BSAFusion: A Bidirectional Stepwise Feature Alignment Network for Unaligned Medical Image Fusion

Huafeng Li[1], **Dayong Su**[1], **Qing Cai**[2*], **Yafei Zhang**[1*]

[1]School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China

[2]School of Information Science and Engineering, Ocean University of China, Qingdao 266100, China

hfchina99@163.com, dayongsu@outlook.com, cq@ouc.edu.cn, zyfeimail@163.com

Abstract

If unaligned multimodal medical images can be simultaneously aligned and fused using a single-stage approach within a unified processing framework, it will not only achieve mutual promotion of dual tasks but also help reduce the complexity of the model. However, the design of this model faces the challenge of incompatible requirements for feature fusion and alignment. To address this challenge, this paper proposes an unaligned medical image fusion method called Bidirectional Stepwise Feature Alignment and Fusion (BSFA-F) strategy. To reduce the negative impact of modality differences on cross-modal feature matching, we incorporate the Modal Discrepancy-Free Feature Representation (MDF-FR) method into BSFA-F. MDF-FR utilizes a Modality Feature Representation Head (MFRH) to integrate the global information of the input image. By injecting the information contained in MFRH of the current image into other modality images, it effectively reduces the impact of modality differences on feature alignment while preserving the complementary information carried by different images. In terms of feature alignment, BSFA-F employs a bidirectional stepwise alignment deformation field prediction strategy based on the path independence of vector displacement between two points. This strategy solves the problem of large spans and inaccurate deformation field prediction in single-step alignment. Finally, Multi-Modal Feature Fusion block achieves the fusion of aligned features. The experimental results across multiple datasets demonstrate the effectiveness of our method.

Code — https://github.com/slrl123/BSAFusion/

Introduction

Multimodal medical image fusion (MMIF) involves the integration of medical image data from different imaging modalities (such as CT, MRI, PET, etc.) to create a new image that contains more comprehensive and accurate lesion information. This technology is of great significance in improving diagnostic accuracy, assisting in the development of treatment plans, promoting medical research and education, and optimizing the utilization of medical resources. As a result, it has garnered the attention of researchers, and a multitude of effective fusion algorithms have been proposed (Li et al. 2018; Tang et al. 2022b; Mu et al. 2023). However, most current methods assume that the source images being fused are strictly aligned at the pixel level. The fusion algorithm can produce the expected results only when this assumption holds true. In real scenarios, however, this assumption is often invalid. To address this issue, registration algorithms are typically used to first align the images to be fused, followed by the fusion process (as shown in Fig.1(a)). Although this two-stage method is effective, cross-modal image registration still faces numerous challenges due to differences in modalities and inconsistencies in features between images.

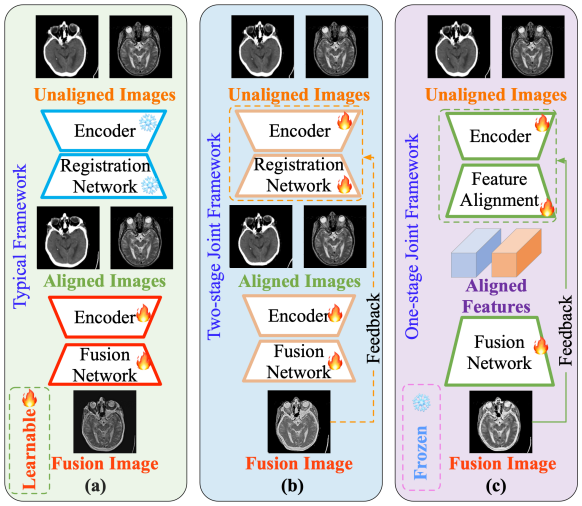


Figure 1: Paradigm of existing unaligned image fusion methods compared to that of our method.

In recent years, researchers have begun exploring the integration of multi-source image registration and fusion into a unified framework to address the aforementioned issues. By leveraging the supervision of fusion results, registration performance can be improved. Based on this idea, joint processing frameworks for image registration and fusion have emerged in recent years (Huang et al. 2022b; Wang et al. 2022; Tang et al. 2022a; Xu et al. 2022). However, these methods are not specifically designed for the registration and fusion of multimodal medical images. Although MURF (Xu, Yuan, and Ma 2023) attempts to integrate mul-

*Corresponding author.

tiple types of source image fusion problems into one framework, this multitasking approach often sacrifices performance in single-task image fusion. In view of this, PAMR-Fuse (Zhong et al. 2023) adopts a similar approach to UMF-CMGR, focusing on the fusion of unregistered multimodal medical images. However, this method depends on an image of one modality to generate a corresponding image in another modality, and its registration performance is often constrained by the quality of the generated image.

In addition, the above methods often adopt a two-stage processing mode (as shown in Fig.1(b)), with registration preceding fusion. This approach typically necessitates the use of a separate, fully-developed image registration model. This is because registration and fusion have incompatible feature requirements, making it challenging to seamlessly embed both into the fusion process through a shared feature encoder. To tackle this issue, single-stage unaligned fusion methods, such as IVFWSR (Li et al. 2024a) and RFVIF (Li et al. 2023a), have been proposed for infrared-visible image fusion. However, these methods only address feature misalignment caused by rigid transformations and are ineffective in handling elastic transformations. In fact, achieving registration and fusion of multimodal medical images using a single-stage processing mode within a joint framework remains challenging. These challenges mainly involve resolving the conflicting requirements of feature extraction for registration and fusion. Typically, feature fusion expects the features to be complementary, while feature matching demands consistency between corresponding features. To achieve simultaneous feature alignment and fusion within a single-stage processing mode, it is crucial to address the aforementioned issues.

Therefore, this paper proposes a single-stage framework for multimodal medical image registration and fusion (as shown in Fig. 1(c)). Unlike traditional two-stage methods, this approach does not require a separate and complete registration process. Instead, it seamlessly embeds the registration steps into the image fusion process, effectively mitigating the increase in model complexity that would result from introducing multiple independent feature encoders. Technically, we innovatively develop an unaligned medical image fusion method called Bidirectional Stepwise Feature Alignment (BSFA). To effectively mitigate the adverse effects of modality differences on cross-modal feature matching, we integrate the Modality Discrepancy-Free Feature Representation (MDF-FR) method into the BSFA framework. MDF-FR achieves global feature integration by appending a Modality Feature Representation Head (MFRH) to each input image. This method significantly reduces the impact of modality differences and inconsistent multimodal information on feature alignment by injecting the head information of the current image into the features of the other images to be fused. As a result, this design effectively preserves the complementary information carried by different images, ensuring both the integrity and diversity of the data. For feature alignment, we propose a bidirectional stepwise deformation field prediction strategy based on the path independence of vector displacement between two points. This strategy effectively addresses the challenges of large-span and inaccurate deformation field predictions encountered in traditional single-step alignment methods, significantly enhancing both the accuracy and efficiency of feature alignment. Finally, through the Multi-Modal Feature Fusion (MMFF) module, the predicted deformation field is applied to multimodal features, achieving precise alignment and effective fusion of input images at the feature level. Overall, the contributions of this paper can be summarized as follows:

- We design a joint implementation framework that integrates feature cross-modal alignment and fusion. By sharing a single feature encoder, it enables the seamless integration of registration and fusion, effectively avoiding the increase in model complexity that would result from introducing additional encoder for registration.
- We propose a modality discrepancy reduction method. This method achieves global feature integration by appending an MFRH to each input image. By incorporating the feature representation of the current image into the features of the other images to be fused, it effectively mitigates the impact of modality differences on feature alignment.
- Based on the path independence of vector displacement between two points, a bidirectional stepwise deformation field prediction strategy is proposed. It effectively addresses the challenges of large spans and inaccurate deformation field predictions encountered in traditional single-step alignment methods.

