GMR
The Original

Enhancing The Accuracy And Robustness Of Multimodal MRI Image Registration Using Hybrid Deep Learning Approaches And Multi-Scale Feature Fusion

Priyanka Chouksey[1*], Manish Kashyap[2], A. Subbarao[3]

[1*] Department of Electronics and communication engineering, Maulana Azad National Institute of Technology, Bhopal, MP, India E-mail(s): chouksey_priyanka@yahoo.in
[2] Department of Electronics and communication engineering, Maulana Azad National Institute of Technology, Bhopal, MP, India Email: manish-kashyap@manit.ac.in
[3] Department of Electronics and communication engineering, Maulana Azad National Institute of Technology, Bhopal, MP, India Email: asrao@manit.ac.in;

Abstract

Purpose: The purpose of the study is to improve the accuracy and the robustness of the multimodal MRI image registration, where the significant variations in the intensity, the resolution, as well as the anatomical structures limit performance of the conventional and the deep learning-based methods. The research addresses the question of whether the hybrid deep learning framework, which integrates the local and the global feature modeling, can achieve the reliable alignment across the different MRI modalities while maintaining the robustness to the noise and the cross-modality variations.
Methods: The hybrid CNN–Transformer based image registration framework is proposed, which integrates the multi-scale feature extraction, the self-attention and the cross-attention mechanisms, the domain adaptation strategies, as well as the spatial transformer networks. The model captures the fine-grained local features as well as the long-range global dependencies. The experiments are conducted on the Osteoarthritis Initiative MRI dataset. The performance is evaluated using the overlap-based, the intensity-based, as well as the deformation-based metrics, which include the Dice Coefficient, the Jaccard Index, the Hausdorff Distance, the Mutual Information, the PSNR, as well as the Bending Energy. The ablation studies, the robustness tests under the noise, as well as the cross-modality evaluations are also performed.
Results: The proposed method achieves the superior registration performance compared to the baseline models as well as the state-of-the-art models, where the Dice Coefficient reaches 0.87 and the Jaccard Index reaches 0.78, while the Hausdorff Distance is reduced to 9.5 mm. The results remain stable under the noisy conditions and across the different MRI sequences, which confirms the strong generalization capability.
Conclusion: The hybrid CNN–Transformer framework provides the accurate, the robust, as well as the generalizable multimodal MRI registration. The integration of the multi-scale features, the attention mechanisms, as well as the domain adaptation significantly enhances the alignment precision.

Keywords: Attention mechanisms; cross-modality evaluation; deep learning hybrid model; image registration; multi-scale feature fusion; MRI.

1 Introduction

In domains like medical imaging, where aligning images from various modalities like MRI, CT, and PET is important for precise diagnosis as well as treatment planning, the multimodal image registration is an essential process. Every imaging modality provides a different perspective on the internal structures of the body, frequently emphasizing particular anatomical or functional features [1]. For example, CT is the best method for seeing bone structures, even though MRI is great for soft tissue contrast. Accurately aligning this disparate data is the challenge of multimodal image registration since it needs to be successfully integrated. But it is precisely this intricacy that makes the search for such alignment are both difficult and beneficial.

The difficult task of multimodal image registration involves aligning images with different visual attributes, including intensity, contrast, resolution, and the type of information they capture. These differences occur because different physical principles underlie each imaging technique. For example, CT uses X-rays, while MRI uses radio waves and magnetic fields. This inherent variability, however, complicates the direct comparison or alignment of images from disparate modalities. The process is further complicated by non-linear deformations in patient anatomy, motion artifacts, and varying

noise levels across modalities. Although these challenges are significant, addressing them is crucial for effective image registration because accurate alignment is essential for subsequent analysis [2][3].
In medical imaging, achieving accurate multimodal registration can greatly improve clinical workflows. It enables doctors to overlay or fuse images from diverse modalities (this helps in detecting diseases, planning surgeries, and tracking treatment progress). Misalignment, on the other hand, can lead to incorrect diagnoses or ineffective treatments. Traditional registration methods based on intensity similarity measures (i.e., mutual information) or feature-based approaches (i.e., key-point matching) have been applied for decades. However, these techniques often struggle with large variations between modalities, non-linear distortions, and the need for significant manual intervention. Because of these challenges, researchers continually seek novel solutions to enhance the accuracy of image registration methods. Although progress has been made, the complexities inherent in the data remain a significant hurdle for practitioners. [4][5][6].

