DMMFnet: A Dual-Branch Multimodal Medical Image Fusion Network Using Super Token and Channel-Spatial Attention

Yukun Zhang[1], Lei Wang[2]*, Muhammad Tahir[3]*, Zizhen Huang [4],
Yaolong Han[5], Shanliang Yang[6], Shilong Liu[7], Muhammad Imran Saeed[8]
School of Computer Science and Technology, Shandong University of Technology, Zibo, 255000, P.R. China[1, 2, 4, 5, 6, 7]
Department of Computer Science, Mohammad Ali Jinnah University, Block-6, P.E.C.H.S, Karachi-75400, Sindh, Pakistan[3, 8]

***Abstract*—Multimodal medical image fusion leverages the correlation between different modal images to enhance the information contained within a single medical image. Existing fusion methods often fail to effectively extract multiscale features from medical images and establish long-distance relationships between deep feature blocks. To address these issues, we propose DMMFnet, an encoder-decoder fusion network that utilizes shared and private encoders to extract shared and private features. DMMFnet is based on super token sampling and channel-spatial attention. The shared encoder and decoder use a transformer structure with super token sampling technology to effectively integrate information from different modalities, improving processing efficiency and enhancing the ability to capture key features. The private encoder consists of invertible neural networks and transformer modules, designed to extract local and global features, respectively. A novel transformer module refines attention distribution and feature aggregation to capture superpixel-level global correlations, ensuring that the network effectively captures essential global information, thereby enhancing the quality of the fused image. Experimental results, comparing DMMFnet with nine leading fusion methods, indicate that DMMFnet significantly improves various evaluation metrics and achieves superior visual effects, demonstrating its advanced fusion capability.**

Keywords*—Medical image fusion; channel-spatial attention; super token sampling; encoder–decoder

I. Introduction

Rapid advancements in medical imaging technology have allowed for the integration of multimodal medical pictures into clinical diagnosis, surgical guidance, and medical research in recent decades [1]. However, different medical images emphasize distinct aspects. For example, Computer Tomography (CT) scans yield accurate images of the bones, they do not capture the fine details of soft tissues. In contrast, Magnetic Resonance Imaging (MRI) offers finely detailed images of the organs' soft tissues, providing substantial clinical diagnostic value [2]. Positron Emission Tomography (PET) images reflect metabolic changes and functional states of lesions through the ingestion of imaging agents. Single Photon Emission Computed Tomography (SPECT) images diagnose a broad spectrum of diseases using varying depth colors to mark affected areas [3, 4]. The usage of the medical images one by one modality to diagnose diseases is not only time-consuming but also requires extensive experience. Therefore, the goal of multimodal medical image fusion techniques is to create a single multimodal image from two modalities, and the outcomes can preserve the meaning, unique characteristics, and information from the original images, such as high-resolution structural data from CT, tissue textures from MRI [5]. Fused images have a richer texture structure and more pronounced lesion areas compared to single-modality medical images [6]. This greatly helps physicians analyze the diseases that are challenging to observe, reducing the misdiagnoses rate and surgical errors.

Usually, there are two types of image fusion tasks: supervised learning fusion and unsupervised learning fusion. Supervised learning is predominantly applied in the domain of multi-focus image fusion [7, 8]. Supervised learning is considered unsuitable for medical image fusion (MMIF) tasks due to the unique nature of medical images [9]. Moreover, due to the characteristics of medical images, the fusion methods designed for other types of images cannot be directly applied to multimodal medical image fusion tasks. According to different computational approaches, medical image fusion methods can be divided into traditional methods and deep learning methods.

Historically, among the traditional fusion methods, multi-scale transform (MST)-based methods, such as wavelet transform [10], pyramid transform [11], and subspace transform [12], have been commonly used. While the tower-based decomposition laid the groundwork for MST-based image fusion research with relatively simple implementation, it lacks directionality and is sensitive to noise, leading to redundancy between the pyramid levels. Wavelet transform offers good time-frequency locality and directionality without information redundancy, but it lacks directional selectivity and translation invariance, failing to fully extract edge information in images. Choosing the appropriate subspace mapping methods for specific fusion tasks remains a significant challenge in MST-based fusion methods.

For image fusion tasks, Pulse-Coupled Neural Networks (PCNN) [13] have received the most attention. Yin et al. proposed a fusion method combining NSST with PCNN (NSST-PAPCNN) for multimodal image fusion tasks. In this approach, NSST is used for feature extraction from multiple levels [14]. Tan et al. developed NST-MSMG [15]. It fuses high-frequency data using PCNN and boundary measurements and fuses low-frequency features utilizing an energy-based fusion rule set. However, PCNN-based approaches nevertheless follow the

*Corresponding Author.

fundamentals of multi-scale transform (MST) methods, which are needed for well-crafted decomposition and fusion rules.

The sparse representation (SR) [16] is widely utilized for image fusion, employing a mechanism that optimizes an extensive dictionary and generates sparse coefficients to achieve effective fusion. Liu et al. integrated multi-scale decomposition with convolutional sparse representation (CSR) [17]. However, methods based on SR have high computational demands, generally employing a complete and redundant dictionary for adaptive sparse representation of images. Furthermore, applying the same decomposition operations to different modalities of the source images may result in unexpected artifacts. In addition, the manual construction of the decomposition strategies and fusion rules make the fusion process complex and time-consuming [18, 19].

Recent advances have seen the adoption of deep learning techniques for multimodal image fusion, aiming to address the shortcomings of conventional fusion methods, all of which can be classified into three main types according to the network architectures: the Auto-encoder-based image fusion, the Convolutional Neural Network [20] (CNN)-based image fusion, and the Generative Adversarial Network [21] (GAN)-based image fusion.

The CNN-based image fusion methods are effective in processing spatial and structural information within image neighborhoods. Although CNN-based models are proficient in extracting local details and inductive biases from images, they lack a comprehensive understanding and learning of global semantic information in images. Additionally, due to their limited receptive field, CNNs inherently find it challenging to capture long-range relationships within images. To deal with these problems, Dosovitskiy et al. introduced the Vision Transformer (ViT) [22], which uses self-attention to conduct global comparisons across all visual tokens. It has shifted the paradigm from CNN-based feature extraction. Subsequent studies [23, 24] have shown that ViTs have potent global dependency learning capabilities in visual content. But recent research [25, 26] has shown that ViTs tend to capture shallow local features with high redundancy. This is due to shallow global attention focusing on a few adjacent tokens and neglecting most distant ones, which heavily hinders the extraction of the texture details in the fused image [27].

