An Adaptive Graph and Hybrid Attention Framework for Unaligned Medical Image Fusion

Wenxuan Feng[1 [0009−0003−7588−3776]], Yufeng Zhou[1 [0009−0008−2864−0604]⋆], and Wenming Cao[1 [0000−0002−8174−6167]]

Shenzhen University, Shenzhen, 518060, China
2410043102@mails.szu.edu.cn, 2250432015@email.szu.edu.cn, wmcao@szu.edu.cn

Abstract. Unaligned multimodal medical image fusion aims to jointly achieve feature alignment and information integration. However, severe modality discrepancies, noise interference, and blurred boundaries often undermine registration stability and fusion quality. In this paper, we propose a Hybrid Attention Graph Fusion Network, termed HAGF-Net, for unaligned multimodal medical image fusion. We find that effective fusion of unaligned multimodal medical images requires not only reliable structural representation, but also robust cross-modal correspondence modeling under uncertain imaging conditions. Therefore, we introduce a Dual Attention Transformer Block (DATB) to enhance structural feature learning by jointly capturing global contextual dependencies and local spatial details. Moreover, we develop a Fuzzy-Aware Graph Registration Block (FGRB), in which Adaptive Feature Graph Modeling (AFGM) dynamically establishes representative structural relationships across modalities to improve correspondence estimation. To further address unreliable responses caused by fuzzy regions, we incorporate Uncertainty-Aware Fuzzy Attention (UFA) together with spatial attention, which suppresses ambiguous features and highlights informative anatomical regions for more accurate deformation estimation. Extensive experiments demonstrate that HAGF-Net achieves superior registration accuracy and fusion quality under complex modality discrepancies and nonlinear deformations. For instance, on the MRI-CT dataset, our method achieves improvements of 0.275 in $Q_{AB/F}$ and 0.395 in Q_S over existing methods.

Keywords: Multimodal Medical Image Fusion · Graph Modeling · Attention Mechanism · Deformable Registration.

1 Introduction

Multimodal medical image fusion (MMIF) aims to integrate complementary information from different imaging modalities, such as CT, MRI, and PET, into a unified representation with enhanced structural details and improved lesion characterization. In recent years, deep learning-based MMIF methods have achieved

⋆ Corresponding author: 2250432015@email.szu.edu.cn

promising performance. CNN-based approaches [1] are effective in extracting local textures and structural details, Transformer-based models [2] improve long-range dependency modeling, and hybrid frameworks [3] attempt to exploit the complementary advantages of convolution and attention mechanisms. However, most of these methods are developed under the assumption that the input images are strictly aligned at the pixel level. This assumption is often violated in real clinical scenarios due to patient motion, modality-specific imaging mechanisms, and local non-rigid anatomical deformations, which significantly limits the practical applicability of existing fusion models.

To address image misalignment, conventional approaches typically adopt a two-stage pipeline that performs image registration prior to fusion [4, 5]. Although this strategy alleviates misalignment to some extent, cross-modal medical image registration remains highly challenging due to substantial modality discrepancies and complex anatomical variations. Recent studies have therefore explored unified frameworks that jointly model registration and fusion within a single architecture [6, 7], aiming to reduce error accumulation and enable end-to-end optimization. However, most existing joint approaches are designed for natural images or simplified multimodal scenarios, and still struggle under severe modality gaps and complex non-rigid deformations in medical images.

The core difficulty is that joint registration and fusion require both cross-modal consistency and modality complementarity, while accurate alignment further depends on jointly modeling global anatomical context and fine-grained local details. Existing architectures remain inadequate for this purpose: CNN-based methods are effective for local feature extraction but limited in long-range dependency modeling, whereas Transformer-based methods can capture global context but often suffer from high computational cost and insufficient spatial correspondence for fine anatomical structures. Restormer [8] provides an efficient restoration-oriented Transformer paradigm for high-resolution image representation; however, it is not explicitly designed to address cross-modal appearance discrepancy or deformation-aware spatial correspondence, which are essential for unaligned MMIF. In addition to feature representation, structural relation modeling is also crucial for reliable deformation estimation, as KNN-based dynamic graph construction is adaptive but computationally expensive [9], while static graph construction is more efficient but lacks input-aware flexibility [10]. These limitations are further exacerbated by fuzzy medical image characteristics, such as intensity inhomogeneity, noise, partial volume effects, and blurred boundaries, which introduce uncertainty and reduce the reliability of deformation estimation [11].

To address these challenges, we propose a Hybrid Attention Graph Fusion Network (HAGF-Net), a unified framework for joint registration and fusion of unaligned multimodal medical images. It consists of a cross-modal feature interaction module, a bidirectional stepwise registration module, and a multimodal feature fusion module. In the cross-modal feature interaction module, we incorporate a Restormer-inspired Dual Attention Transformer Block (DATB) to enhance global–local feature representation by integrating channel attention with

spatial attention. Based on the enhanced features, we develop a Fuzzy-Aware Graph Registration Block (FGRB) for robust cross-modal correspondence modeling. Within FGRB, Adaptive Feature Graph Modeling (AFGM) constructs efficient and input-adaptive structural relationships across modalities, overcoming the limitations of both static and KNN-based graph construction. Furthermore, Uncertainty-Aware Fuzzy Attention (UFA) and Spatial Attention (SPA) are incorporated to suppress ambiguous responses, emphasize reliable anatomical regions, and progressively refine deformation estimation. The aligned features are then fed into the fusion module, enabling unified optimization of registration and fusion.