1.1 Motivation

Multimodal image registration is still a difficult task, even with deep learning advancements. Convolutional neural networks (CNNs), which are good at capturing local features, are often used in traditional methods.
Conversely, they grapple with overarching structures and long-range dependencies (a significant hurdle). This challenge proves particularly critical because there exists a necessity for global alignment to synchronize images from diverse modalities. Furthermore, these methodologies typically focus on single-scale data: a limitation that hampers their capacity to adjust to varying resolutions and complex details. Current deep learning techniques frequently surpass classical methods; but, they still grapple with issues related to generalization [7]. Models that are trained on specific modality pairs like MRI-CT tend to perform inadequately on alternative pairs like PET-MRI, because of their inflexible architecture. Additionally, many of these models overlook multi-scale information, which is indispensable for accurately capturing both fine and broad structural details, thereby resulting in misalignments within critical regions.
Furthermore, attention mechanisms, which enhance performance in other domains, remain underutilized in image registration. These mechanisms can help models focus on vital regions, improving accuracy. Lastly, traditional models struggle with domain adaptation, as variations between datasets and imaging devices can degrade performance. This highlights the need for a hybrid framework that integrates multiple deep-learning techniques to effectively address the complexities of multimodal image registration [28].

1.2 Problem-Definition

The task of multimodal image registration can be formally defined as follows: Given two images I_1 and I_2 , each from different modalities, the goal is to find a spatial transformation T that aligns I_1 to I_2 such that corresponding anatomical structures in both images are accurately aligned. It is represented by the equation [8]:

$$I_1' = T(I_1) \quad (1)$$

where I_1' is a transformed image, ideally aligned with I_2.
The transformation T can be rigid or non-rigid. The primary challenge is to determine the correct transformation that reduces the misalignment between the images and variations in intensity, resolution, and structure between the two modalities.

1.3 Contribution

This study offers a novel framework that makes several significant contributions to address the limitations of existing multimodal image registration techniques:

Hybrid Deep Learning Model. To combine the benefits of CNNs and Transformers, we suggest a hybrid deep learning model. Local spatial features are crucial for accurate registration at smaller scales, and CNNs excel at detecting them. Transformers, on the other hand, aid in capturing the image's long-term relationships and general context. By combining these two strategies, our model efficiently makes use of both local as well as global data, improving accuracy and dependability throughout the registration process, particularly in difficult anatomical regions [9].

Multi-Scale Feature Extraction and Fusion. We incorporate a multi-scale feature extraction mechanism to address the challenge of varying image resolutions. As a result, the model can identify both fine-grained details (such as tiny tumors or anatomical borders) and coarse-grained, more expansive structures (such as organs or large tissue regions) [10][11]. By extracting and combining features at different scales, our model ensures that crucial information is maintained during the registration process, leading to more accurate alignment across modalities.

Attention Mechanisms for Enhanced Focus. We have included attention mechanisms in this registration model to focus on the important areas in the images where the entire area is not equally important for accurate registration. Using this mechanism, it can highlight regions that is useful clinically, such as tumor locations or key anatomical features [12].

Domain Adaptation for Variability Handling. Variability among different imaging modalities and datasets presents a major challenge in multimodal image registration. To ensure that the model can adapt to various imaging situations, we implement domain adaptation techniques [13]. Our framework addresses the differences between modalities by using

adversarial training and domain-specific normalization layers, enhancing the model's ability to withstand variations in imaging conditions and data distributions.

Advanced Evaluation with Novel Metrics. We thoroughly assess our framework using a variety of metrics, including new metrics that offer deeper insights into structural alignment, such as Hausdorff Distance and Bending Energy, as well as conventional registration measures, such as the Dice Coefficient and Mutual Information.
Our thorough evaluation of the model's resilience to different kinds of noise and occlusions guarantees that our strategy works in both ideal and real-world clinical settings.
This introduction serves as a foundation for the research paper by identifying shortcomings of both conventional techniques and current deep learning methods. Our hybrid model presents a holistic solution to the challenges of multimodal image registration; however, it could be applied in areas beyond medical imaging, because it addresses underlying issues effectively. Although the model shows promise, its application is still limited by certain factors [14].