GAN is a type of deep learning model consisting of two modules: the generator and the discriminator. Ma et al. applied GANs to multimodal medical image fusion tasks [28]. Hung et al. introduce a multi-generator method for image fusion [29]. However, GAN-based image fusion approaches are prone to training instabilities and gradient vanishing issues [30]. Moreover, GAN architectures lose structural details due to down-sampling in pooling layers, which results in inefficient utilization of image information. The auto-encoder-based image fusion utilizes an unsupervised neural network model that comprises an encoder and a decoder. Deep Fuse [31] was one of the first methods in this domain. Li et al. introduced DenseNet and nested connections to improve the feature extraction capability of encoders [32, 33]. Furthermore, Jian et al. enhanced a self-encoder-based fusion framework by integrating an attention mechanism, aiming to extract features that are more interpretable [34]. However, due to the characteristics of GAN models, image fusion methods based on GANs are prone to instability during training. Additionally, GAN-based methods predominantly rely on CNN architectures, their limited ability to capture global information often results in insufficient fusion.

Although the above multimodal image fusion methods have obtained quite good results, several of the aforementioned problems persist. To deal with them, we propose an encoder-decoder network that uses the Invertible Neural Networks (INN) [35] and transformer module. The proposed method demonstrates superior feature extraction capabilities compared to existing method. By utilizing two distinct feature extractors to capture features at varying frequencies and then separately fusing these features during the fusion stage, our method preserves the original image's texture and structural information to the greatest extent possible. This approach is more effective in achieving the objectives of the MMIF tasks.

Here are the primary contributions of this study:

1) In order to effectively extract complementary information from the input images, a novel transformer module has been designed. The spatial and channel attention mechanisms are utilized to capture super-pixel-level global dependencies, resulting in a significant enhancement of the fusion image quality as demonstrated by both subjective and objective experiments.

2) The Context Broadcasting (CB) technique is employed in the transformer layer. This integration ensures consistent attention at each layer of the transformer model, thereby reducing the density of attention maps. Furthermore, the consistent attention mechanisms facilitate easier overall optimization of the model, aiding in more effective learning and representation of the complex relationships within the input data.

3) Extensive experiments on medical and biological image fusion demonstrate that the DMMFnet outperforms nine advanced fusion methods in terms of both quantitative metrics and visual assessment.

The organization of the paper is as follows: Section I introduces the research background and the contributions of this work. Section II provides a detailed description of the proposed DMMFnet. Section III presents and discusses the experimental results. Section IV concludes the paper.

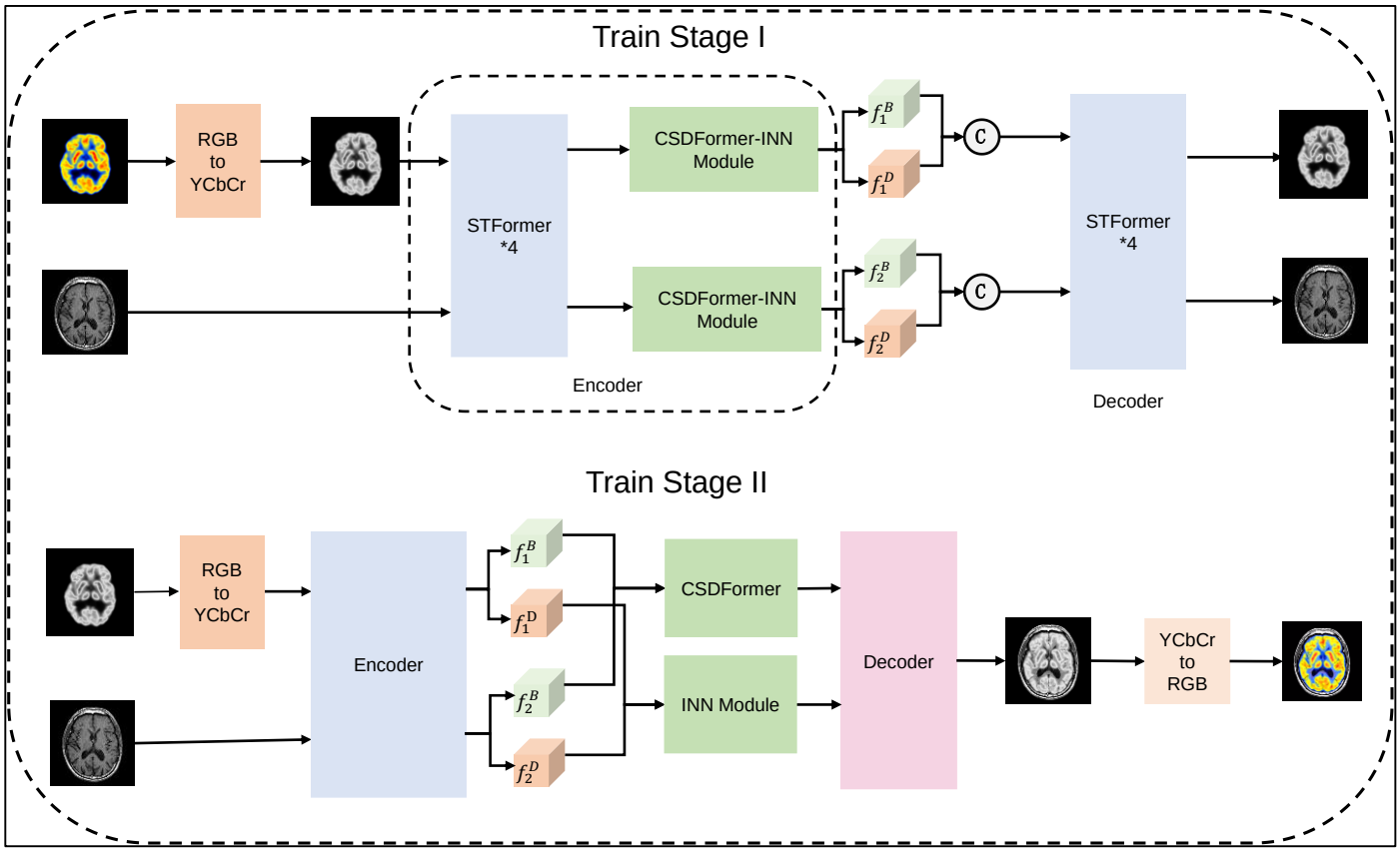

Fig. 1. The architecture of our DMMFnet method.

II. The Proposed Fusion Network

The detailed architecture of the proposed fusion framework is shown in Fig. 1. In general, CNN captures local features of the input image, whereas the transformer focuses on global features [36]. Therefore, we designed an encoder with a dual-branch structure. We used the INN to extract local features of the image and the transformer branch to extract global features of the image, then fused them separately. Finally, the DMMFnet comprises three modules: a fusion layer designed to combine different features, a decoder is used to rebuild the image and create the fused image, while an encoder is used to extract features.