Related Work

For MMIF, deep learning has been widely used due to its ability to effectively extract statistical information from large datasets. Based on the types of feature extraction networks used, existing medical image fusion methods can be classified into CNN-based, Transformer-based, and hybrid methods. Among them, CNN-based methods mainly focus on network architecture design. Commonly used frameworks for MMIF in CNN-based methods include residual connections (Gu et al. 2024), skip connections (Di et al. 2024), dense connections (Zuo, Zhang, and Yang 2021), and Network Architecture Search (Ye et al. 2023). Additionally, there are dynamic meta-learning method (Huang et al. 2022a) and medical semantic-guided two-branch method (Wen et al. 2023). However, these methods are often limited by the shortcomings of CNNs in modeling long-distance dependencies. Transformer (Vaswani et al. 2017) has addressed this limitation, resulting in methods such as FATMusic (Zhao et al. 2023b) and MATR (Tang et al. 2022b). Given the complementary strengths of CNN and Transformer in feature extraction (Li et al. 2024b), researchers have proposed hybrid methods like DesTrans (Song et al. 2024), DFENet (Li et al. 2023b) and MRSC-Fusion (Xie et al. 2023).

In recent years, various methods have been developed for multimodel image fusion, including MMIF, such as U2Fusion (Xu et al. 2020), Cddfuse (Zhao et al. 2023a), DDFM (Zhao et al. 2023b), EMMA (Zhao et al. 2024), QuadzBayer(Zheng et al. 2024) and HFT(Chen et al. 2023).

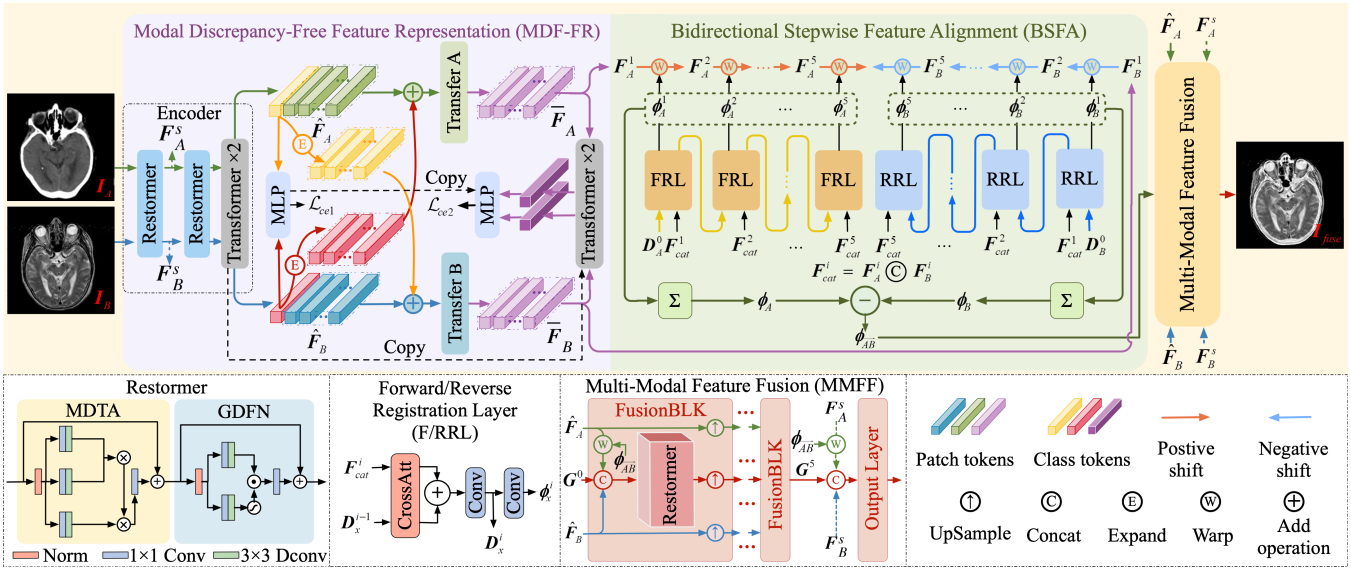


Figure 2: Overall framework of the proposed method. The unaligned multimodal medical image pairs $\{\boldsymbol{I}_A, \boldsymbol{I}_B\}$ are processed through the MDF-FR module, yielding features $\{\boldsymbol{F}_A^s, \boldsymbol{F}_B^s\}$ and $\{\hat{\boldsymbol{F}}_A, \hat{\boldsymbol{F}}_B\}$. Additionally, modality-specific feature representation heads, denoted as $\hat{\boldsymbol{f}}_A$ and $\hat{\boldsymbol{f}}_B$, are generated. These heads are utilized to minimize the modality disparities between $\{\hat{\boldsymbol{F}}_A, \hat{\boldsymbol{F}}_B\}$. Within the BSFA, a progressive deformation field prediction, denoted as $\phi_{\overrightarrow{AB}}$, is carried out based on the modality-discrepancy-mitigated features $\{\bar{\boldsymbol{F}}_A, \bar{\boldsymbol{F}}_B\}$. Finally, the features $\{\bar{\boldsymbol{F}}_A, \bar{\boldsymbol{F}}_B\}$, $\{\boldsymbol{F}_A^s, \boldsymbol{F}_B^s\}$, along with the predicted deformation field $\phi_{\overrightarrow{AB}}$, are fed into the MMFF module to generate the final fused result.