2 Related Works

Aligning images from various modalities is the goal of multimodal image registration, an essential component of medical imaging. This field has undergone significant change, moving from traditional techniques that relied on local feature extraction and statistical measures to more recent deep learning-based strategies.
Although hybrid approaches (that leverage deep learning, attention mechanisms, and multi-scale feature extraction) have been employed in recent developments, traditional methods often center on structural similarities. Such inherent challenges like anatomical differences between modalities, image noise, and intensity variations are effectively addressed by these modern techniques. This section reviews critical contributions to the field and identifies methodological flaws; however, it situates our work within this rapidly evolving framework.

2.1 Current Approaches

Recently, Wang et al. (2024) created the VDD-Reg model to align retinal images from optical coherence tomography angiography (OCTA) and erythrocyte-mediated angiography (EMA) [15]. However, it is essential to understand in detail the underlying algorithms to appreciate the complexities. Even if there is a great deal of room for improvement in accuracy, challenges remain. This intricate technical interplay underscores the significance of further research in this field. This model employs segmentation techniques to effectively manage large variations in vessel density; however, it also highlights the potential of combining traditional methods with deep learning for specific medical applications.
Mok et al. (2024) presented a structural representation learning approach that focuses on deformable multimodal medical image registration and may apply to a variety of imaging modalities [16]. Their method uses contrastive learning and Deep Neighborhood Self-similarity (DNS), which allows the model to provide useful image representations. This occurs without the necessity of pre-aligned training datasets or anatomical annotations; however, it raises questions about potential limitations. Although promising, the efficacy of such methods remains to be fully established. Correcting problems with anatomical alignment and local noise alterations is made much easier with this method.
Numerous existing methodologies face challenges when confronted with complex cases which exhibit significant nonlinear variations between modalities. To tackle these issues, Deng et al. (2024) have introduced RCV-Net: an unsupervised registration model that amalgamates convolutional block attention modules (CBAM) and a ResNet architecture [17].
Furthermore, Yang et al. (2024) developed adjacent self-similarity 3D convolution (ASTC) to address issues (such as radiation variations, illumination changes and image noise) during registration [18]. ASTC generates strong multidimensional features by applying local nearby self-similarity technique; however, these features are then enhanced further using 3D convolution. This increases the accuracy of point matching, but it also raises questions about the computational efficiency involved. Although effective, one must consider the potential drawbacks of such methods (especially in real-time applications) because they may introduce complexity that could hinder practical implementation. However, the complexity of the interactions involved in these processes may produce unexpected results.
Even with these advancements, there is still a notable need for enhanced methodologies. Difficulties in capturing fine details and managing significant deformations persist, emphasizing the need for multi-scale feature extraction and attention mechanisms to improve performance across various datasets.

2.2 Attention Mechanisms and Multi-Scale Feature Extraction

Attention mechanisms are vital in contemporary multimodal image registration; they enable models to concentrate on critical regions (particularly in medical imaging, where specific anatomical features hold great significance). Wu et al. (2023) developed an attention-based glioma grading network (AGGN) that employs multi-branch convolutions to extract both shallow and deep features, enhancing accuracy which occurs through a dual-domain weighting of channel and spatial information [19]. Lin et al. (2024) proposed a multi-branch and multi-scale neural network for medical image fusion, incorporating attention mechanisms to optimize feature extraction and semantic understanding [20]. however, this improvement facilitates complex registration tasks. Although multi-scale feature extraction addresses variations in image texture and scale, many current methods still struggle to generalize across modalities with substantial structural differences. Therefore, this reveals a pressing need for further advancements in multimodal registration techniques.

2.3 Domain Adaptation for Multimodal Registration

Domain adaptation is essential for addressing variability across imaging modalities (especially when images are sourced from diverse scanners or protocols). Recent innovations strive to enhance the generalization of registration models. Chen et al. (2024) introduced IFSrNet, a multi-scale feature-guided registration network for multispectral images that employs Cycle Generative Adversarial Network (CycleGAN) to create pseudo-infrared images from visible ones [21]. This model utilizes a multi-head attention module to address inconsistencies between reference and target images (thus improving accuracy).