The encoder consists of three components: a Shared Feature Extractor (SFE) based on STFormer, a Global Feature Extractor (GFE) based on CSDFormer, and a Local Feature Extractor (LFE) based on INN. Using PET-MRI image fusion as an example, we have defined some symbols to explain the entire fusion process: The paired PET and MRI input images are denoted as P and M, respectively. SFE, GFE, and LFE are indicated by $S(\,)$, $G(\,)$, and $L(\,)$, respectively. To extract shared features from the inputs is the aim of the SFE. This process is offered in Eq. (1):

$$f_P^S = S(P), f_M^S = S(M) \quad (1)$$

The model's Shared Feature Encoder includes the STFormer, which is based on super token sampling attention blocks [37], and the Gated-Dconv Feed-Forward Network (GDFN) module [38]. Please refer to the original paper for more information on the structure of the super token sampling attention and GDFN. The schematic of the STFormer is depicted in Fig. 2.

By incorporating super tokens into the transformer and utilizing sparse associative learning to sample super tokens from visual tokens, we can effectively capture global dependencies through self-attention on these super tokens. The reason for selecting super token sampling attention blocks in the shared feature extraction step is that they efficiently capture global dependencies by decomposing global attention into a multiplication of sparse associative mappings and low-dimensional attention. This reduces computational complexity while retaining key image information.

The GDFN block is intended to merge features from various sources, specifically images from different modalities that exhibit unique characteristics within certain frequency ranges. The DMMFnet network can flexibly process and merge these features through the GDFN block.

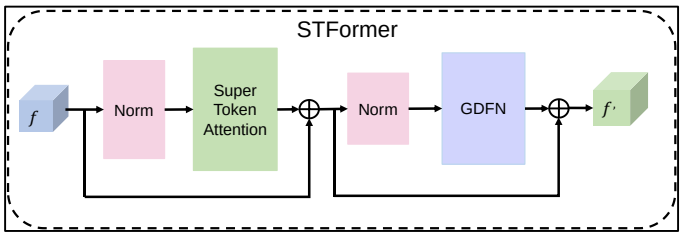

Fig. 2. The schematic of the shared feature encoder.

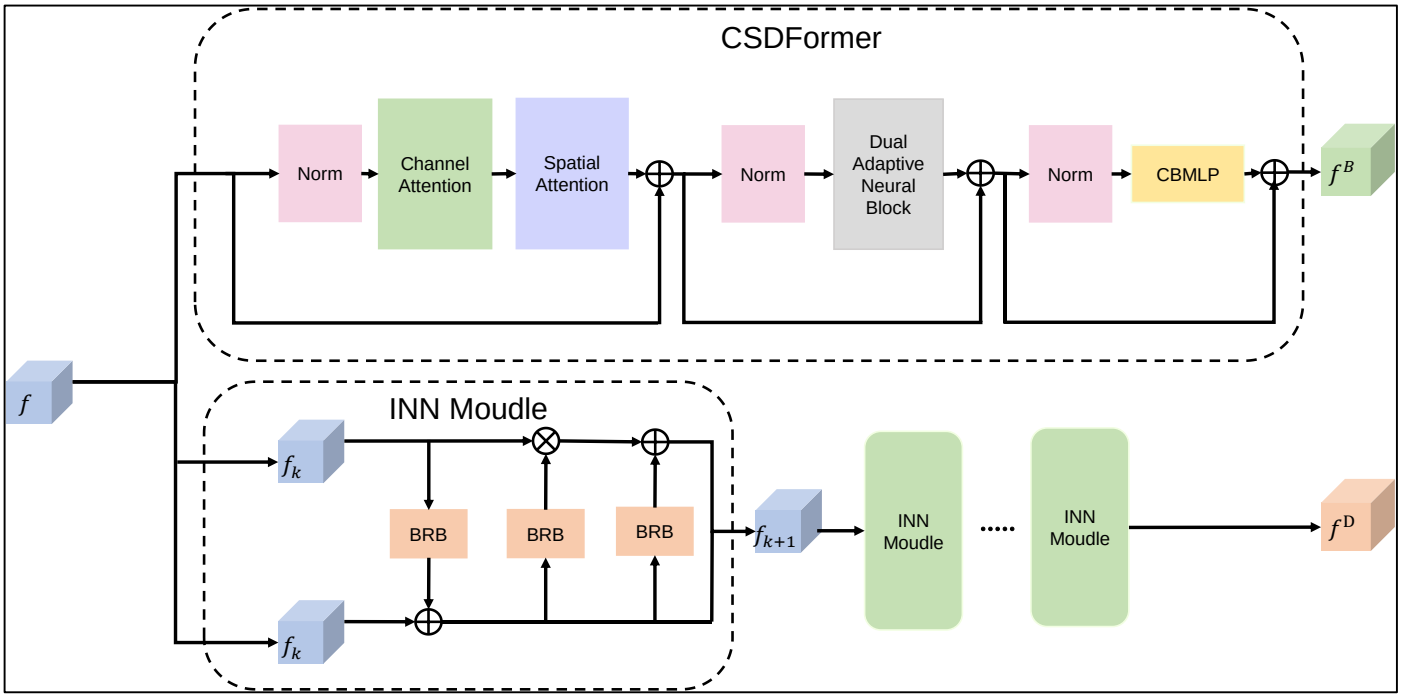

Fig. 3. The framework of the Global Feature Extractor (GFE) and Local Feature Extractor (LFE).

As illustrated in Fig. 3, the private feature encoder consists of two branches: GFE and LFE. The GFE extracts global features, while the LFE extracts local features.

The GFE branch concentrates on low-frequency global features by finely tuning the attention distribution. It effectively captures global dependencies at the super-pixel level through spatial and channel attention. By incorporating a Dual Adaptive Neural Block (DA), it adaptively encapsulates global features from superpixels to pixels, optimizing feature focus and refinement to ensure the capture of subtle changes and patterns crucial to the fusion process. In designing the CSDFormer, a Context Broadcasting (CB) technique was employed in the MLP layer. This technique involves manually inserting uniform attention into every layer of the ViT model, providing the necessary dense interactions and lowering the concentration level of attention maps across all layers. CB also enhances its capacity and generalization ability with negligible cost [39]. which is formulated as Eq. (2).

$$f_P^B = G(f_P^S), f_M^B = G(f_M^S) \tag{2}$$

The LFE branch is focused on the lossless extraction of local high-frequency features. Given that edge and texture details are crucial for image fusion tasks, the INN module is utilized to preserve as many image details as possible. The INN module aims to mitigate information loss by mutually generating input and output features through a reversible design, and it can retain the high-frequency information of the medical image almost without any loss. The process is offered in Eq. (3) below.

$$f_P^D = L(f_P^S), f_M^D = L(f_M^S) \tag{3}$$

A. Fusion Layer

The fusion layer comprises the basic feature fusion layer and the deep feature fusion layer, which respectively combine the basic and deep features. The fusion of basic and deep features is akin to the extraction of basic and deep features in the encoder. Therefore, CSDFormer and INN blocks are employed for the basic and deep fusion layers. The fusion process can be expressed by Eq. (4).

$$f^B = F_B(f_P^B, f_M^B), f^D = F_D(f_P^D, f_M^D) \tag{4}$$

F_B and $F_D$represent the basic and deep feature fusion layers, respectively.