Although these methods are effective, they all assume that the source images to be fused have already been registered. However, in practical applications, this assumption often does not hold true. Therefore, when dealing with unaligned multimodal medical images, these methods cannot be directly applied and require additional image registration models to align the images for fusion. This not only increases model complexity, hindering deployment in computationally constrained environments, but also results in fusion failures if the registration model fails. To address these issues, methods for joint registration and fusion have been developed. Typical examples include ReCoNet (Huang et al. 2022b), UMF-CMGR (Wang et al. 2022), SuperFusion (Tang et al. 2022a), as well as others like RFNet (Xu et al. 2022), MURF (Xu, Yuan, and Ma 2023), IVFWSR (Li et al. 2024a), and MERF (Hong, Zhang, and Ma 2024). However, these methods are not specifically designed for multimodal medical images and do not exhibit the expected advantages in this domain. Although PAMRFuse (Zhong et al. 2023) is designed for MMIF, its performance is limited by the quality of the generated images. To overcome these challenges, we propose an unaligned MMIF scheme that integrates registration and fusion, allowing the two tasks to complement each other within a single-stage framework.

Proposed Method

Overview

As shown in Fig. 2, the proposed method consists of three core components: MDF-FR, BSFA, and MMFF. The goal of MDF-FR is to eliminate modality discrepancies between unaligned multimodal medical image pairs $\boldsymbol{I}_A, \boldsymbol{I}_B$. To address the conflicting requirements of feature alignment and feature fusion, we introduce an MFRH for each input image within MDF-FR. This mechanism reduces the impact of modality discrepancies on feature alignment by injecting MFRH of the current image into the features of other modality image. BSFA predicts the deformation field between features of unaligned images, facilitating subsequent alignment. To tackle challenges posed by large displacements and difficult deformation field predictions inherent in unidirectional methods, BSFA employs a bidirectional, gradually aligned deformation field prediction strategy based on the path independence of vector displacement between two points. Finally, MMFF aligns the features by applying the predicted deformation field and then constructs the fused image based on the aligned features.

Modality Discrepancy-Free Feature Representation

In MDF-FR, we utilize a network consisting of Restormer and Transformer layers as the encoder for extracting features from unaligned image pairs $\{\boldsymbol{I}_A, \boldsymbol{I}_B\}$. The structure of the Restormer is exhibited in Fig. 2 (Zamir et al. 2022). For input images $\boldsymbol{I}_A$ and $\boldsymbol{I}_B$, the features output by the first Restormer layer are denoted as $\boldsymbol{F}_A^s \in \mathbb{R}^{C\times H\times W}$ and $\boldsymbol{F}_B^s \in \mathbb{R}^{C\times H\times W}$, respectively, where C, H, and W represent the number of channels, height, and width of the feature maps. Since the shallow features $\boldsymbol{F}_A^s$ and $\boldsymbol{F}_B^s$ contain the underlying details of the image, we directly feed them into the multimodal feature fusion layer for feature alignment and fusion

to retain more edge details in the fusion results. The features output from the second Restormer layer are then fed into two Transformer layers to extract the features $\bar{\boldsymbol{F}}_A$ and $\bar{\boldsymbol{F}}_B$, which are used for modality discrepancy elimination and deformation field prediction. The results output by the Transformer layers are denoted as $\hat{\boldsymbol{F}}_A = [\hat{\boldsymbol{f}}_A^1, \hat{\boldsymbol{f}}_A^2, \cdots, \hat{\boldsymbol{f}}_A^P] \in \mathbb{R}^{P \times W'}$ and $\hat{\boldsymbol{F}}_B = [\hat{\boldsymbol{f}}_B^1, \hat{\boldsymbol{f}}_B^2, \cdots, \hat{\boldsymbol{f}}_B^P] \in \mathbb{R}^{P \times W'}$, along with the modality feature representation heads $\hat{\boldsymbol{f}}_A \in \mathbb{R}^{1 \times W'}$ and $\hat{\boldsymbol{f}}_B \in \mathbb{R}^{1 \times W'}$, where P is the total number of patches that the feature output by the second Restormer layer is divided into, and W' is the length of the vector. In the proposed method, $\hat{\boldsymbol{f}}_A$ and $\hat{\boldsymbol{f}}_B$ are used to describe the modality categories of the input images. To ensure that $\hat{\boldsymbol{f}}_A$ and $\hat{\boldsymbol{f}}_B$ contain the modal information of the input images, we feed them into an MLP, with the output aiming to minimize the cross-entropy loss defined in Eq. (1):

$$\mathcal{L}_{ce1} = CE\left(\boldsymbol{y}_A, [0, 1]\right) + CE\left(\boldsymbol{y}_B, [1, 0]\right) \tag{1}$$

where CE represents cross entropy, and $\boldsymbol{y}_A$ and $\boldsymbol{y}_B$ represent the results predicted by the MLP.

Due to the significant modality discrepancies between $\hat{\boldsymbol{F}}_A$ and $\hat{\boldsymbol{F}}_B$, cross-modal matching and deformation field prediction based on these features face substantial challenges. To address this issue, existing methods typically extract shared modality features directly from $\hat{\boldsymbol{F}}_A$ and $\hat{\boldsymbol{F}}_B$ for deformation field prediction. Although this approach is effective, it may lead to the loss of non-shared or modality-specific information, thereby reducing the expressive power of the features. In contrast, if we directly inject the corresponding modality information into $\hat{\boldsymbol{F}}_A$ and $\hat{\boldsymbol{F}}_B$, we can not only mitigate the impact of modality discrepancies on deformation field prediction but also prevent the loss of non-shared information caused by extracting shared information. The modal feature representation heads obtained by minimizing the loss in Eq.(1), which represent global features $\hat{\boldsymbol{f}}_A$ and $\hat{\boldsymbol{f}}_B$, are not affected by misaligned source images. Therefore, they can be directly injected into $\hat{\boldsymbol{F}}_A$ and $\hat{\boldsymbol{F}}_B$ to reduce the modality discrepancies between the two:

$$\begin{aligned} \tilde{\boldsymbol{F}}_A &= [\hat{\boldsymbol{f}}_A^1 + \hat{\boldsymbol{f}}_B, \hat{\boldsymbol{f}}_A^2 + \hat{\boldsymbol{f}}_B, \cdots \hat{\boldsymbol{f}}_A^P + \hat{\boldsymbol{f}}_B] \\ \tilde{\boldsymbol{F}}_B &= [\hat{\boldsymbol{f}}_B^1 + \hat{\boldsymbol{f}}_A, \hat{\boldsymbol{f}}_B^2 + \hat{\boldsymbol{f}}_A, \cdots \hat{\boldsymbol{f}}_B^P + \hat{\boldsymbol{f}}_A] \end{aligned} \tag{2}$$

