J. Lei, Y. Lei, and S. Lu, "Image fusion technique based on non-subsampled contourlet transform and adaptive unit-fast-linking pulse-coupled neural network," *IET Image Process.*, vol. 5, no. 2, pp. 113–121, 2011.

[29] P. Geng, Z. Wang, Z. Zhang, and Z. Xiao, "Image fusion by pulse couple neural network with shearlet," *Opt. Eng.*, vol. 51, no. 6, pp. 1–6, 2012.

[30] B. Escalante-Ramírez, "The Hermite transform as an efficient model for local image analysis: An application to medical image fusion," *Comput. Electr. Eng.*, vol. 34, no. 2, pp. 99–110, Mar. 2008.

[31] J. Du and W. Li, "Two-scale image decomposition based image fusion using structure tensor," *Int. J. Imag. Syst. Technol.*, vol. 30, no. 2, pp. 271–284, Jun. 2020.

[32] S. Zhang and F. He, "DRCDN: Learning deep residual convolutional dehazing networks," *Vis. Comput.*, vol. 36, no. 9, pp. 1797–1808, Sep. 2020.

[33] H. Li, F. He, Y. Chen, and Y. Pan, "MLFS-CCDE: Multi-objective large-scale feature selection by cooperative coevolutionary differential evolution," *Memetic Comput.*, vol. 13, no. 1, pp. 1–18, Mar. 2021.

[34] Y. Liang, F. He, X. Zeng, and J. Luo, "An improved loop subdivision to coordinate the smoothness and the number of faces via multi-objective optimization," *Integr. Comput.-Aided Eng.*, vol. 29, no. 1, pp. 23–41, Dec. 2021.

[35] G. Piella, "A general framework for multiresolution image fusion: From pixels to regions," *Inf. Fus.*, vol. 4, no. 4, pp. 259–280, Dec. 2003.

[36] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[37] K. Han *et al.*, "A survey on visual transformer," 2020, *arxiv:2012.12556*.

[38] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[39] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[40] J. Chen *et al.*, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[41] H. Chen *et al.*, "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12299–12310.

[42] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

[43] 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.

[44] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–16.

[45] M. Yin, X. Liu, Y. Liu, and X. Chen, "Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 1, pp. 49–64, Jan. 2019.

[46] W. Tang, Y. Liu, J. Cheng, C. Li, H. Peng, and X. Chen, "A phase congruency-based green fluorescent protein and phase contrast image fusion method in nonsubsampled shearlet transform domain," *Microsc. Res. Technique*, vol. 83, no. 10, pp. 1225–1234, May 2020.

[47] W. Tang, Y. Liu, J. Cheng, C. Li, and X. Chen, "Green fluorescent protein and phase contrast image fusion via detail preserving cross network," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 584–597, 2021.

[48] W. Tang, L. Wang, and Y. Liu, "Green fluorescent protein and phase contrast image fusion via dual attention residual network," in *Proc. IEEE Int. Conf. Med. Imag. Phys. Eng. (ICMIPE)*, Nov. 2021, pp. 1–6.

[49] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Inf. Fusion*, vol. 42, pp. 158–173, Jul. 2018.

[50] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, Dec. 2021.

[51] X. Lin, L. Ma, W. Liu, and S. Chang, "Context-gated convolution," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 701–718.

[52] 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.

[53] S. Zhao, Y. Wang, Z. Yang, and D. Cai, "Region mutual information loss for semantic segmentation," 2019, *arXiv:1910.12037*.

[54] Q. Xie, M. Zhou, Q. Zhao, Z. Xu, and D. Meng, "MHF-Net: An interpretable deep network for multispectral and hyperspectral image fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1457–1473, Mar. 2022.

[55] M. Hossny, S. Nahavandi, and D. Creighton, "Comments on 'Information measure for performance of image fusion,'" *Electron. Lett.*, vol. 44, no. 18, pp. 1066–1067, Aug. 2008.

[56] N. Cvejic, C. N. Canagarajah, and D. R. Bull, "Image fusion metric based on mutual information and Tsallis entropy," *Electron. Lett.*, vol. 42, no. 11, pp. 626–627, May 2006.