B. Decoder

The different features extracted and processed in the previous phase are used as input to the decoder *DC()*. The reconstructed image from training stage I and the fused image from training stage II are the outputs of *DC()*. The corresponding formula is as follows:

$$\Sigma\tau\alpha\gamma\varepsilon\mathrm{I}:\ P^* = DC(f_P^B,\ f_P^D), M^* = \mathrm{DC}(f_M^B,\ f_M^D) \tag{5}$$

$$\Sigma\tau\alpha\gamma\varepsilon\mathrm{II}: FUSE = DC(f^B, f^D) \tag{6}$$

Due to the input includes cross-modal and multi-frequency features, ensuring consistency between the decoder and the shared encoder enables the decoder to better understand and exploit the feature representations provided by the encoder, leading to improved fusion results. Therefore, we employ STFormer blocks as the fundamental units for the decoder.

C. Two-stage Training

A significant obstacle in medical image fusion tasks is the absence of a definitive ground truth due to the expensive and privacy-sensitive nature of the data sources, rendering advanced supervised learning methods ineffective. Therefore, we employ a two-stage learning approach to end-to-end train the DMMFnet.

Training stage I: In this phase, we initially feed the paired PET and MRI images $\{P, M\}$ into the Shared Feature Extractor (SFE) to extract their shared features $\{f_P^S, f_P^S\}$.Subsequently, each image is processed through the GFE based on the CSDFormer structure and the LFE based on INN separately. The basic features $\{f_P^B, f_M^B\}$ and detail features $\{f_P^D, f_M^D\}$ are extracted from the two modalities. The basic and detailed features within the same modality are merged (such as $\{f_P^B, f_P^D\}$ for PET or $\{f_M^B, f_M^D\}$ for MRI) and transmitted to the decoder for the reconstruction of the original PET or MRI image.

Training stage II: We continue to use paired PET and MRI images $\{P, M\}$ as the input. However, this time, we fed them into the encoder that was trained in Training Stage I. This enables further decomposition of features. Afterwards, we input the base features $\{f_P^B, f_M^B\}$ and detail features $\{f_P^D, f_M^D\}$ individually into fusion layers F_B and F_B .After the feature fusion process, the fused features $\{f^B, f^D\}$ are inputted into the decoder, which generates the fused image $FUSE$.

D. Loss Function

In training phase I, the total loss L_{total} is offered in Eq. (7) below.

$$L_{total} = \alpha_1 L_{pet} + \alpha_2 L_{mri} + \alpha_3 L_{decomp} \tag{7}$$

L_{pet} and L_{mri} represent the reconstruction losses for the two types of medical images. Since the model in the first stage can be regarded as a process of decomposition followed by synthesis, information loss inevitably occurs during both the decomposition and synthesis stages. L_{decomp} denotes the feature decomposition loss, while α_1 , α_2 , α_3 are adjustment parameters. The overall loss function in the first stage is designed to ensure that information is maintained throughout the encoding and decoding procedures. Each loss function is as follows.

$$L_{pet/mri} = L_{int}^{I}(I, I^*) + \sigma L_{ssim}(I, I^*) \tag{8}$$

$$L_{decomp} = \frac{\left(CC(f_P^D, f_M^D)\right)^2}{CC(f_P^B, f_M^B)+\varepsilon} \tag{9}$$

$$L_{int}^{I} = \frac{1}{HW}||I - I^*|| \tag{10}$$

where I denotes the original image before reconstruction, and I^* represents the image reconstructed in the first stage. Here, to ensure the positivity of this term, the operator CC is the correlation coefficient and ε is assigned a value of 1.01.

In training phase II, the total loss L_{total} is offered in Eq. (11) below.

$$L_{total} = \alpha_1 L_{ssim} + \alpha_2 L_{test} + \alpha_3 L_{int}^{II} \tag{11}$$

L_{ssim}represents the structural loss, measuring the similarity between two images. L_{test} stands for texture loss, while L_{int}^{II} denotes intensity loss. α_1, α_2, α_3 are adjustment parameters. The overall loss function in the second stage aims to train the fusion network weights while simultaneously adjusting the the first stage's trained encoder and decoder. Each loss function is as follows.

$$L_{int}^{II} = \frac{1}{HW}||I_f - Max(I_{pet}, I_{mri})|| \tag{12}$$

$$L_{text} = \frac{1}{HW}||\,|\nabla I_f| - Max(|\nabla I_{pet}|, |\nabla I_{mri}|)|| \tag{13}$$

$$L_{ssim} = \gamma_1(1 - ssim(I_f, I_{pet})) + \gamma_2(1 - ssim(I_f, I_{mri})) \tag{14}$$

H and W stand for the original image's height and width, respectively. γ_1 and γ_2 are the adjustment parameters used to set the importance of information from each image. In the context of multimodal medical image fusion, the most critical aspect is effectively preserving information about lesion regions and texture details. We consider the structural information of MRI images to be more important in the MMIF. Through experimentation, we have found that the best fusion results are achieved when γ_1= 0.45 and γ_2= 0.55.

III. Experimental Results and Discussion

A. The Dataset and Experimental Setting

The present study involves two fusion tasks: biological and medical image fusion. The medical image tasks include MRI/CT fusion, MRI/PET fusion, and MRI/SPECT fusion. All data can be obtained from [40]. The MRI/CT brain training set consists of 160 pairs, with 24 pairs in the test set; the MRI/PET brain training set comprises 245 pairs, with 24 pairs in the test set; and the MRI/SPECT brain training set includes 333 pairs, with 24 pairs in the test set. The biological image fusion tasks include Green Fluorescent Protein (GFP) and Phase Contrast (PC) images fusion, with data sourced from [41]. The training set comprises 130 pairs, while the test set consists of 18 pairs. During the preprocessing stage, the training images are cropped to a size of 128 × 128.

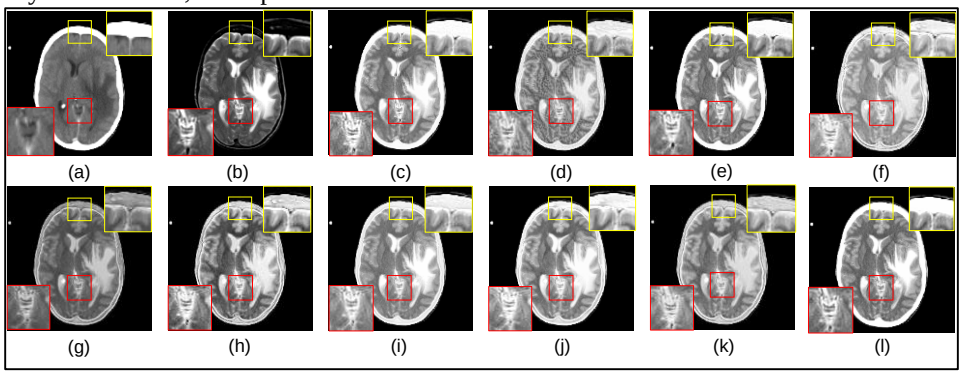

Fig. 4. The comparison of CT-MRI fusion results.