To ensure that the features processed by Eq. (2) effectively eliminate differences between modalities, we use two Transfer blocks, namely TransferA and TransferB. Each block is composed of two Transformer layers, and the parameters are not shared between the blocks, allowing for further extraction of features $\bar{\boldsymbol{F}}_A$ and $\bar{\boldsymbol{F}}_B$ necessary for predicting the deformation field. Additionally, to determine whether the features output by TransferA and TransferB exhibit modality specificity, we replicate the Transformer layer in the encoder and the MLP behind the encoder. We then sequentially pass the features $\bar{\boldsymbol{F}}_A$ and $\bar{\boldsymbol{F}}_B$ through the Transformer layer and the MLP, respectively, to identify the modality category of features $\bar{\boldsymbol{F}}_A$ and $\bar{\boldsymbol{F}}_B$. To achieve this, we utilize the cross-entropy loss function in Eq. (3) to update the parameters in TransferA and TransferB:

$$\mathcal{L}_{ce2} = CE\left(\boldsymbol{y}_A^*, [0.5, 0.5]\right) + CE\left(\boldsymbol{y}_B^*, [0.5, 0.5]\right) \tag{3}$$

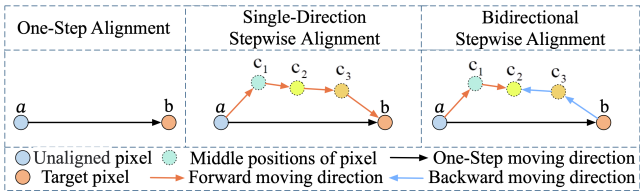


Figure 3: Schematic of different alignment methods.

where $\boldsymbol{y}_A^*$ and $\boldsymbol{y}_B^*$ are the predicted results of the MLP for the modal categories of $\bar{\boldsymbol{F}}_A$ and $\bar{\boldsymbol{F}}_B$. In this process, we introduce TransferA and TransferB to further extract features from $\bar{\boldsymbol{F}}_A$ and $\bar{\boldsymbol{F}}_B$ that are helpful for deformation field prediction. After processing with TransferA and TransferB, we directly reuse the Transformer layer and the MLP to ensure that the features $\bar{\boldsymbol{F}}_A$ and $\bar{\boldsymbol{F}}_B$ used for deformation field prediction no longer exhibit modality discrepancies.

Feature Alignment

Path independence of vector displacement: Since $\bar{\boldsymbol{F}}_A$ and $\bar{\boldsymbol{F}}_B$ have effectively eliminated the modality discrepancies under the constraint of Eq. (3), it is suitable to predict the deformation field between the features of the input images $\boldsymbol{I}_A$ and $\boldsymbol{I}_B$ using $\bar{\boldsymbol{F}}_A$ and $\bar{\boldsymbol{F}}_B$. However, when the offset between $\boldsymbol{I}_A$ and $\boldsymbol{I}_B$ is large, directly predicting the deformation field between them becomes challenging. As shown in Fig. 3, during the pixel alignment process, pixel a moves to the location of pixel b. This can be achieved by applying the deformation field between pixels a and b to a. The deformation field between pixels a and b, which includes the direction and distance of spatial movement from the location of a to that of b, can be regarded as a vector, denoted as $\overrightarrow{ab}$. From Fig. 3, it can be seen that the alignment between the locations of a and b can be achieved in a single step, directly moving from the location of a to that of b, or by passing through intermediate points c_1, c_2, and c_3 to reach the location of b.

Due to

$$\overrightarrow{ab} = \overrightarrow{ac_1} + \overrightarrow{c_1c_2} + \overrightarrow{c_2c_3} + \overrightarrow{c_3b}, \tag{4}$$

this indicates that the deformation field from point a to point b can be accumulated from the deformation fields from a to c_1, c_1 to c_2, c_2 to c_3, and c_3 to b. This demonstrates that the vector from point a to point b is independent of the path taken from point a to point b. In this paper, we refer to this characteristic as the path independence of vector displacement between two points. The deformation field prediction method based on this theory is called the unidirectional progressive prediction method.

As our goal is to align points a and b in their spatial locations, this can be achieved by simultaneously moving both a and b towards an intermediate location. The bidirectional progressive alignment, as shown in Fig.2, allows point a to reach the intermediate point c_2 through point c_1, while point b can reach c_2 through point c_3 in the opposite direction, thereby achieving feature alignment at the location of point c_2. At this point, the deformation field used by pixel a to move to the position of pixel b can be expressed as a vector:

$$\overrightarrow{ab} = \overrightarrow{ac_2} - \overrightarrow{bc_2} = (\overrightarrow{ac_1} + \overrightarrow{c_1c_2}) - (\overrightarrow{bc_3} + \overrightarrow{c_3c_2}) \tag{5}$$

This bidirectional alignment method effectively captures the interrelationships between images and reduces cumulative errors. Furthermore, if alignment in one direction encounters issues, alignment in the opposite direction can compensate, thereby enhancing the overall robustness of the alignment process. Based on these considerations, this paper proposes the Bidirectional Stepwise Feature Alignment method.

Bidirectional Stepwise Feature Alignment: As shown in Fig. 2, the proposed BSFA predicts the deformation fields of the input image features $\bar{\boldsymbol{F}}_A$ and $\bar{\boldsymbol{F}}_B$ from two directions. Both the forward and reverse predictions of BSFA involve five layers of deformation field prediction operations, corresponding to the insertion of five intermediate nodes between the two input source images, achieving spatial alignment of their features by the fifth layer. In our method, the modules responsible for predicting the forward and reverse deformation fields are referred to as the Forward Registration Layer (FRL) and the Reverse Registration Layer (RRL), respectively, as illustrated in Fig. 2. For the initial FRL, its inputs are $\boldsymbol{F}_{cat}^1 = \text{concat}(\boldsymbol{F}_A^1, \boldsymbol{F}_B^1)$ and $\boldsymbol{D}_A^0 \in \mathbb{R}^{W\times\frac{H}{2}\times\frac{H}{2}}$, and the outputs are the deformation field $\boldsymbol{\phi}_A^1$ and $\boldsymbol{D}_A^1$. Here, $\boldsymbol{F}_A^1 \in \mathbb{R}^{W\times\frac{H}{2}\times\frac{H}{2}}$ and $\boldsymbol{F}_B^1 \in \mathbb{R}^{W\times\frac{H}{2}\times\frac{H}{2}}$ denote the reshaped versions of $\bar{\boldsymbol{F}}_A$ and $\bar{\boldsymbol{F}}_B$, $\boldsymbol{D}_A^0$ is initialized to $\boldsymbol{F}_A^1$, and K represents the number of layers in both FRL and RRL. The first RRL takes $\boldsymbol{F}_{cat}^1$ and $\boldsymbol{D}_B^0$ as inputs, and its outputs include the deformation field $\boldsymbol{\phi}_B^1$ and $\boldsymbol{D}_B^1$, with $\boldsymbol{D}_B^0$ initialized to $\boldsymbol{F}_B^1$. In the i-th FRL and i-th RRL, their inputs are $\left\{\boldsymbol{F}_{cat}^i, \boldsymbol{D}_A^{i-1}\right\}$ and $\left\{\boldsymbol{F}_{cat}^i, \boldsymbol{D}_B^{i-1}\right\}$, respectively, where

$$\begin{aligned} \boldsymbol{F}_{\text{cat}}^i &= \text{concat}\left(\boldsymbol{F}_A^i, \boldsymbol{F}_B^i\right) \\ \boldsymbol{F}_j^i &= \uparrow_{\times 2}\left(\boldsymbol{W}\left(\boldsymbol{F}_j^{i-1} \mid \boldsymbol{\phi}_j^{i-1}\right)\right), \quad (j = A, B) \end{aligned} \tag{6}$$

In Eq. (6), $\boldsymbol{W}$ represents the Warp operation (Jaderberg et al. 2015), which adjusts the spatial position of pixels according to the deformation field $\boldsymbol{\phi}_A^{i-1}$. The symbol "$\uparrow_{\times 2}$" denotes a $2\times$ upsampling operation.