[57] Q. Wang, Y. Shen, and J. Q. Zhangb, "A nonlinear correlation measure for multivariable data set," *Phys. D, Nonlinear Phenomena*, vol. 200, nos. 3–4, pp. 287–295, 2005.

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

[59] J. Zhao, R. Laganiere, and Z. Liu, "Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement," *Int. J. Innov. Comput. Inf. Control*, vol. 3, no. 6, pp. 1433–1447, 2007.

[60] H. Chen and P. K. Varshney, "A human perception inspired quality metric for image fusion based on regional information," *Inf. Fusion*, vol. 8, no. 2, pp. 193–207, Apr. 2007.

[61] M. Hossny, S. Nahavandi, D. Creighton, and A. Bhatti, "Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters," *Electron. Lett.*, vol. 46, no. 18, pp. 266–1268, 2010.
[62] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Inf. Fusion*, vol. 14, pp. 127–135, Apr. 2013.
[63] Z. Wang, E. P. Simoncelli, and A. C. Bovil, "Multi-scale structural similarity for image quality assessment," in *Proc. IEEE Conf. Signals Syst. Comput.*, vol. 2, Nov. 2003, pp. 1398–1402.
[64] J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, and N. Canagarajah, "Pixel-and region-based image fusion with complex wavelets," *Inf. Fusion*, vol. 8, no. 2, pp. 119–130, 2007.
[65] Y. Cao, S. Li, and J. Hu, "Multi-focus image fusion by nonsubsampled shearlet transform," in *Proc. 6th Int. Conf. Image Graph.*, Aug. 2011, pp. 17–21.
[66] W. Tang, Y. Liu, C. Zhang, J. Cheng, H. Peng, and X. Chen, "Green fluorescent protein and phase-contrast image fusion via generative adversarial networks," *Comput. Math. Methods Med.*, vol. 2019, pp. 1–11, Dec. 2019.
[67] O. A. Koroleva, M. L. Tomlinson, D. Leader, P. Shaw, and J. H. Doonan, "High-throughput protein localization in arabidopsis using agrobacterium-mediated transient expression of GFP-ORF fusions," *Plant J.*, vol. 41, no. 1, pp. 162–174, Nov. 2005.
[68] S. Borgeaud, L. C. Metzger, T. Scrignari, and M. Blokesch, "The type VI secretion system of vibrio cholerae fosters horizontal gene transfer," *Science*, vol. 347, no. 6217, pp. 63–67, Jan. 2015.

Wei Tang received the B.E. degree from the Wannan Medical College, Wuhu, China, in 2018, and the M.S. degree in biomedical engineering from the Hefei University of Technology, Hefei, China, in 2021. She is currently pursuing the Ph.D. degree with the School of Computer Science, Wuhan University, Wuhan, China. Her current research interests include image processing, computer vision, and information fusion.

Fazhi He (Member, IEEE) received the bachelor's, master's, and Ph.D. degrees from the Wuhan University of Technology. He was a Postdoctoral Researcher at the State Key Laboratory of CAD&CG, Zhejiang University; a Visiting Researcher at the Korea Advanced Institute of Science and Technology; and a Visiting Faculty Member of The University of North Carolina at Chapel Hill. He is currently a Professor with the School of Computer Science, Wuhan University. His research interests are artificial intelligence, intelligent computing, computer graphics, image processing, computer-aided design, computer supported cooperative work, and co-design of software/hardware.

Yu Liu (Member, IEEE) received the B.S. and Ph.D. degrees from the Department of Automation, University of Science and Technology of China, Hefei, China, in 2011 and 2016, respectively.

He is currently an Associate Professor with the Department of Biomedical Engineering, Hefei University of Technology, Hefei. His research interests include image processing, computer vision, information fusion, and machine learning. In particular, he is interested in image fusion, image restoration, visual recognition, and deep learning. He is serving as an Editorial Board Member for *Information Fusion*.

Yansong Duan was born in 1975. He received the M.S. and Ph.D. degrees from Wuhan University, Wuhan, China, in 2009 and 2016, respectively. He is currently an Associate Professor with the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include photogrammetry, image processing, matching, 3D city reconstruction, computer vision, and high performance computing.