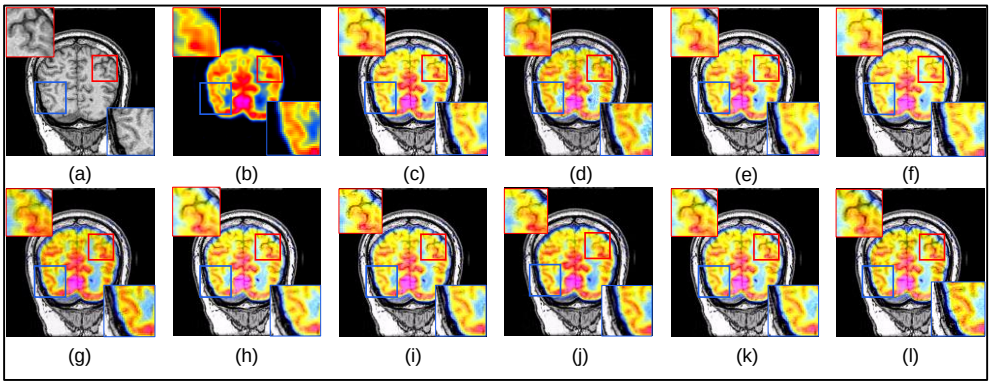

Fig. 5. The comparison of PET-MRI fusion results.

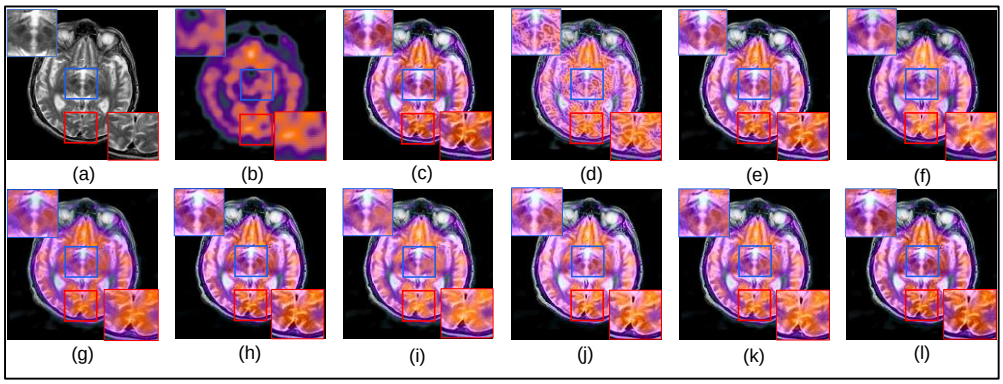

Fig. 6. The comparison of SPECT-MRI fusion results.

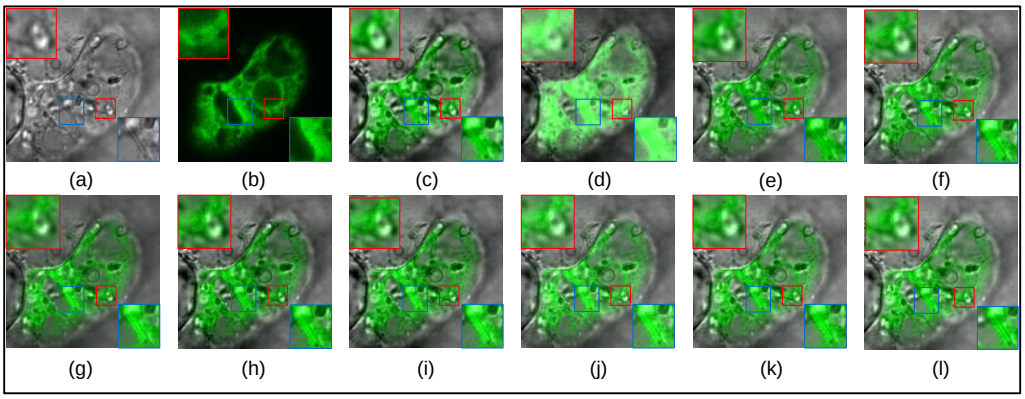

Fig. 7. The comparison of GFP-PC fusion results.

B. The Evaluation Metrics

The quantitative metrics must be used in order to compare the fusion results objectively. The fusion results are quantitatively assessed from six dimensions: entropy of information (EN) [42], Spatial Frequency (SF) [43], mutual information (MI) [44], Structural Similarity Index Measure (SSIM) [45], Visual Information Fidelity (VIF) [46], and edge-based similarity measurement $Q^{AB/F}$ [47]. Finally, calculate the average value of each metric for the respective methods for objective analysis.

C. The Subjective Analysis

For evaluation, we choose three fusion methods based on traditional methods and six state-of-the-art deep learning-based fusion methods., including PSO–NSST [48], LLPACM [49], PCNN–NSST 50], SDnet [51], EMFusion [52], U2Fusion [6], SwinFusion [53], CDDFuse [54], and CoCoNet [55]. The default values supplied by the authors of these image fusion algorithms are the configurations for the parameter settings. SwinFusion was originally designed to handle only MRI-PET images, necessitating the retraining of the model for other modal images.

The fusion results for CT-MRI, MRI-PET, MRI-SPECT, and GFP-PC are shown in Fig. 4, Fig. 5, Fig. 6, and Fig. 7, respectively. In each figure, (a) and (b) are the original images. (c)-(l) represent the fused images using the PSO – NSST, LLPACM, PCNN – NSST, SDnet, EMFusion, U2Fusion, SwinFusion, CDDFuse, CoCoNet and the DMMFnet method.

In Fig. 4, Fig. 4 (a) and Fig. 4 (b) represent the original CT and MRI images. Fig. 4 (g) is notably dark, obscuring detailed information. Fig. 4 (f), Fig. 4 (h), and Fig. 4 (k) lose some information from the original CT images, with the brain's outline becoming indistinct. Fig. 4 (d) is overly blurred, affecting clarity and detail. Fig. 4 (i) and Fig. 4 (j) are excessively bright, which hinders clear information presentation. Fig. 4 (c) and Fig. 4 (e) yield acceptable results but suffer from noise-like artifacts that affect structural details. Due to our proposed feature extraction strategy, the DMMFnet effectively preserves most of the structural information from the source images and excellently maintains the edges.

As shown in Fig. 5, Fig. 5 (a) Fig. 5 (b) represent the original MRI and PET images, respectively. Fig. 5 (e) and Fig. 5 (f) appear overly bright, resulting in unclear visual effects. Fig. 5 (g) is too dark, leading to poor preservation of MRI information. Fig. 5 (d) fails to effectively extract and preserve information, resulting in severe color distortion. Although Fig. 5 (i) and Fig. 5 (j) retain color information well, there is still some loss of MRI information at the edges. Due to our proposed fusion method, the DMMFnet successfully integrates the main information from the source images into the fused image and accurately portrays lesion information.