After correcting the input image features $\boldsymbol{F}_a^i$ and $\boldsymbol{F}_b^i$ using the predicted deformation fields in both directions, cross-modal alignment of the features is achieved at an intermediate location. However, the progressive alignment process causes features to gradually move from their original positions. Directly fusing the intermediate aligned features does not allow the input source images to guide the fusion process effectively, which can hinder the improvement of fusion quality. Based on the path independence of vector displacement between two points mentioned earlier, we can construct a transformation that directly aligns the features of source image $\boldsymbol{I}_a$ with those of source image $\boldsymbol{I}_b$, using the predicted deformation field at each stage. According to the principle in Eq. (5), the deformation field that achieves the alignment of $\boldsymbol{I}_a$ and $\boldsymbol{I}_b$ features can be expressed as:

$$\begin{aligned} &\boldsymbol{\phi}_A = \sum_{i=1}^{K} \uparrow_2 \left(2^{K-i}\boldsymbol{\phi}_A^i\right), \boldsymbol{\phi}_B = \sum_{i=1}^{K} \uparrow_2 \left(2^{K-i}\boldsymbol{\phi}_B^i\right) \\ &\boldsymbol{\phi}_{\overrightarrow{AB}} = \boldsymbol{\phi}_A - \boldsymbol{\phi}_B \end{aligned} \tag{7}$$

To ensure the quality of the deformation field, we introduce smoothing loss $\mathcal{L}_{smooth}$:

$$\mathcal{L}_{smooth} = \sum_{i=1}^{K} 10^{i-K}\left(\|\nabla\boldsymbol{\phi}_A^i\|_2 + \|\nabla\boldsymbol{\phi}_B^i\|_2\right) \tag{8}$$

and consistency loss $\mathcal{L}_{consis}$ for model updating:

$$\mathcal{L}_{} = \mathcal{L}\left(\boldsymbol{I}, \boldsymbol{W}(\boldsymbol{I}, \boldsymbol{\phi})\right) + \|\boldsymbol{I} - \boldsymbol{W}(\boldsymbol{I}, \boldsymbol{\phi})\| \tag{9}$$

where $\mathcal{L}_{ssim}$ represents structural similarity (SSIM) loss, $\boldsymbol{I}_A'$ is the label image after $\boldsymbol{I}_A$ is strictly aligned to $\boldsymbol{I}_B$ at the pixel level.

Multimodal Feature Fusion

As shown in Fig. 2, after obtaining $\boldsymbol{\phi}_{\overrightarrow{AB}}$, we send it along with features $\boldsymbol{F}_A^s$ and $\boldsymbol{F}_B^s$, as well as features $\hat{\boldsymbol{F}}_A$ and $\hat{\boldsymbol{F}}_B$, to FusionBLK for feature fusion processing. We first transform the features $\hat{\boldsymbol{F}}_A \in \mathbb{R}^{P\times W}$ and $\hat{\boldsymbol{F}}_B \in \mathbb{R}^{P\times W}$ into $\hat{\boldsymbol{F}}_A^r \in \mathbb{R}^{W\times\frac{H}{2}\times\frac{H}{2}}$ and $\hat{\boldsymbol{F}}_B^r \in \mathbb{R}^{W\times\frac{H}{2}\times\frac{H}{2}}$, and send them to FusionBLK, where J is the number of FusionBLKs. For the i-th FusionBLK, its inputs are $\hat{\boldsymbol{F}}_1^r$, $\hat{\boldsymbol{F}}_B^r$, and $\boldsymbol{G}^{i-1}$, and the output is $\boldsymbol{G}^i$:

$$\begin{aligned} \boldsymbol{G}^i = \uparrow_{\times 2} \boldsymbol{E}_r\Bigg(&\text{concat}\left(\boldsymbol{W}\left(\uparrow_{\times 2}\left(\hat{\boldsymbol{F}}_A^r\right)\Big|_{\downarrow_{\times 2}}\left(\frac{\boldsymbol{\phi}_{\overrightarrow{AB}}}{2^{J-i}}\right)\right),\right. \\ &\left.\uparrow_{\times 2}\left(\hat{\boldsymbol{F}}_B^r\right), \boldsymbol{G}^{i-1}\right)\Bigg) \end{aligned} \tag{10}$$

where $\boldsymbol{E}_r$ represents the encoder composed of Restormers, and $\boldsymbol{G}^0$ is the zero matrix when $i = 1$. The output of the last layer, i.e., the J-th FusionBLK, is denoted as $\boldsymbol{G}^J$. Subsequently, we concatenate $\boldsymbol{G}^J$ and $\boldsymbol{F}_B^s$ with the corrected result $\tilde{\boldsymbol{F}}_A^s = \boldsymbol{W}(\boldsymbol{F}_A^s \mid \boldsymbol{\phi}_{\overrightarrow{AB}})$ of $\boldsymbol{F}_A^s$, and send the concatenated result to a reconstruction layer composed of a Restormer, a convolutional layer, and a sigmoid activation function to obtain the fused image $\boldsymbol{I}_{fuse}$. The entire fusion process is depicted as "MMFF" in Fig.2, where it represents the sequence of steps involved in the fusion.