Extensive experiments on unaligned multimodal medical image datasets demonstrate that the proposed method consistently outperforms state-of-the-art fusion and joint registration–fusion approaches in terms of both alignment accuracy and fusion quality. These results validate the effectiveness of the proposed hybrid attention and graph-based design in handling complex modality discrepancies and anatomical variations. In summary, the main contributions of this work are three-fold:

- We propose a Hybrid Attention Graph Fusion Network (HAGF-Net) for unaligned multimodal medical image fusion, in which registration and fusion are jointly optimized within a unified single-stage architecture.
- We introduce a Dual Attention Transformer Block (DATB) as a feature enhancement unit in the cross-modal feature interaction module, and further develop a Fuzzy-Aware Graph Registration Block (FGRB) to jointly enhance global–local feature representation and reliable cross-modal structural correspondence modeling.
- We further develop an adaptive uncertainty-aware deformation refinement strategy within FGRB, where Adaptive Feature Graph Modeling (AFGM) enables input-adaptive relation construction, while Uncertainty-Aware Fuzzy Attention (UFA) and Spatial Attention (SPA) improve the robustness and spatial precision of deformation estimation.

2 Method

As shown in Fig. 1, HAGF-Net consists of three components: a Cross-Modal Feature Interaction Module, a Bidirectional Stepwise Registration Module, and a Multi-Modal Feature Fusion Module. The first module enhances shared feature representations for unaligned multimodal images via global–local modeling and cross-modal interaction. Based on the enhanced features, the registration module progressively estimates deformation fields for feature alignment. To improve robustness under large modality discrepancies, the Fuzzy-Aware Graph Registration Block (FGRB) integrates adaptive graph modeling and uncertainty-aware attention to refine deformation estimation. Finally, the fusion module warps the features with the estimated deformation and generates the fused image.

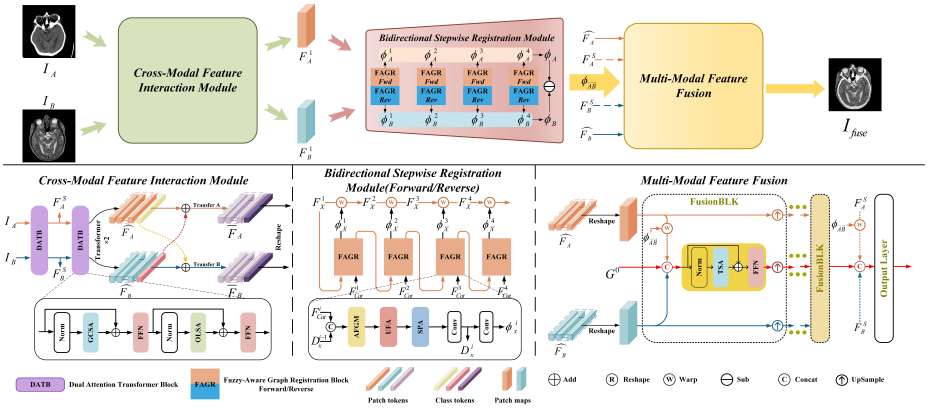

Fig. 1. Overview of HAGF-Net. For an unaligned multimodal image pair $\{I_A, I_B\}$, the Cross-Modal Feature Interaction Module first extracts shared features $\{F_A^S, F_B^S\}$ and interaction-enhanced features $\{\bar{F}_A, \bar{F}_B\}$. The Bidirectional Stepwise Registration Module then progressively estimates deformation fields for feature alignment. Finally, the aligned features, together with the shared representations and the predicted deformation field $\phi_{\overrightarrow{AB}}$, are fed into the Multi-Modal Feature Fusion Module to generate the fused image I_{fuse}.

2.1 Dual Attention Transformer Block

Accurate fusion of unaligned multimodal medical images requires both global anatomical consistency and fine-grained local structural correspondence. However, convolution-based operators are usually limited in long-range dependency modeling, while conventional Overlapping Local Spatial Attention suffers from high computational cost when applied to high-resolution dense prediction tasks. To obtain efficient and discriminative structural representations, we incorporate a *Dual Attention Transformer Block* (DATB) into the cross-modal feature interaction module. Motivated by the efficient high-resolution representation paradigm of Restormer [8], DATB is introduced for cross-modal feature interaction. It combines channel-wise global dependency modeling with spatial attention-based local structural modeling, thereby providing more reliable anatomical representations for subsequent deformation estimation and fusion.

As illustrated in Fig. 1, DATB integrates *Global Channel-Structural Attention* (GCSA) and *Overlapping Local Spatial Attention* (OLSA) within a unified framework. GCSA is used to efficiently capture global dependencies along the channel dimension, while OLSA enhances local spatial interaction for fine structural correspondence. Given an input feature $F_{\text{in}} \in \mathbb{R}^{H\times W\times C}$, DATB can be formulated as

$$F_G = F_{\text{in}} + \text{FFN}(\text{GCSA}(\text{LN}(F_{\text{in}}))), \quad F_O = F_G + \text{FFN}(\text{OLSA}(\text{LN}(F_G))). \quad (1)$$

where LN(·) denotes layer normalization. In this formulation, F_G captures the intermediate representation after channel-wise attention and nonlinear transformation, while F_O denotes the refined feature produced after spatial attention.