As shown in Fig.6, Fig.6 (a) and Fig.6 (b) are the original MRI and SPECT images. In contrast to PET, SPECT images have much sparser intensity information, leading to varying degrees of information loss across different deep learning methods. Fig. 6 (d) is overly blurred, while Fig. 6 (f) and Fig. 6 (i) retain too much color information, resulting in overly bright images that obscure details. Fig. 6 (g) suffers from severe deficiencies in color information. The DMMFnet outperforms all other deep learning methods.

As shown in Fig.7, Fig.7 (a) and Fig.7 (b) are the original GFP and PC images. Due to the higher resolution of GFP and PC images compared to medical images, the feature information is more apparent, and deep learning methods generally yield better visual results. In Fig.7 (d), significant feature loss is observed. In Fig.7 (c) and Fig.7 (e), the cell structures are preserved, but there is some loss of color information. Conversely, in Fig.7 (i), Fig.7 (j), and Fig.7 (k), the cell structures are not clearly preserved. Additionally, Fig.7 (g) is too dark, resulting in poor visual quality. DMMFnet preserves both the cell structures and color information more comprehensively.

D. The Objective Analysis

In MRI-CT image fusion, our algorithm possesses four optimal metrics (EN, SF, SSIM, and $Q^{AB/F}$) and one suboptimal metric (VIF), with the SF metric performing significantly better than the others. As shown in Table I.

For MRI-PET image fusion, compared to other algorithms, our algorithm achieves the highest scores in SF, MI, SSIM, and $Q^{AB/F}$, while being second best in VIF as seen in Table II.

In MRI-SPECT image fusion, our algorithm has two optimal metrics (VIF and $Q^{AB/F}$) and one suboptimal metric, surpassing all deep learning methods, as seen in Table III.

TABLE I. THE OBJECTIVE EVALUATION OF CT-MRI FUSION IMAGES

Fusion Methods	Evaluation metrics					
	EN	SF	MI	SSIM	VIF	$Q^{AB/F}$
PSO–NSST	4.5	34.25	**2.09**[1]	**1.36**[1]	0.5	0.5
LLPACM	4.62	32.9	1.98	1.28	0.39	0.45
PCNN–NSST	4.58	36.07	2.07	1.34	0.47	**0.57**[2]
SDnet	4.66	34.77	**2.26**[2]	1.29	**0.51**[2]	0.52
EMFusion	4.62	26.7	2.04	**1.35**[2]	0.43	0.5
U2Fusion	4.64	35.77	1.94	1.34	0.4	0.51
SwinFusion	**4.69**[1]	33.6	1.92	**1.35**[2]	**0.61**[1]	0.51
CDDFuse	4.62	35.3	2.08	1.33	0.49	0.53
CoConet	**4.68**[2]	**36.15**[2]	1.92	1.33	0.43	0.54
Proposed	**4.69**[1]	**38.52**[1]	2.04	**1.36**[1]	**0.51**[2]	**0.62**[1]

TABLE II. THE OBJECTIVE EVALUATION OF PET-MRI FUSION IMAGES

Fusion Methods	Evaluation metrics					
	EN	SF	MI	SSIM	VIF	$Q^{AB/F}$
PSO–NSST	5.32	**38.03**[2]	2.75	**1.27**[2]	0.65	0.71
LLPACM	5.32	33.67	2.29	1.22	0.63	0.72
PCNN–NSST	5.44	37.37	2.45	1.26	0.65	0.71
SDnet	5.43	37.29	2.44	1.26	0.67	0.69
EMFusion	5.38	32.9	2.26	**1.27**[2]	**0.68**[2]	0.7
U2Fusion	5.35	38.01	**2.77**[2]	1.26	0.67	**0.73**[2]
SwinFusion	**5.52**[2]	36.49	2.12	1.23	**0.71**[1]	0.68
CDDFuse	5.44	37.68	2.43	1.25	0.62	0.67
CoConet	**5.58**[1]	37.89	2.63	1.26	0.67	**0.74**[1]
Proposed	5.4	**38.05**[1]	**2.81**[1]	**1.28**[1]	**0.68**[2]	**0.74**[1]

TABLE III. THE OBJECTIVE EVALUATION OF SPECT-MRI FUSION IMAGES

Fusion Methods	Evaluation metrics					
	EN	SF	MI	SSIM	VIF	$Q^{AB/F}$
PSO–NSST	**5.54**[2]	**27.93**[1]	**3.07**[1]	1.25	**0.84**[2]	0.71
LLPACM	5.33	26.47	2.31	1.08	0.46	0.57
PCNN–NSST	**5.56**[1]	**27.16**[2]	**3.02**[2]	**1.3**[1]	0.74	0.7
SDnet	5.2	23.77	2.27	1.25	0.63	0.69
EMFusion	5.2	20.11	2.42	1.26	0.71	0.71
U2Fusion	5.08	25.45	2.8	1.25	0.81	**0.74**[2]
SwinFusion	5.07	22.67	2.12	**1.28**[2]	**0.87**[1]	0.64
CDDFuse	5.2	25.98	2.69	1.24	0.75	0.72
CoConet	5.18	25.27	2.89	1.27	0.7	0.71
Proposed	5.09	25.83	**3.07**[1]	**1.28**[2]	**0.87**[1]	**0.75**[1]

In GFP-PC image fusion, our algorithm possesses four optimal metrics (EN, MI, SSIM, VIF, and $Q^{AB/F}$), as seen in Table IV.

TABLE IV. THE OBJECTIVE EVALUATION OF GFP-PC FUSION IMAGES

Fusion Methods	Evaluation metrics					
	EN	SF	MI	SSIM	VIF	$Q^{AB/F}$
PSO–NSST	6.81	11.48	2.15	0.7	0.75	0.58
LLPACM	4.78	10.9	1.62	0.64	0.65	0.37
PCNN–NSST	6.75	11.56	2.73	0.79	0.87	0.54
SDnet	6.81	12.73	3.05	0.74	0.63	**0.64**[2]
EMFusion	6.54	11.82	2.25	0.85	0.83	0.58
U2Fusion	6.63	**12.97**[1]	2.79	0.77	0.94	0.62
SwinFusion	6.76	**12.83**[2]	3.49	**0.89**[2]	**1.06**[2]	0.61
CDDFuse	**6.85**[2]	12.11	2.28	0.88	1.05	0.6
CoConet	6.84	12.18	**3.56**[2]	0.88	1.02	**0.65**[1]
Proposed	**6.89**[1]	12.04	**3.63**[1]	**0.9**[1]	**1.07**[1]	0.61

Due to the inherently low original resolution of SPECT images, up-sampling is required before fusion, introducing a significant amount of non-uniform pixel noise. This noise interference adversely affects the results for metrics such as EN and SF, making deep learning methods perform less favorably on general evaluation metrics for SPECT-MRI modality fusion compared to traditional methods. However, our method better preserves information from both modalities, surpassing other deep learning approaches in objective analysis.