To ensure structural consistency between the fused image and the source images, we use structure loss $\mathcal{L}_{struct}$ to optimize the network parameters:

$$\mathcal{L}_{struct} = \mathcal{L}_{ssim}\left(\boldsymbol{I}_{fuse}, \tilde{\boldsymbol{I}}_A\right) + \mu\,\mathcal{L}_{ssim}\left(\boldsymbol{I}_{fuse}, \boldsymbol{I}_B\right) \tag{11}$$

where $\tilde{\boldsymbol{I}}_A = \boldsymbol{W}(\boldsymbol{I}_A \mid \boldsymbol{\phi}_{\overrightarrow{AB}})$, and μ is a hyperparameter used to adjust the influence of the two SSIM losses in $\mathcal{L}_{struct}$. To ensure good contrast in the fusion results, we introduce a pixel intensity loss:

$$\mathcal{L}_{inten} = \left\|\boldsymbol{I}_{fuse} - \max(\tilde{\boldsymbol{I}}_A, \boldsymbol{I}_B)\right\|_1 \tag{12}$$

At the same time, gradient loss is also introduced to prevent the loss of edge details in the source image:

$$\mathcal{L}_{grad} = \left\|\nabla\boldsymbol{I}_{fuse} - \max(\nabla\tilde{\boldsymbol{I}}_A, \nabla\boldsymbol{I}_B)\right\|_1 \tag{13}$$

Therefore, the total loss of this approach is:

$$\begin{aligned} \mathcal{L}_{total} = &\mathcal{L}_{ce1} + \mathcal{L}_{ce2} + \mathcal{L}_{consis} + \mathcal{L}_{smooth} \\ &+ \mathcal{L}_{struct} + \mathcal{L}_{grad} + \lambda\mathcal{L}_{inten} \end{aligned} \tag{14}$$

where λ is a hyperparameter used to adjust the contribution of $\mathcal{L}_{inten}$ to the total loss.

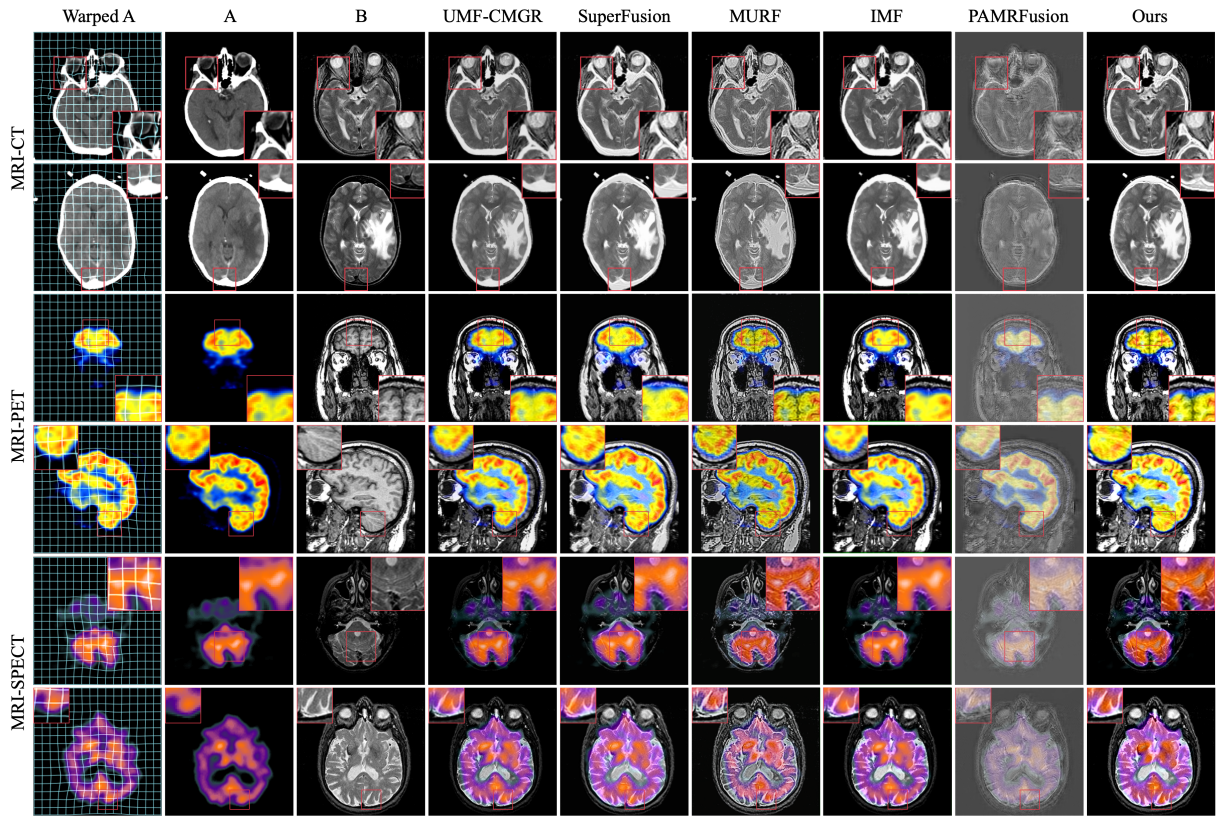


Figure 4: Visual Comparison of Fusion Results: Joint Registration and Fusion Method vs. Our Method. The first column shows the deformed image to be fused, the second column displays the corresponding label, and the third column presents the MRI image to be fused. Columns 4 to 9 show the results obtained by different fusion methods.

Experiments

Experimental Setup

Dataset and Implementation Details: We follow the protocols of existing methods (Huang et al. 2024; Wen et al. 2023) and train the model using CT-MRI, PET-MRI, and SPECT-MRI datasets from Harvard[1]. These datasets consist of 144, 194, and 261 strictly registered image pairs, respectively, each with a size of 256×256. To simulate misaligned image pairs as collected in real-world scenarios, we designate the MRI images as the reference and apply a mixture of rigid and non-rigid deformations to the non-MRI images, thereby creating the required training set. Additionally, the same deformations are applied to 20, 55, and 77 strictly registered image pairs to construct an unaligned test set. To augment the data, we apply these mixed deformations randomly in each epoch, along with random rotations and flips to increase the diversity of the training samples.

During the training process, we adopt an end-to-end approach, training for 3,000 epochs on each dataset with a batch size of 32. We use the Adam optimizer (Jimmy Ba 2015) to update model parameters, starting with an initial learning rate of 5×10^{-5}. The learning rate is dynamically adjusted using a Cosine Annealing Learning Rate (LR) scheduler (Loshchilov and Hutter 2017), decreasing to 5×10^{-7} over time. Two hyperparameters are set in the loss function: λ is set to 0.5, and μ is updated after each epoch based on the fusion results and the SSIM between the two source images, calculated as $\mu = \sum_{n=1}^{N} \mathcal{L}_{ssim}^{(n)}(\boldsymbol{I}_{fuse}, \boldsymbol{I}_B) / \sum_{n=1}^{N} \mathcal{L}_{ssim}^{(n)}(\boldsymbol{I}_{fuse}, \tilde{\boldsymbol{I}}_A)$, where N is the number of training samples in each epoch. The proposed method is implemented using the PyTorch framework and trained on a single NVIDIA GeForce RTX 4090 GPU.

Evaluation Metrics: To objectively evaluate the performance of fusion methods, we selected five commonly used image quality metrics: Gradient-based Fusion Performance ($Q_{AB/F}$) (Xydeas, Petrovic et al. 2000), Chen-Varshney Metric (Q_{CV}) (Chen and Varshney 2007), Visual Information Fidelity (Q_{VIF}) (Han et al. 2013), Structure-based Metric (Q_S) (Piella and Heijmans 2003), and Structural Similarity Index Measure (Q_{SSIM}) (Wang et al. 2004). Among these evaluation metrics, a lower value of Q_{CV} indicates better quality of the fused image, while higher values of the other metrics indicate better fusion quality.