The GCSA branch follows the preliminary multi-Dconv head transposed attention (MDTA) used in Restormer [8]. As shown in Fig. 2(a), the query Q, key K, and value V are first generated by a 1×1 point-wise convolution followed by a 3×3 depth-wise convolution. After reshaping Q and K, GCSA computes a channel-wise attention matrix of size $C \times C$, rather than a spatial attention matrix of size $HW \times HW$. Therefore, it can aggregate global contextual information with reduced computational complexity, which is suitable for dense medical image representation. Although GCSA provides efficient global context modeling, channel-wise attention alone is insufficient to describe fine local anatomical correspondence under non-rigid misalignment. Therefore, we further introduce OLSA to enhance spatial interaction. In OLSA, in Fig. 2(c), the query tokens are obtained from non-overlapping windows of size $M \times M$, whereas the key and value tokens are extracted from enlarged overlapping windows of size $M_O \times M_O$. For each query window, attention is computed between the query tokens and their corresponding overlapping contextual tokens. This asymmetric window construction enables the model to exploit neighboring spatial context while maintaining local attention efficiency. The FFN adopts a gated depth-wise feed-forward design, following the spirit of GDFN in Restormer. As shown in Fig. 2(b), the input feature is projected into two parallel branches with point-wise convolutions and depth-wise convolutions, followed by element-wise multiplication. A final 1×1 convolution projects the gated feature back to the original channel dimension. This gated mechanism controls the information flow by suppressing less informative responses and allowing useful structural features to be propagated to subsequent layers.

Base on DATB, the Cross-Modal Feature Interaction Module aims to enhance shared representations for the unaligned multimodal image pair $\{I_A, I_B\}$. The shared encoder first extracts shallow shared features $\{F_A^S, F_B^S\}$ and deep features, while DATB strengthens their structural representation by jointly modeling channel-wise global dependencies and spatial local details. To ensure that the encoded global representations preserve modality-discriminative information, the modality-specific tokens are fed into an MLP and supervised by

$$\mathcal{L}_{ce1} = CE(p_A, [0, 1]) + CE(p_B, [1, 0]), \tag{2}$$

where p_A and p_B denote the predicted modality labels, CE denotes soft-label cross-entropy. Based on these modality-aware representations, the transfer branches further exchange complementary information across modalities, yielding interaction-enhanced features $\{\hat{F}_A, \hat{F}_B\}$. The transferred features are then processed by two dedicated transformation branches to obtain $\{\bar{F}_A, \bar{F}_B\}$ for subsequent deformation estimation. To encourage these transformed features to become modality-invariant, we further impose

$$\mathcal{L}_{ce2} = CE(p_A^*, [0.5, 0.5]) + CE(p_B^*, [0.5, 0.5]), \tag{3}$$

where p_A^* and p_B^* denote the modality predictions of the transformed features. In this way, the proposed module reduces modality discrepancy while preserving complementary information for subsequent deformation estimation.

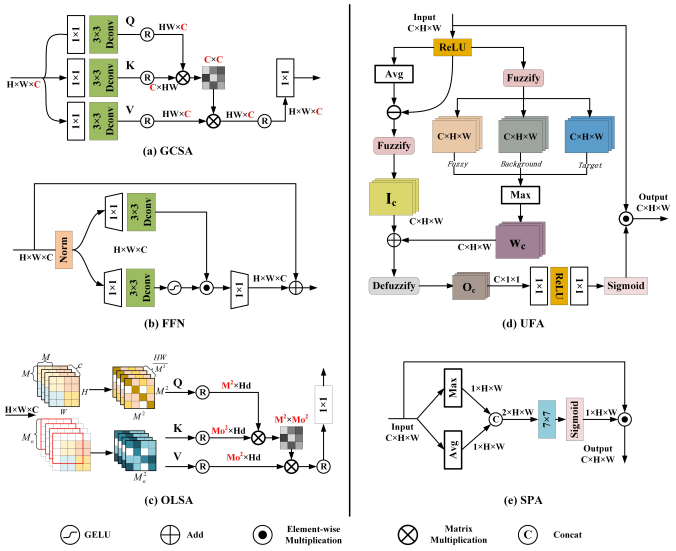

Fig. 2. Structures of the proposed modules: (a) Global Channel-Structural Attention (GCSA) , (b) Feed-Forward Network (FFN), (c) Overlapping Local Spatial Attention (OLSA) , (d) Uncertainty-Aware Fuzzy Attention (UFA), and (e) SPatial Attention (SPA).

2.2 Fuzzy-Aware Graph Registration Block (FGRB)

As shown in Fig. 1, the proposed Bidirectional Stepwise Registration Module (BSRM) progressively predicts deformation fields for the interaction-enhanced features $\bar{F}_A$ and $\bar{F}_B$ from two directions. Instead of estimating a large deformation field in a single step, BSRM decomposes registration into multiple bidirectional refinement stages, thereby gradually reducing cross-modal structural discrepancies and improving alignment stability. At each stage, a Fuzzy-Aware Graph Registration Block (FGRB) serves as the basic deformation estimation unit. Specifically, the i-th forward and reverse branches take $\{F_{\text{cat}}^i, D_A^{i-1}\}$ and $\{F_{\text{cat}}^i, D_B^{i-1}\}$ as inputs, respectively, where

$$F_{\text{cat}}^i = \text{concat}(F_A^i, F_B^i), \quad F_j^i = \uparrow_{\times 2} \left(\mathcal{W}(F_j^{i-1}, \phi_j^{i-1})\right), \quad j \in \{A, B\}. \tag{4}$$