E. The Ablation Study

As the test cases, we chose 30 pairs of images at random from each modality for ablation experiments. We chose peak EN, MI, VIF, and SSIM as the metrics. The configuration of each experiment is shown in Table V.

TABLE V. THE OBJECTIVE EVALUATION OF GFP-PC FUSION IMAGES

Experiment	Configurations
Experiment 1	Replace the shared encoder STFormer with Restormer [38], keeping other parts unchanged.
Experiment 2	Not utilize context broadcasting technology.
Experiment 3	The feature extraction component in the encoder uses only the transformer module.
Experiment 4	The feature extraction component in the encoder uses only the INN module.
Experiment 5	Change the two-stage training to direct training.
Experiment 6	Removal of L_{decomp} from the loss functions used in the first phase of training.
Experiment 7	Sets γ_1 and γ_2 to {0.5, 0.5}, indicates that both images are equally important.

From Fig. 8, it can be observed that the fusion evaluation metrics using dual-branch feature extractors in the fusion module are better than those obtained by using either the CNN or the Transformer alone as the feature extractor. This indirectly confirms that dual-branched feature extractors improve the ability of the network to extract features, beneficial for subsequent bottom-level visual tasks in image fusion, and that context broadcasting technology significantly aids in improving fusion effects. In summary, Table VI proves the effectiveness and soundness of our network and loss function design.

TABLE VI. THE RESULTS OF THE ABLATION EXPERIMENT

Experiment	Evaluation metrics				
	EN	SF	MI	VIF	SSIM
Experiment 1	4.93	24.43	2.69	.0.75	1.24
Experiment 2	5.07	24.45	2.96	0.81	1.25
Experiment 3	4.86	22.19	2.21	0.85	1.21
Experiment 4	5.01	23.81	2.35	0.72	1.22
Experiment 5	4.89	22.11	2.42	0.71	1.26
Experiment 6	5.07	25.76	3.06	0.87	1.25
Experiment 7	5.06	25.83	3.07	0.85	1.26
Proposed	**5.07**	**25.83**	**3.07**	**0.87**	**1.27**

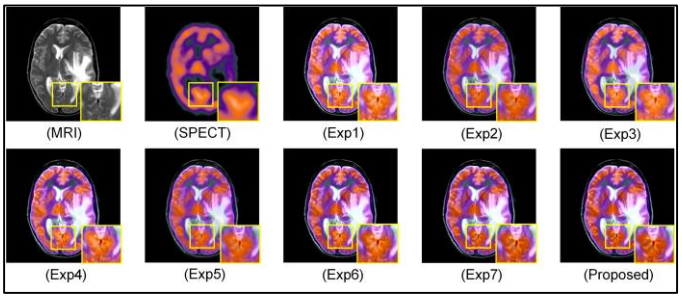

Fig. 8. The ablation experiment and the corresponding zoomed-in details of each fused image.

IV. CONCLUSION AND FUTURE WORK

This study proposes a multimodal medical image fusion network, DMMFnet. For the extraction of shared features in multimodal images, a new transformer module based on super token sampling is constructed, which effectively captures global dependencies. This module not only significantly improves the processing speed of the model but also ensures the effective capture and preservation of key features, thereby enhancing the ability to recognize and integrate important information in medical images. In addition, the proposed CSDformer module further optimizes feature extraction and fusion. By introducing the Context Broadcast strategy, the much-needed dense interaction is achieved, which greatly improves the ability to capture detailed features. Although DMMFnet has achieved satisfactory fusion results, it does not present a notable advantage in computational efficiency due to the limitations of the Transformer architecture.

Future work: Future improvements should focus on refining the network to achieve better results at a lower computational cost. Enhancements in this direction would make the DMMFnet more practical and efficient for real-world applications.

ACKNOWLEDGMENT

We extend our gratitude to the Shandong Provincial Natural Science Foundation, the Key R&D Program of Shandong Province, P.R China, and Mohammad Ali Jinnah University, Karachi, Pakistan.

RESEARCH FUNDING

This research was funded by the Shandong Provincial Natural Science Foundation through project ZR2021MF017 and

the Key R&D Program of Shandong Province, China, under project 2023RKY01015.

AUTHORS' CONTRIBUTION

In this paper, each author contributed equally to the research and development process. Yukun Zhang and Lei Wang collaborated on the conceptualization and design of the study, as well as the implementation of methodologies. Muhammad Tahir, Muhammad Imran Saeed, and Zizhen Huang were responsible for extensive data collection, analysis, and interpretation, significantly contributing to the empirical aspects of the research. Yaolong Han provided valuable insights and expertise in the theoretical framework and literature review, ensuring the study's rigor and coherence. Shanliang Yang refined the manuscript through editing and formatting, ensuring clarity and coherence in presenting the results. The authors worked collaboratively to draft and revise the manuscript, incorporating feedback and suggestions to produce the final version.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper. All authors have reviewed and approved the manuscript and have no financial or personal relationships that could inappropriately influence or bias the content of the paper.

REFERENCES

[1] 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 Transactions on Instrumentation and Measurement, vol. 68, pp. 49-64, 2018.

[2] Z. Ding, H. Li, Y. Guo, D. Zhou, and Y. Liu, "M4fnet: Multimodal medical image fusion network via multi-receptive-field and multi-scale feature integration," Computers in Biology and Medicine, vol. 159, pp. 106923, 2023.

[3] W. Li, R. Li, J. Fu, and X. Peng, "MSENet: A multi-scale enhanced network based on unique features guidance for medical image fusion," Biomedical Signal Processing and Control, vol. 74, pp. 103534, 2022.

[4] Y. Zhang, M. Jin, and G. Huang, "Medical image fusion based on improved multi-scale morphology gradient-weighted local energy and visual saliency map," Biomedical Signal Processing and Control, vol. 74, pp. 103535, 2022.

[5] G. Zhang, R. Nie, J. Cao, L. Chen, and Y. Zhu, "FDGNet: A pair feature difference guided network for multimodal medical image fusion," Biomedical Signal Processing and Control, vol. 81, pp. 104545, 2023.

[6] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 1, pp. 502-518, 2020.

[7] 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 Transactions on Multimedia, vol. 21, no. 8, pp. 1982-1996, 2019.

[8] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," Information Fusion, vol. 36, pp. 191-207, 2017.

[9] L. Chen, X. Wang, Y. Zhu, and R. Nie, "Multi-level difference information replenishment for medical image fusion," Applied Intelligence, vol. 53, pp. 4579-4591, 2023.

[10] Y. Yang, D.S. Park, S. Huang, and N. Rao, "Medical image fusion via an effective wavelet-based approach," EURASIP Journal on Advances in Signal Processing, vol. 2010, pp. 1-13, 2010.

[11] W. Wang and F. Chang, "Multi-focus image fusion method based on Laplacian pyramid," Journal of Computers, vol. 6, no. 12, pp. 2559-2566, 2011.