Comparison With the State-of-the-art Methods

The common approach to solving the problem of unaligned multi-source image fusion is to first perform registration on

[1]http://www.med.harvard.edu/aanlib/

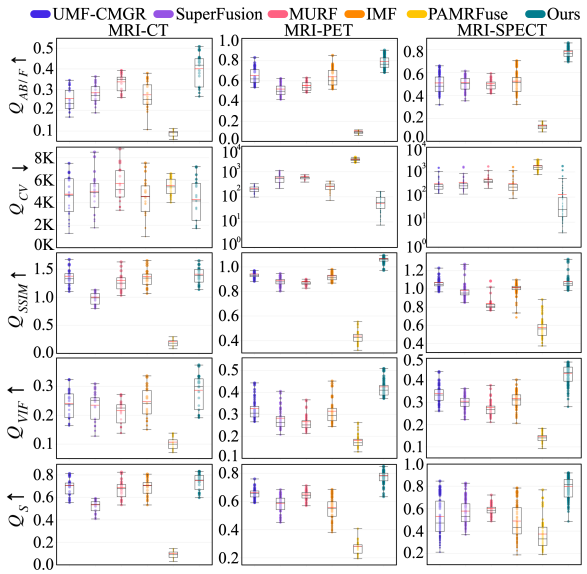
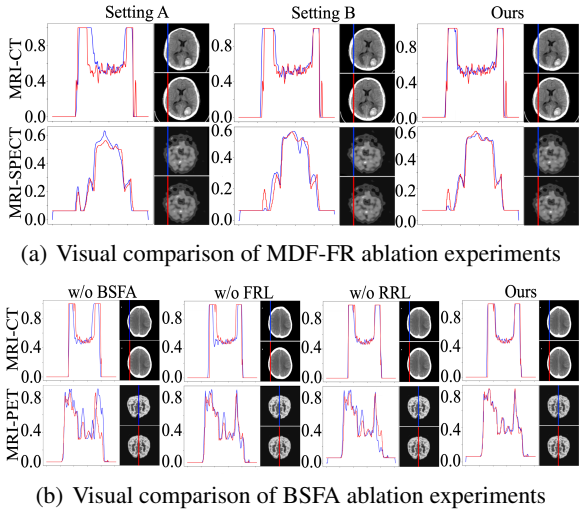


Figure 5: Comparison of objective evaluation results: Joint registration and fusion vs. the proposed method. The black line denotes the median, and the red line denotes the mean.

the images to be fused, and then fuse them. We refer to this method as "Registration+Fusion." In addition to conventional methods, a two-stage approach has been developed that combines registration and fusion into a single process, referred to as "Joint Registration and Fusion." To verify the superiority of our method, we compare it with these two approaches. Since our method is most closely related to "Joint Registration and Fusion," we focus on the comparison with this method here. Due to space limitations, the comparison between our method and the "Registration+Fusion" method is included in the supplementary materials[2].

Specifically, we compared the performance of our method with five joint registration and fusion methods: UMF-CMGR, SuperFusion, MURF, IMF (Wang et al. 2024), and PAMRFuse. The first four methods are specifically designed for the registration and fusion of multimodal images and are applicable to MMIF. PAMRFuse, on the other hand, is a method specifically proposed for the fusion of unregistered medical images. Fig. 4 presents a visual comparison of the fusion results generated by different methods. It is evident that our proposed method demonstrates significant advantages in feature alignment, contrast preservation, and detail retention. This indicates that, compared to existing two-stage joint processing frameworks, our method exhibits stronger performance, primarily due to its ability to seamlessly integrate registration and fusion tasks into a unified process. Additionally, we created box plots of test metrics for each method to visually analyze performance differences. As shown in Fig. 5, our method achieved the best mean performance across all metrics. Compared to the IMF method, which also demonstrates excellent performance, our bidirectional alignment strategy yielded significantly better fusion results, outperforming IMF's unidirectional alignment strategy.

[2]https://arxiv.org/abs/2412.08050



(a) Visual comparison of MDF-FR ablation experiments

(b) Visual comparison of BSFA ablation experiments

Figure 6: Ablation experiments of MDF-FR and BSFA. The red curve represents the pixel values of the label, and the blue curve shows the pixel values after registration.

Ablation Study

Effectiveness of MDF-FR: To verify the effectiveness of MDF-FR, we designed Setting A and Setting B. In Setting A, neither MFRH swapping nor the $\mathcal{L}_{ce2}$ was used. In contrast, Setting B did not involve MFRH swapping but did use the $\mathcal{L}_{ce2}$. The experimental results, as shown in Fig. 6(a), indicate that the proposed method achieves better alignment and fusion performance when MDF-FR is included.

Effectiveness of BSFA: To verify the effectiveness of BSFA, we conducted several experiments. First, we removed BSFA entirely to assess its impact on overall alignment performance and fusion results. Next, we removed the FRL from BSFA, retaining only the RRL. Finally, we kept only the FRL and removed the RRL. The alignment results shown in Fig. 6(b) indicate that the proposed method achieves excellent alignment only when BSFA is fully implemented.

Conclusion

This paper presents a one-stage multimodal medical image registration and fusion framework. Unlike traditional two-stage methods, it reduces model complexity with a shared feature encoder. By incorporating MDF-FR, the framework addresses modal differences in cross-modal feature alignment. The MFRH for each input integrates global image features, retaining complementary information across modalities. Additionally, a bidirectional stepwise alignment strategy predicts deformation fields using vector displacement principles. The method preserves fused information's integrity and diversity and shows potential for clinical applications requiring precise and efficient registration and fusion.

Acknowledgments

This work was supported in part by the National Science Foundation of China under Grant 62161015, Grant 62471448, and Grant 62102338; in part by the Yunnan Fundamental Research Projects under Grant 202301AV070004; in part by Shandong Provincial Natural Science Foundation under Grant ZR2024YQ004; and in part by TaiShan Scholars Youth Expert Program of Shandong Province under Grant No.tsqn202312109.

References

Chen, H.; and Varshney, P. K. 2007. A human perception inspired quality metric for image fusion based on regional information. *Information fusion*, 8(2): 193–207.

Chen, R.; Zheng, B.; Zhang, H.; Chen, Q.; Yan, C.; Slabaugh, G.; and Yuan, S. 2023. Improving dynamic hdr imaging with fusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 340–349.

Di, J.; Guo, W.; Liu, J.; Ren, L.; and Lian, J. 2024. AMM-Net: A multimodal medical image fusion method based on an attention mechanism and MobileNetV3. *Biomedical Signal Processing and Control*, 96: 106561.