In a related endeavor, Yang et al. (2024) developed a network for underwater image enhancement that integrates spatial attention and feature aggregation, enhancing adaptability across multiple domains [22]. These studies highlight the growing importance of domain adaptation in multimodal registration; however, challenges continue to exist in effectively managing significant domain shifts within complex datasets (because these shifts hinder the registration process), indicating a necessity for ongoing research in this field.

2.4 Positioning Our Approach

Our proposed methodology enhances existing approaches by merging transformers with CNNs (convolutional neural networks), which facilitates superior feature extraction and long-range dependency modeling. We effectively capture both global as well as local features and integrate multi-scale feature extraction with attention mechanisms, thereby overcoming the challenges posed by substantial deformations and complex anatomical structures.

However, we additionally employ domain adaptation strategies to ensure dependable performance across diverse imaging modalities (which is crucial for applications about medical imaging). Although rigorous testing—alongside the application of sophisticated evaluation metrics—distinguishes our work, they also provide a more comprehensive assessment of model performance than traditional metrics (such as mean squared error and mutual information). This research bears implications that extend beyond its theoretical contributions as a consequence.

While deep learning, attention mechanisms, and domain adaptation have led to significant advancements in the field of multimodal image registration, issues with feature generalization, domain variability, and handling large deformations still exist [23]. To fill these gaps, we have developed a hybrid deep learning approach that combines the advantages of CNNs, transformers, multi-scale feature extraction, and domain adaptation. This provides a new approach to the advanced multimodal image registration model, but it is important to take implementation complexity into account.

3 Methodology

We propose a multimodal image registration framework that integrates some of the state-of-the-art techniques for improved alignment of images from different modality. In order tackle issues led to by difference of modality, our methodology uses a combination of deep learning, multi-scale feature extraction, attention mechanisms, domain adaptation and spatial transformer networks. We'll break out each of these components with explanations and the corresponding mathematical formulations as well as the rationale for the design choices.

3.1 Overall Architecture

Our method consists of a few modules consists of a feature integration module, a multi-scale feature integration attention mechanism, a fusion layer, domain adaptation strategies, and a STN-based registration module.

Fig.1. shows how a trajectory of the model looks like. Two input images from different modalities are initially fed through the feature extraction module to produce discriminative representations. The multi-scale feature fusion block then takes these features as input to extract features on multiple scales. The attention mechanisms guarantee that the most relevant parts of the image will be attended to by the model. We apply domain-adaptation approaches to reduce such modality gap. Finally, the spatial transformer network calculates the transformation needed to align the images.

Let I_A and I_B represent input images from modality A and modality B, respectively. The goal is to estimate a transformation matrix $\boldsymbol{T}$ that aligns I_B to I_A, such that [24]:

$$I'_B = T(I_B; \mathbf{T}) \quad (2)$$

where T is the transformation function, and I'_B is the registered version of I_B.

Feature Extraction Module. In the endeavor to extract features from multimodal images, we utilize a synthesis of Convolutional Neural Networks (CNNs) and Transformers. CNNs are especially proficient at identifying local patterns such as edges and textures whereas Transformers excel in modeling long-range dependencies [1]. This effectiveness is largely due to their self-attention mechanism; however, the integration of these methodologies presents unique challenges. Although CNNs provide intricate spatial information, Transformers are instrumental in elucidating contextual relationships across the entirety of the image, because they can simultaneously concentrate on various segments of the input.