Here, $\mathcal{W}(\cdot)$ denotes the warp operation and $\uparrow_{\times 2}$ denotes $2\times$ upsampling. Within each FGRB, Adaptive Feature Graph Modeling (AFGM), Uncertainty-Aware Fuzzy Attention (UFA), and Spatial Attention (SPA) are sequentially introduced

to model adaptive cross-modal structural relationships, suppress uncertain responses, and enhance informative spatial regions. Based on the refined features, FGRB predicts the updated deformation representation D_x^i and the intermediate deformation field ϕ_x^i. After K stages of bidirectional refinement, the final deformation fields are obtained by aggregating the intermediate predictions:

$$\phi_A = \sum_{i=1}^{K} \phi_A^i, \quad \phi_B = \sum_{i=1}^{K} \phi_B^i, \quad \phi_{\overrightarrow{AB}} = \phi_A - \phi_B. \tag{5}$$

Registration is supervised by smoothness and consistency losses:

$$\mathcal{L}_{smooth} = \sum_{i=1}^{K} \left(\|\nabla \phi_A^i\|_2 + \|\nabla \phi_B^i\|_2 \right), \tag{6}$$

$$\mathcal{L}_{consis} = \mathcal{L}_{ssim}\big(I'_A,\, W(I_A, \phi_{\overrightarrow{AB}})\big) + \left\| I'_A - W(I_A, \phi_{\overrightarrow{AB}}) \right\|_1, \tag{7}$$

where $\mathcal{L}_{ssim}$ denotes the structural similarity (SSIM) loss, and I'_A is the label image obtained by strictly aligning I_A to I_B at the pixel level.

Adaptive Feature Graph Modeling (AFGM) We adopt Adaptive Feature Graph Modeling (AFGM): as an efficient dynamic graph construction method to replace the computationally expensive KNN-based graph construction. AFGM is built on the axial graph paradigm of SVGA [12], but differs from static graph construction by generating a graph that varies with input images. Specifically, given an input feature map X, we first partition it into four quadrants and perform a diagonal flip to obtain a spatially rearranged feature representation while preserving structural correspondence. By computing the Euclidean distance between the original and flipped features at corresponding positions, a representative set of distance samples is obtained, from which the mean μ and standard deviation σ of node-wise distances are estimated. Accordingly, an adaptive threshold is defined as $\theta = \mu - \sigma$. Using this threshold, candidate neighbors are generated along both horizontal and vertical directions via rolling operations. For each candidate pair, the feature distance is computed as

$$D = \|X - X_{\text{rolled}}\|. \tag{8}$$

A binary mask is then defined by comparing D with the adaptive threshold θ, so that only valid connections are preserved for graph construction. Based on the selected neighbors, node relationships are modeled by the relative feature difference $X_{\text{rolled}} - X$, and the resulting graph responses are aggregated to obtain the graph-enhanced feature X_{final}. Finally, the original feature X and X_{final} are concatenated along the channel dimension and fused by a 1×1 convolution to produce the final output.

To embed this mechanism into the registration network, we further develop a lightweight AFGM Block, which consists of an input projection, a dynamic axial graph convolution, an output projection, and a residual connection. Given

Algorithm 1 Adaptive Feature Graph Modeling

Given: K, the axial connection interval; X, the input feature map; $L_d \in \{H, W\}$, the axis length corresponding to direction d

$X_{\mathrm{q}} \leftarrow \mathrm{QuadrantFlip}(X)$
$dist \leftarrow \mathrm{norm}(X, X_{\mathrm{q}})$
$\mu \leftarrow \mathrm{mean}(dist)$
$\sigma \leftarrow \mathrm{std}(dist)$
$thr \leftarrow \mu - \sigma$
$X_{\mathrm{final}} \leftarrow 0$
for $d \in \{H, W\}$ **do**
  for $k = K, 2K, \ldots, < L_d$ **do**
    $X_{\mathrm{rolled}} \leftarrow \mathrm{Roll}(X, k, d)$
    $dist \leftarrow \mathrm{norm}(X, X_{\mathrm{rolled}})$
    $mask \leftarrow \mathbf{1}(dist < thr)$
    $X_d \leftarrow mask \cdot (X_{\mathrm{rolled}} - X)$
    $X_{\mathrm{final}} \leftarrow \max(X_d, X_{\mathrm{final}})$
  end for
end for
return $\mathrm{Conv}(\mathrm{Concat}(X, X_{\mathrm{final}}))$

an input feature $X \in \mathbb{R}^{C\times H\times W}$, we first map it into a graph-friendly embedding space through a learnable 1×1 convolutional projection with batch normalization, denoted by W_{in}. Then, a dynamic graph operator, termed AdaGraph, is applied according to Algorithm 1. The resulting graph features are subsequently projected back to the original channel space by another 1×1 convolutional projection with batch normalization, denoted by W_{out}, and fused with the input via a residual connection:

$$Y = W_{out}(\mathrm{AdaGraph}(W_{in}(X))) + X. \tag{9}$$

This design enables input-adaptive graph construction and improves long-range structural modeling compared with fixed-neighbor methods.

Uncertainty-Aware Fuzzy Attention and Spatial Refinement Although AFGM is capable of dynamically establishing more representative cross-modal structural relationships, the resulting relation-enhanced features may still be affected by modality discrepancies, blurred boundaries, and noise. To further suppress unreliable responses and improve the robustness of deformation estimation, we introduce an Uncertainty-Aware Fuzzy Attention (UFA) module, as illustrated in Fig. 2(d), to perform uncertainty-aware recalibration on the relation-enhanced features.