Gu, X.; Wang, L.; Deng, Z.; Cao, Y.; Huang, X.; and Zhu, Y.-m. 2024. Adaptive spatial and frequency experts fusion network for medical image fusion. *Biomedical Signal Processing and Control*, 96: 106478.

Han, Y.; Cai, Y.; Cao, Y.; and Xu, X. 2013. A new image fusion performance metric based on visual information fidelity. *Information Fusion*, 14(2): 127–135.

Hong, W.; Zhang, H.; and Ma, J. 2024. MERF: A Practical HDR-Like Image Generator via Mutual-Guided Learning Between Multi-Exposure Registration and Fusion. *IEEE Transactions on Image Processing*, 33: 2361–2376.

Huang, J.; Li, X.; Tan, H.; and Cheng, X. 2024. Generative Adversarial Network for Trimodal Medical Image Fusion using Primitive Relationship Reasoning. *IEEE Journal of Biomedical and Health Informatics*, 1–13.

Huang, W.; Zhang, H.; Quan, X.; and Wang, J. 2022a. A two-level dynamic adaptive network for medical image fusion. *IEEE Transactions on Instrumentation and Measurement*, 71: 1–17.

Huang, Z.; Liu, J.; Fan, X.; Liu, R.; Zhong, W.; and Luo, Z. 2022b. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *European conference on computer Vision (ECCV)*, 539–555. Springer.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2017–2025.

Jimmy Ba, D. K. 2015. Adam: a method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Li, H.; He, X.; Tao, D.; Tang, Y.; and Wang, R. 2018. Joint medical image fusion, denoising and enhancement via discriminative low-rank sparse dictionaries learning. *Pattern Recognition*, 79: 130–146.

Li, H.; Liu, J.; Zhang, Y.; and Liu, Y. 2024a. A Deep Learning Framework for Infrared and Visible Image Fusion Without Strict Registration. *International Journal of Computer Vision*, 132: 1625–1644.

Li, H.; Zhao, J.; Li, J.; Yu, Z.; and Lu, G. 2023a. Feature dynamic alignment and refinement for infrared–visible image fusion: Translation robust fusion. *Information Fusion*, 95: 26–41.

Li, S.; Tu, Y.; Xiang, Q.; and Li, Z. 2024b. MAGIC: Rethinking Dynamic Convolution Design for Medical Image Segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9106–9115.

Li, W.; Zhang, Y.; Wang, G.; Huang, Y.; and Li, R. 2023b. DFENet: A dual-branch feature enhanced network integrating transformers and convolutional feature learning for multimodal medical image fusion. *Biomedical Signal Processing and Control*, 80: 104402.

Loshchilov, I.; and Hutter, F. 2017. Sgdr: Stochastic gradient descent with warm restarts. In *the proceedings of the International Conference on Learning Representations (ICLR)*.

Mu, P.; Wu, G.; Liu, J.; Zhang, Y.; Fan, X.; and Liu, R. 2023. Learning to Search a Lightweight Generalized Network for Medical Image Fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7): 5921–5934.

Piella, G.; and Heijmans, H. 2003. A new quality metric for image fusion. In *Proceedings 2003 International Conference on Image Processing (ICIP)*, volume 3, III–173. IEEE.

Song, Y.; Dai, Y.; Liu, W.; Liu, Y.; Liu, X.; Yu, Q.; Liu, X.; Que, N.; and Li, M. 2024. DesTrans: A medical image fusion method based on transformer and improved DenseNet. *Computers in Biology and Medicine*, 174: 108463.

Tang, L.; Deng, Y.; Ma, Y.; Huang, J.; and Ma, J. 2022a. SuperFusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12): 2121–2137.

Tang, W.; He, F.; Liu, Y.; and Duan, Y. 2022b. MATR: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Transactions on Image Processing*, 31: 5134–5149.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 5998–6008.

Wang, D.; Liu, J.; Fan, X.; and Liu, R. 2022. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, 3508–3515.

Wang, D.; Liu, J.; Ma, L.; Liu, R.; and Fan, X. 2024. Improving Misaligned Multi-modality Image Fusion with One-stage Progressive Dense Registration. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Wen, J.; Qin, F.; Du, J.; Fang, M.; Wei, X.; Chen, C. P.; and Li, P. 2023. MSGFusion: Medical semantic guided two-branch network for multimodal brain image fusion. *IEEE Transactions on Multimedia*, 26: 944–957.

Xie, X.; Zhang, X.; Ye, S.; Xiong, D.; Ouyang, L.; Yang, B.; Zhou, H.; and Wan, Y. 2023. MRSCFusion: Joint residual Swin transformer and multiscale CNN for unsupervised multimodal medical image fusion. *IEEE Transactions on Instrumentation and Measurement*, 72: 5026917.

Xu, H.; Ma, J.; Jiang, J.; Guo, X.; and Ling, H. 2020. U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 502–518.

Xu, H.; Ma, J.; Yuan, J.; Le, Z.; and Liu, W. 2022. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19679–19688.

Xu, H.; Yuan, J.; and Ma, J. 2023. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 12148–12166.

Xydeas, C. S.; Petrovic, V.; et al. 2000. Objective image fusion performance measure. *Electronics letters*, 36(4): 308–309.

Ye, S.; Wang, T.; Ding, M.; and Zhang, X. 2023. F-DARTS: Foveated differentiable architecture search based multimodal medical image fusion. *IEEE Transactions on Medical Imaging*, 42(11): 3348–3361.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient Transformer for High-Resolution Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5728–5739.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; and Van Gool, L. 2023a. 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 (CVPR)*, 5906–5916.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Zhang, K.; Xu, S.; Chen, D.; Timofte, R.; and Van Gool, L. 2024. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 25912–25921.

Zhao, Z.; Bai, H.; Zhu, Y.; Zhang, J.; Xu, S.; Zhang, Y.; Zhang, K.; Meng, D.; Timofte, R.; and Van Gool, L. 2023b. DDFM: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8082–8093.

Zheng, B.; Li, H.; Chen, Q.; Wang, T.; Zhou, X.; Hu, Z.; and Yan, C. 2024. QuadzBayer Joint Demosaicing and Denoising Based on Dual Encoder Network with Joint Residual Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7552–7561.

Zhong, Y.; Zhang, S.; Liu, Z.; Zhang, X.; Mo, Z.; Zhang, Y.; Hu, H.; Chen, W.; and Qi, L. 2023. Unsupervised Fusion of Misaligned PAT and MRI Images via Mutually Reinforcing Cross-Modality Image Generation and Registration. *IEEE Transactions on Medical Imaging*, 1702–1714.

Zuo, Q.; Zhang, J.; and Yang, Y. 2021. DMC-fusion: Deep multi-cascade fusion with classifier-based feature synthesis for medical multi-modal images. *IEEE Journal of Biomedical and Health Informatics*, 25(9): 3438–3449.