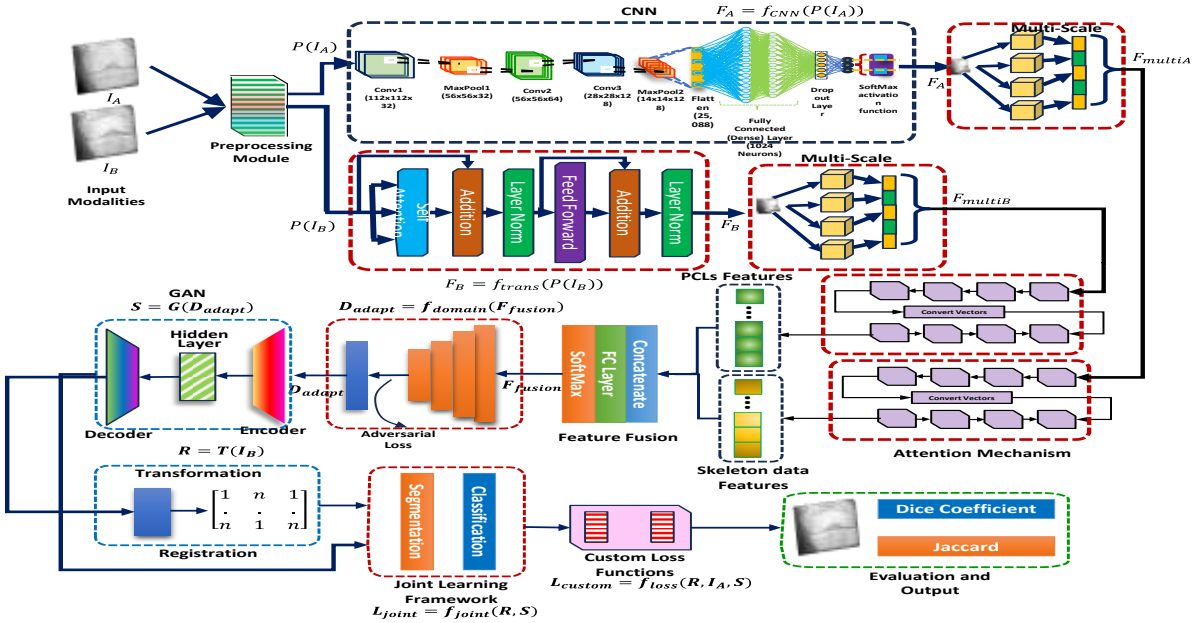

Fig. 1. Overall architecture of Hybrid Multimodal Registration.

CNN-based Feature Extraction. For each input image I, the CNN extracts hierarchical features at multiple levels [25]:

$$F_i = CNN_i(I) \quad (3)$$

where F_i denotes the feature maps extracted at the $i-th$ layer of the CNN. These features capture local information at various spatial resolutions, enabling fine-grained analysis of the input image.

Transformer-based Feature Extraction. The output of the CNN is then passed through a Transformer network to model global relationships [26]:

$$H_i = Transformer(F_i) \quad (4)$$

where H_i represents the globally contextualized feature map obtained through the self-attention mechanism of the Transformer. The self-attention matrix A in the Transformer is computed as [27]:

$$A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (5)$$

where Q and K are the query and key matrices, respectively, and d_k is the dimensionality of the keys. This operation allows the model to weigh the importance of different feature locations relative to each other, irrespective of spatial distance.

Multi-Scale Feature Fusion. For the integration of fine features with broader contextual information, multi-scale feature fusion is a crucial procedure (Smith, 2020). Our method allows us to represent the picture on many scales by using characteristics that are taken from different layers of the CNN-Transformer stack [28].

Let F_s denote features from a shallow layer and F_d from a deeper layer. We fuse these features using a weighted sum:

$$F_{fused} = w_s F_s + w_d F_d \quad (6)$$

where the contributions of each feature map are balanced by learnable weights $w_s$and w_dThis combination guarantees that both high-level semantic information and low-level details are included in the final representation. For additional refinement, a multi-scale convolution operation is used to the fused features [29].

$$F_{refined} = \sum_{i=1}^{n} Conv_{k_i}(F_{fused}) \quad (7)$$

where $Conv_{k_i}$ denotes convolution with kernel size k_i, and n is the number of scales. This allows the network to integrate features across different resolutions.

Attention Mechanisms. Attention mechanisms are crucial for focusing the model's capacity on the most relevant regions of the image. We use both self-attention and cross-attention in our architecture.

Self-Attention. Self-attention is applied to features from each modality independently, allowing the model to capture relationships within the same modality [30]:

$$A_s = softmax\left(\frac{Q_s K_s^T}{\sqrt{d_s}}\right) V_s \quad (8)$$

where Q_s, K_s,V_s are the query, key, and value matrices for the feature maps of modality A or B.