Given an input feature map $X \in \mathbb{R}^{B\times C\times H\times W}$, we first adopt a bounded ReLU activation to stabilize the training process: $X' = \mathrm{ReLU6}(X)$, which constrains the feature values within a stable range. Subsequently, three learnable sigmoid-based fuzzy membership functions are employed to model the degrees of membership for target, ambiguous, and background regions, respectively. Specifically, let X'_c denote the feature of the c-th channel, the fuzzification process is defined as:

$$\mu_l(X'_c) = \sigma\big(-g_l(X'_c + s_l)\big), \quad l \in \{1, 2, 3\}, \tag{10}$$

where s_l and g_l are learnable parameters, $\sigma(\cdot)$ denotes the sigmoid function, and μ_l represents the membership degree of the l-th fuzzy set. According to fuzzy entropy theory, uncertainty reaches its maximum when the membership value is close to 0.5, whereas values closer to 0 or 1 indicate higher certainty. Based

on this observation, we define the uncertainty weight as: $w_l = 2\left|\mu_l(X'_c) - 0.5\right|$, and take the maximum response across the three fuzzy sets to obtain the final uncertainty weight at each spatial location:

$$W_c = \max_{l\in\{1,2,3\}} w_l. \tag{11}$$

This design suppresses ambiguous regions while assigning larger weights to more reliable structural regions. To further characterize the relationship between feature responses and the channel-wise statistical center, we compute the spatial mean m_c for each channel and construct an additional fuzzification branch:

$$I_c = \sigma\big(k(X'_c - m_c)\big), \tag{12}$$

where k is a learnable parameter. This branch captures the relative deviation of features with respect to the channel-wise mean. Then, the fuzzy response I_c and the uncertainty weight W_c are fused via a fuzzy algebraic sum:

$$P_c = I_c \oplus W_c = I_c + W_c - I_c W_c. \tag{13}$$

Based on the fused representation, a weighted defuzzification strategy is adopted to aggregate spatial information and obtain a channel-wise global descriptor:

$$O_c = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W} P_c(i,j)X'_c(i,j)}{\sum_{i=1}^{H}\sum_{j=1}^{W} P_c(i,j)}. \tag{14}$$

Here, $O_c \in \mathbb{R}^{B\times C\times 1\times 1}$ can be regarded as the channel representation after uncertainty-aware weighting. Finally, two 1×1 convolution layers followed by a non-linear activation are applied to generate channel attention weights, which are used to recalibrate the input features:

$$A_c = \sigma\big(\mathrm{Conv}_2(\mathrm{ReLU}(\mathrm{Conv}_1(O_c)))\big), \quad Y = X \odot A_c, \tag{15}$$

where $\odot$ denotes element-wise multiplication and Y is the output feature of UFA.

The above process enables UFA to adaptively enhance reliable responses while suppressing ambiguous or noisy ones from an uncertainty-aware perspective. However, channel-wise recalibration alone is insufficient to fully capture the spatial distribution of key anatomical structures. To further emphasize local regions relevant to deformation estimation, we introduce a Spatial Attention (SPA) module, as shown in Fig. 2(e), after UFA for spatially selective enhancement. Specifically, SPA extracts complementary spatial cues via channel-wise max pooling and average pooling, and generates a spatial attention map:

$$\mathrm{SPA}(x) = \sigma\big(\mathrm{Conv}([C_{\max}(x),\, C_{\mathrm{avg}}(x)])\big) \odot x, \tag{16}$$

where $C_{\max}(\cdot)$ and $C_{\mathrm{avg}}(\cdot)$ denote max pooling and average pooling, respectively. Through this design, SPA further enhances informative structural regions, providing more precise spatial representations for subsequent deformation estimation.

2.3 Multimodal Feature Fusion

As illustrated in Fig. 1, given the estimated deformation field $\phi_{\overrightarrow{AB}}$, the deep features $\hat{F}_A, \hat{F}_B$ are first rearranged into spatial feature maps and then fed into the multimodal feature fusion module together with the shallow features F_s^A, F_s^B. The fusion module performs progressive feature integration in a coarse-to-fine manner. At each fusion stage, the feature of modality A is spatially aligned under the guidance of the deformation field and then aggregated with the corresponding feature of modality B and the fusion output from the previous stage. The output of the i-th FusionBLK is denoted as G^i, which represents the intermediate fusion representation at the current scale and is passed to the next stage for further refinement. G^0 is initialized as a zero tensor the output of the last FusionBLK is denoted as G^4. Subsequently, G^4 is concatenated with F_s^B and $\widetilde{F}_s^A = \mathcal{W}(F_s^A, \phi_{\overrightarrow{AB}})$. The resulting representation is then fed into the reconstruction layer to generate the fused image I_{fuse}.

To ensure the quality of the fused image, we adopted three fusion-related losses, including a structural loss, an intensity loss, and a gradient loss. The structural loss is defined as

$$\mathcal{L}_{struct} = \mathcal{L}_{ssim}(I_{fuse}, \hat{I}_A) + \mu\mathcal{L}_{ssim}(I_{fuse}, I_B), \tag{17}$$

where $\hat{I}_A = \mathcal{W}(I_A, \phi_{\overrightarrow{AB}})$ and μ is a balancing coefficient. Intensity loss and gradient loss are jointly defined as

$$\mathcal{L}_{inten} = \|I_{fuse} - \max(\hat{I}_A, I_B)\|_1,\ \mathcal{L}_{grad} = \|\nabla I_{fuse} - \max(\nabla\hat{I}_A, \nabla I_B)\|_1. \tag{18}$$

which enforce intensity consistency and preserve structural details, respectively. The overall objective function is formulated as

$$\mathcal{L}_{total} = \mathcal{L}_{ce1} + \mathcal{L}_{ce2} + \mathcal{L}_{smooth} + \mathcal{L}_{consis} + \mathcal{L}_{struct} + \mathcal{L}_{grad} + \lambda\mathcal{L}_{inten}, \tag{19}$$

where λ are the balance coefficients.