[12] Y. Liu, X. Chen, R.K. Ward, and Z.J. Wang, "Medical image fusion via convolutional sparsity based morphological component analysis," IEEE Signal Processing Letters, vol. 26, no. 3, pp. 485-489, 2019.

[13] Z. Wang, Y. Ma, F. Cheng, and L. Yang, "Review of pulse-coupled neural networks," Image and Vision Computing, vol. 28, no. 1, pp. 5-13, 2010.

[14] 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 Transactions on Instrumentation and Measurement, vol. 68, no. 1, pp. 49-64, 2018.

[15] W. Tan, P. Tiwari, H.M. Pandey, C. Moreira, and A.K. Jaiswal, "Multimodal medical image fusion algorithm in the era of big data," Neural Computing and Applications, pp. 1-21, 2020.

[16] B. Yang and S. Li, "Multifocus image fusion and restoration with sparse representation," IEEE Transactions on Instrumentation and Measurement, vol. 59, no. 4, pp. 884-892, 2009.

[17] B. Yang and S. Li, "Image fusion with convolutional sparse representation," IEEE Signal Processing Letters, vol. 23, no. 12, pp. 1882-1886, 2016.

[18] H. Xu and J. Ma, "EMFusion: An unsupervised enhanced medical image fusion network," Information Fusion, vol. 76, pp. 177-186, 2021.

[19] C. Wang, R. Nie, J. Cao, X. Wang, and Y. Zhang, "IGNFusion: An unsupervised information gate network for multimodal medical image fusion," IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 4, pp. 854-868, 2022.

[20] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel, "Backpropagation applied to handwritten zip code recognition," Neural Computation, vol. 1, no. 4, pp. 541-551, 1989.

[21] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A.A. Bharath, "Generative adversarial nets," Advances in Neural Information Processing Systems, vol. 27, 2014.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, 2020.

[23] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in Proceedings of the International Conference on Machine Learning (PMLR), 2021, pp. 10347-10357.

[24] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021, pp. 32-42.

[25] K. Li, Y. Wang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "Uniformer: Unified transformer for efficient spatiotemporal representation learning," arXiv preprint arXiv:2201.04676, 2022. https://doi.org/10.48550/arXiv.2201.04676.

[26] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" Advances in Neural Information Processing Systems, vol. 34, pp. 12116-12128, 2021.

[27] 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 Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021, pp. 10012-10022.

[28] 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 Transactions on Image Processing, vol. 29, pp. 4980-4995, 2020.

[29] J. Huang, Z. Le, Y. Ma, F. Fan, H. Zhang, and L. Yang, "MGMDcGAN: Medical Image Fusion Using Multi-Generator Multi-Discriminator Conditional Generative Adversarial Network," IEEE Access, vol. 99, pp. 1-1, 2020.

[30] W. Li, Y. Zhang, G. Wang, and Y. Huang, "DFENet: A dual-branch feature enhanced network integrating transformers and convolutional feature learning for multimodal medical image fusion," Biomedical Signal Processing and Control, vol. 80, pp. 104402, 2023.

[31] K. Ram Prabhakar, V. Sai Srikar, and R. Venkatesh Babu, "Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure

image pairs," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4714-4722.

[32] H. Li and X.J. Wu, "Dense-Fuse: A Fusion Approach to Infrared and Visible Images," IEEE Transactions on Image Processing, vol. 28, pp. 2614-2623, 2018.

[33] H. Li, X.J. Wu, and J. Kittler, "RFN-Nest: An end-to-end residual fusion network for infrared and visible images," Information Fusion, vol. 73, pp. 72-86, 2021.

[34] L. Jian, X. Yang, Z. Liu, G. Jeon, M. Gao, and D. Chisholm, "SEDRFuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion," IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-15, 2020.

[35] J. Cui, L. Zhou, F. Li, and Y. Zha, "Visible and infrared image fusion by invertible neural network," in China Conference on Command and Control, Singapore: Springer Nature Singapore, 2022, pp. 133-145.

[36] N. Park and S. Kim, "How do vision transformers work?" arXiv preprint arXiv:2202.06709, 2022. https://doi.org/10.48550/arXiv.2202.06709.

[37] H. Huang, X. Zhou, J. Cao, R. He, and T. Tan, "Vision transformer with super token sampling," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22690-22699.

[38] S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, and M.H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5718-5729.

[39] N. Hyeon-Woo, K. Yu-Ji, B. Heo, D. Han, S.J. Oh, and T.H. Oh, "Scratching visual transformer's back with uniform attention," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2023, pp. 5807-5818.

[40] K. A. Johnson, J. A. (n.d.) Becker, The whole brain atlas. Available: http://www.med.harvard.edu/aanlib/ home.html.

[41] R. Tsien, "The green fluorescent protein," Annu. Rev. Biochem., vol. 67, pp. 509–544, 1998.

[42] J.W. Roberts, J.A. Van Aardt, and F.B. Ahmed, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," Journal of Applied Remote Sensing, vol. 2, no. 1, pp. 023522, 2008.

[43] A.M. Eskicioglu and P.S. Fisher, "Image quality measures and their performance," IEEE Transactions on Communications, vol. 43, no. 12, pp. 2959-2965, 1995.

[44] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," Electronics Letters, vol. 38, no. 7, pp. 313–315, 2002.

[45] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, 2004.

[46] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," Information Fusion, vol. 14, no. 2, pp. 127-135, 2013.

[47] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," Information Fusion, vol. 45, pp. 153-178, 2019

[48] Y. Gao, S. Ma, J. Liu, Y. Liu, and X. Zhang, "Fusion of medical images based on salient features extraction by PSO optimized fuzzy logic in NSST domain," Biomedical Signal Processing and Control, vol. 69, pp. 102852, 2021.

[49] W. Li, J. Du, Z. Zhao, and J. Long, "Fusion of medical sensors using adaptive cloud model in local Laplacian pyramid domain," IEEE Transactions on Biomedical Engineering, vol. 66, no. 4, pp. 1172-1183, 2018.

[50] W. Tan, P. Tiwari, H.M. Pandey, C. Moreira, and A.K. Jaiswal, "Multimodal medical image fusion algorithm in the era of big data," Neural Computing and Applications, pp. 1-21, 2020.

[51] H. Zhang and J. Ma, "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion," International Journal of Computer Vision, vol. 129, no. 10, pp. 2761-2785, 2021.

[52] H. Xu and J. Ma, "EMFusion: An unsupervised enhanced medical image fusion network," Information Fusion, vol. 76, pp. 177-186, 2021.

[53] 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 Journal of Automatica Sinica, vol. 9, no. 7, pp. 1200-1217, 2022.

[54] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool, "CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5906-5916.

[55] J. Liu, R. Lin, G. Wu, R. Liu, Z. Luo, and X. Fan, "CoConet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion," International Journal of Computer Vision, pp. 1-28, 2023.