Cross-Attention. Cross-attention mechanisms are employed to align features across different modalities. For features F_A from modality A and F_B from modality B, cross-attention is formulated as [18] [31][32]:

$$A_{AB} = Softmax\left(\frac{Q_A K_B^T}{\sqrt{d_k}}\right) V_B \quad (9)$$

where Q_A is the query matrix from modality A and K_B, V_B are the key and value matrices from modality B. This enables the model to focus on corresponding regions across modalities.

Domain Adaptation. Domain adaptation is necessary to address the differences between the two modalities. We use adversarial training to ensure that the features extracted from the two modalities are aligned in a shared latent space.
Adversarial Training. We introduce a discriminator D that distinguishes between features from modality A and modality B. The goal of the feature extractor is to fool the discriminator, thus aligning the feature distributions of the two modalities. The adversarial loss is given by [33]:

$$\mathcal{L}_{adv} = \mathbb{E}[\log D(F_A)] + \mathbb{E}\big[\log\big(1 - D(F_B)\big)\big] \quad (10)$$

The feature extractor is trained to minimize $\mathcal{L}_{adv}$, while the discriminator is trained to maximize it.

Normalization. To further reduce modality discrepancies, we apply adaptive instance normalization (AdaIN) to the extracted features. AdaIN aligns the statistical distributions (mean and variance) of features across modalities [19] [34]:

$$AdaIN(F_A, F_B) = \sigma_B \left(\frac{F_A - \mu_A}{\sigma_A}\right) + \mu_B \quad (11)$$

Where μ_A, σ_A are the mean and standard deviation of F_A, and μ_B, σ_B are the corresponding statistics for F_B.

Registration Module. The core of the registration process is the Spatial Transformer Network (STN), which estimates the transformation **T** required to align the input images. The STN consists of a localization network that outputs the parameters of a transformation matrix. Given an input feature map FFF, the STN predicts a set of parameters θ that define the transformation [35]:

$$T = STN(F) = \theta \quad (12)$$

This transformation is then applied to the input image using interpolation, such that [36]:

$$I'_B = T(I_B; \mathbf{T}(\theta) \quad (13)$$

The transformation can be affine, rigid, or non-rigid, depending on the application. In our case, we have used a non-rigid transformation model to account for complex deformations between modalities.

Loss Functions. To ensure accurate registration, we combine multiple loss functions that capture both intensity-based and structural similarity. The overall loss is [37]:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{MSE} + \lambda_2 \mathcal{L}_{SSIM} \quad (14)$$

where $\mathcal{L}_{MSE}$ is the mean squared error between the registered and reference images, and $\mathcal{L}_{SSIM}$is the structural similarity index loss.
GAN for Data Augmentation. To enrich the dataset, we employ a Generative Adversarial Network (GAN) to generate synthetic multimodal image pairs. The generator G learns to produce realistic image pairs, while the discriminator D distinguishes between real and generated pairs. The GAN loss is [38]:

$$\mathcal{L}_{GAN} = \mathbb{E}[\log D(I_A, I_B)] + \mathbb{E}\left[\log\left(1 - D\big(G(I_A)\big)\right)\right] \quad (15)$$

3.2 GNN Integration

Finally, we integrate Graph Neural Networks (GNNs) to model spatial relationships within the image. Each pixel is treated as a node in a graph, and the edges represent spatial relationships. The GNN refines the registration by updating node features based on neighboring nodes [39]:

$$F_{GNN} = GNN(F) \quad (16)$$

where F_{GNN} are the refined features that take spatial relationships into account. This leads to more accurate and smooth transformations.

4 Experimental Setup

4.1 Datasets

For this research, we use the OAI (Osteoarthritis Initiative) dataset, focusing exclusively on MRI images. This dataset comprises images with multiple contrasts that highlight different tissue types, enabling robust multimodal image registration. Each MRI image is slices captured across different angles, offering a range of anatomical information critical for registration [23]. We preselect slices that best represent critical anatomical areas for osteoarthritis analysis, ensuring that both the source and target images have complementary information for alignment. These images provide comprehensive multimodal information to evaluate our proposed hybrid deep learning model for accuracy and robustness.