3 Experiments and Results

3.1 Experimental Setup

Dataset and Implementation Details: We conduct experiments on the MRI-CT and MRI-PET datasets from Harvard[1], containing 144 and 194 strictly aligned image pairs, respectively, at a resolution of 256×256. Unaligned samples are generated by taking MRI as the reference and applying mixed rigid and non-rigid deformations to the paired non-MRI images. The same strategy is used to construct the test sets from 20 MRI-CT pairs and 55 MRI-PET pairs. During training, random deformations, rotations, and flips are further used for data augmentation.

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

The model is trained end-to-end for 3000 epochs with a batch size of 32 using Adam [13]. The initial learning rate is set to 5×10^{-5} and decayed to 5×10^{-7} via cosine annealing [14]. We fix λ to 0.5 and update μ after each epoch according to

$$\mu = \frac{\sum_{n=1}^{N} \mathcal{L}_{\text{ssim}}^{(n)}(I_{\text{fuse}}, I_B)}{\sum_{n=1}^{N} \mathcal{L}_{\text{ssim}}^{(n)}(I_{\text{fuse}}, \tilde{I}_A)}, \tag{20}$$

where N is the number of training samples in each epoch. The implementation is based on PyTorch and runs on a single NVIDIA GeForce RTX 3090 GPU.

Evaluation Metrics: To quantitatively assess fusion performance, we utilize five representative image quality metrics, including Gradient-based Fusion Performance ($Q_{AB/F}$) [15], Chen–Varshney Metric (Q_{CV}) [16], Structure-based Metric (Q_S) [17], Structural Similarity Index Measure (Q_{SSIM}) [18], and Visual Information Fidelity (Q_{VIF}) [19]. Among these metrics, a lower value of Q_{CV} indicates better fusion quality, whereas higher values of the remaining metrics correspond to superior performance.

3.2 Comparison With the State-of-the-art Methods

Existing methods can be broadly divided into two categories: Registration and Fusion and Joint Registration and Fusion. Since our method belongs to the latter, we mainly compare with representative joint methods, including SuperFusion[6], MURF[7],UMF-CMGR[4], IMF[5], and BASFusion[20].

Fig. 3 shows qualitative comparisons of fusion results on MRI-CT and MRI-PET datasets. Existing methods often exhibit structural misalignment, blurred boundaries, or insufficient preservation of modality-specific informative regions, particularly in the enlarged local areas. In contrast, the proposed method produces fused images with clearer anatomical contours, more accurate structural alignment, and better retention of salient functional and structural details. This indicates that the proposed framework is more capable of jointly preserving cross-modal complementary information while maintaining structural consistency.

Table 1 reports the quantitative comparison results. On the MRI-CT dataset, our method achieves the best performance on all five metrics, including $Q_{AB/F}$ (0.6105), Q_{CV} (0.049), Q_S (0.936), Q_{SSIM} (1.567), and Q_{VIF} (0.709). On the MRI-PET dataset, our method also attains the best results on $Q_{AB/F}$ (0.7231), Q_S (0.967), and Q_{VIF} (0.823), while maintaining competitive performance on the remaining metrics. Compared with representative joint registration–fusion methods such as BASFusion and IMF, the proposed method consistently yields superior or more balanced performance, demonstrating stronger capability in structural preservation, information transfer, and visual fidelity. These results confirm the robustness and effectiveness of HAGF-Net for unaligned multimodal medical image fusion.

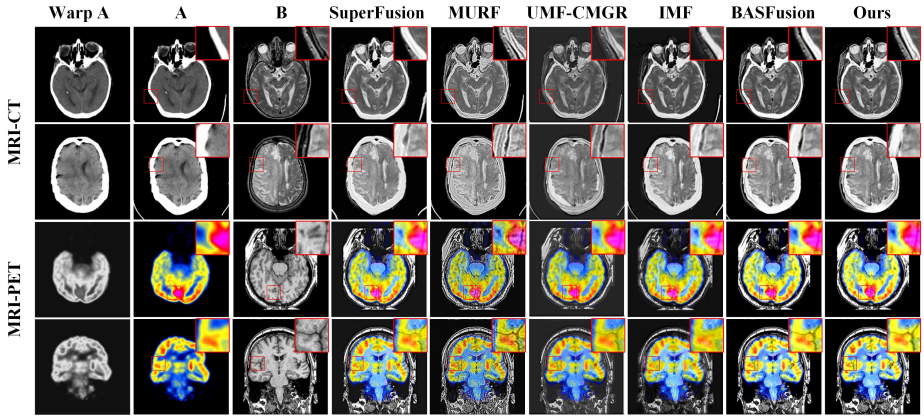

Fig. 3. Qualitative comparison of different methods on MRI-CT and MRI-PET examples. The first three columns show the deformed non-MRI image, the corresponding original CT/PET image, and the MRI image, respectively, while the remaining columns present the fusion results of competing methods and our method.

Table 1. Quantitative comparison on MRI-CT and MRI-PET datasets. The best results are shown in bold.

Method	MRI-CT					MRI-PET				
	$Q_{AB/F}\uparrow$	$Q_{CV}\downarrow$	$Q_S\uparrow$	$Q_{SSIM}\uparrow$	$Q_{VIF}\uparrow$	$Q_{AB/F}\uparrow$	$Q_{CV}\downarrow$	$Q_S\uparrow$	$Q_{SSIM}\uparrow$	$Q_{VIF}\uparrow$
SuperFusion	0.3355	0.13	0.541	0.982	0.472	0.6716	0.005	0.349	0.663	0.632
MURF	0.439	0.112	0.844	1.368	0.483	0.5707	0.022	0.382	0.645	0.56
UMF-CMGR	0.4085	0.110	0.139	0.452	0.304	0.4828	0.028	0.295	0.544	0.342
IMF	0.4719	0.056	0.891	1.472	0.613	0.6153	0.005	0.694	1.148	0.681
BASFusion	0.5649	0.051	0.932	1.557	0.698	0.7169	0.003	0.947	1.409	0.800
Ours	**0.6105**	**0.049**	**0.936**	**1.567**	**0.709**	**0.7231**	**0.003**	**0.967**	**1.435**	**0.823**

3.3 Ablation Study

To verify the effectiveness of each component, we conduct ablation experiments on the MRI-CT and MRI-PET datasets, as reported in Table 2. DATB improves Q_{VIF} on MRI-PET from 0.800 to 0.817, indicating its effectiveness in enhancing global–local feature representation. Since FAGR consists of AFGM, UFA, and SPA, we further evaluate its internal components. With AFGM alone, $Q_{AB/F}$ on MRI-CT reaches 0.5815, while the complete FAGR further improves it to 0.6102. This demonstrates that AFGM establishes cross-modal structural correspondences, whereas UFA and SPA further refine registration by suppressing unreliable responses and enhancing spatial constraints. In addition, Config B achieves better overall performance, suggesting that applying UFA before SPA is more effective. Overall, these results validate the effectiveness and complementarity of the proposed components.

To further support the quantitative ablation results, we visualize the registration performance on the MRI-CT and MRI-PET datasets in Fig. 4. The baseline shows obvious deviations between the label and registered intensity profiles, indi-

Table 2. Ablation study on MRI-CT and MRI-PET datasets. Config A: SPA→UFA; Config B: UFA→SPA. The best results are shown in bold.

Modality				MRI-CT					MRI-PET				
DATB	AFGM	UFA	SPA	$Q_{AB/F}\uparrow$	$Q_{CV}\downarrow$	$Q_S\uparrow$	$Q_{SSIM}\uparrow$	$Q_{VIF}\uparrow$	$Q_{AB/F}\uparrow$	$Q_{CV}\downarrow$	$Q_S\uparrow$	$Q_{SSIM}\uparrow$	$Q_{VIF}\uparrow$
–	–	–	–	0.5649	0.051	0.932	1.558	0.698	0.7169	0.003	0.947	1.409	0.800
✓	–	–	–	0.5713	0.049	0.932	1.561	0.706	0.7196	0.003	0.961	1.430	0.817
–	✓	–	–	0.5815	0.049	0.935	1.566	0.706	0.7149	0.003	0.967	1.428	0.820
✓	✓	–	–	0.5815	0.050	0.933	1.566	0.699	0.7200	0.003	0.966	1.430	**0.827**
✓	✓	Config A		0.6094	0.050	0.936	1.566	0.703	0.7204	0.003	0.962	1.426	0.816
✓	✓	Config B		**0.6102**	**0.049**	**0.938**	**1.567**	**0.710**	**0.7231**	**0.003**	**0.967**	**1.435**	0.823

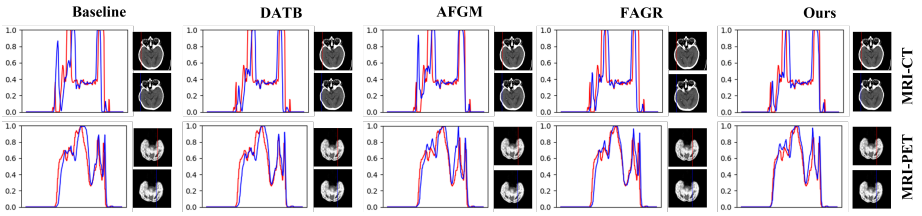

Fig. 4. The red curve represents the pixel values of the label, and the blue curve shows the pixel values after registration.

cating limited alignment capability. Introducing DATB reduces these deviations by enhancing both local details and global contextual information. Compared with AFGM, FAGR further improves the profile overlap, since UFA and SPA suppress unreliable correspondences and impose spatial consistency on the deformation field. The full model achieves the closest match between the red and blue curves, demonstrating the most accurate registration. These results confirm that the proposed modules are complementary and jointly contribute to robust unaligned medical image fusion.

To further evaluate the efficiency of AFGM, we compare it with CNN-, KNN-, and SVGA-based variants on the MRI-CT dataset in terms of NMI, parameters, and average CPU testing time. As shown in Fig. 5, AFGM achieves the highest NMI with the fewest parameters, indicating its superior ability to model cross-modal structural correspondences with a compact design. Although CNN has the shortest testing time, its lower NMI reflects insufficient representation capability. KNN obtains competitive accuracy but introduces considerable parameter and time costs, while SVGA yields lower accuracy than AFGM with only a slightly smaller testing time. Overall, AFGM provides the best trade-off among registration accuracy, model complexity, and computational efficiency.

4 Conclusion and Future Work

In this paper, we propose a unified framework for multimodal image fusion that enhances feature representation and cross-modal correspondence. By modeling global–local structures and adaptive cross-modal relationships, the method

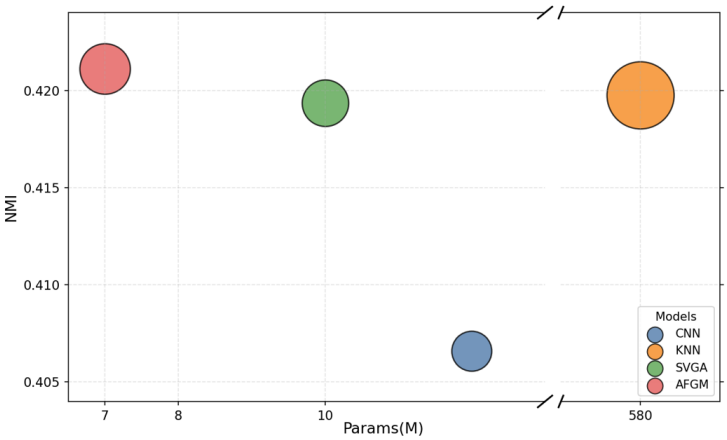

Fig. 5. Complexity–accuracy trade-off on the MRI-CT dataset. The x-axis denotes the number of parameters, and the y-axis represents the NMI value between the model-registered image and the strictly registered reference image, where a higher NMI indicates better registration accuracy. The bubble size indicates the average CPU testing time.

achieves more accurate alignment and better structural detail preservation. Uncertainty aware refinement further improves robustness under challenging conditions. Extensive experiments demonstrate that HAGF-Net consistently outperforms state-of-the-art approaches across different datasets and evaluation metrics, showing strong robustness and effectiveness. These results highlight the effectiveness of combining adaptive relation modeling with uncertainty-aware refinement for reliable multimodal alignment and fusion. Despite these promising results, HAGF-Net still has limitations in handling large or complex deformations. Future work will focus on improving its capability in such scenarios and enhancing efficiency for practical applications.

References

1. X. Gu, L. Wang, Z. Deng, Y. Cao, X. Huang, and Y.-m. Zhu, "Adaptive spatial and frequency experts fusion network for medical image fusion," *Biomedical Signal Processing and Control*, vol. 96, p. 106478, 2024.
2. Z. Zhao, H. Bai, Y. Zhu, J. Zhang, S. Xu, Y. Zhang, K. Zhang, D. Meng, R. Timofte, and L. Van Gool, "Ddfm: Denoising diffusion model for multi-modality image fusion," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 8082–8093.
3. W. Li, Y. Zhang, G. Wang, Y. Huang, and R. Li, "Dfenet: A dual-branch feature enhanced network integrating transformers and convolutional feature learning

for multimodal medical image fusion," *Biomedical Signal Processing and Control*, vol. 80, p. 104402, 2023.
4. D. Wang, J. Liu, X. Fan, and R. Liu, "Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration," *arXiv preprint arXiv:2205.11876*, 2022.
5. D. Wang, J. Liu, L. Ma, R. Liu, and X. Fan, "Improving misaligned multi-modality image fusion with one-stage progressive dense registration," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 11, pp. 10 944–10 958, 2024.
6. L. Tang, Y. Deng, Y. Ma, J. Huang, and J. Ma, "Superfusion: A versatile image registration and fusion network with semantic awareness," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 12, pp. 2121–2137, 2022.
7. H. Xu, J. Yuan, and J. Ma, "Murf: Mutually reinforcing multi-modal image registration and fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 10, pp. 12 148–12 166, 2023.
8. 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/CVF conference on computer vision and pattern recognition*, 2022, pp. 5728–5739.
9. K. Han, Y. Wang, J. Guo, Y. Tang, and E. Wu, "Vision gnn: An image is worth graph of nodes," *Advances in neural information processing systems*, vol. 35, pp. 8291–8303, 2022.
10. M. Munir, W. Avery, M. M. Rahman, and R. Marculescu, "Greedyvig: Dynamic axial graph construction for efficient vision gnns," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6118–6127.
11. Y. Jiang, K. Zhao, K. Xia, J. Xue, L. Zhou, Y. Ding, and P. Qian, "A novel distributed multitask fuzzy clustering algorithm for automatic mr brain image segmentation," *Journal of medical systems*, vol. 43, no. 5, p. 118, 2019.
12. M. Munir, W. Avery, and R. Marculescu, "Mobilevig: Graph-based sparse attention for mobile vision applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2211–2219.
13. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
14. I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.
15. C. S. Xydeas and V. Petrovic, "Objective image fusion performance measure," *Electronics letters*, vol. 36, no. 4, pp. 308–309, 2000.
16. H. Chen and P. K. Varshney, "A human perception inspired quality metric for image fusion based on regional information," *Information fusion*, vol. 8, no. 2, pp. 193–207, 2007.
17. G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Proceedings 2003 international conference on image processing (Cat. No. 03CH37429)*, vol. 3. IEEE, 2003, pp. III–173.
18. 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.
19. 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.
20. H. Li, D. Su, Q. Cai, and Y. Zhang, "Bsafusion: A bidirectional stepwise feature alignment network for unaligned medical image fusion," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 39, no. 5, 2025, pp. 4725–4733.