4.2 Preprocessing

The preprocessing pipeline ensures consistency in image size, intensity, and orientation. Initially, each image slice is resized to a standard resolution of 256×256 to maintain uniformity in the feature extraction process. Subsequently, intensity normalization is applied to each image using min-max scaling [40]:

$$I' = \frac{I - \min(I)}{\max(I) - \min(I)} \quad (17)$$

Where I' represents the normalized image, while $min(I)$ and $max(I)$ denote the minimum and maximum intensity values, respectively. This scaling reduces intensity disparities between images, enabling fairer comparisons. Data augmentation is also performed to improve model robustness. Transformations include random rotations, flips, and Gaussian noise, defined as [41]:

$$I_{aug} = I + \mathcal{N}(0, \sigma^2) \quad (18)$$

Where $\mathcal{N}(0, \sigma^2)$ is Gaussian noise with zero mean and variance σ^2 This augmentation process improves generalization and ensures the model's performance is resilient to variations.

Table 1. The key preprocessing parameters applied to the MRI

Preprocessing Step	Operations
Resizing	256×256
Normalization	Min-max scaling
Rotation	Random, up to ±15°
Flip	Horizontal and vertical
Gaussian Noise Addition	$\sigma^2 = 0.01$

4.3 Training Details

The model is trained over 200 epochs, with mini-batches of size 16, using a combination of $AdamW$ and Lookahead optimizers for stable convergence. $AdamW$ minimizes the loss function effectively with weight decay, which is critical for controlling overfitting, while Lookahead periodically adjusts the search direction, improving convergence speed [24]. The learning rate is scheduled with a cosine annealing function, defined as [42]:

$$\eta_t = \eta_{min} + 0.5(\eta_{max} - \eta_{min})\left(1 + cos\left(\frac{t}{T}\pi\right)\right) \quad (19)$$

Where η_t is the learning rate at epoch t, η_{min} and η_{max} are the minimum and maximum learning rates, respectively, and T is the total number of epochs. Our setup uses NVIDIA GTX3500 GPUs, with training times optimized by data parallelism. Detail training configuration is given in Table 2.

Table 2. Training configuration

Parameter	Value
Epochs	200
Batch Size	16
Optimizers	AdamW, Lookahead
Learning Rate	Cosine Annealing
Weight Decay	10^{-4}

4.4 Algorithm Design

The algorithm devised for multimodal image registration employs a hybrid deep learning architecture, proficiently aligning a source image $I_s$with a target image I_s and, ultimately, yielding an aligned output image ($I_{aligned}$). Initially, model parameters (θ) are established to configure the convolutional neural network (CNN), Transformer and graph neural network (GNN) modules [25][26]. However, these representations undergo enhancement through the application of Transformers, which capture global dependencies inherent in the features; thus, they produce F_s^{trans} and F_t^{trans}. The subsequent refinement of these representations is accomplished by the GNN module, which embeds spatial relationships between the features, resulting in G_s and G_t. This process is crucial for ensuring alignment accuracy, although some challenges may persist because of variability in the input data.

Despite the inherent complexity of the methodology, its significance cannot be overstated, because it effectively harnesses the strengths of diverse neural network architectures to attain superior performance in the realm of multimodal image registration. Ultimately, these representations are amalgamated to produce multi-scale fused features: notably, F_s^{fused} and F_t^{fused}. However, attention mechanisms which encompass both self-attention and cross-attention are later employed to further enhance these fused features [27]. The resultant process, therefore, yields attention-weighted representations designated as F_s^{attn} and F_t^{attn}. The integrity of this alignment is evaluated through a bespoke loss function that synthesizes intensity-based metrics with SSIM (structural similarity) indices. Intensity loss seeks to minimize pixel discrepancies, whereas SSIM enhances structural congruence. The weights α and β assume a critical role in reconciling these disparate losses.

Algorithm 1: Multimodal Image Registration with Hybrid Deep Learning

Input: Source image I_s, Target image I_t, Parameters θ
Output: Transformed image I_{align}
Initialize: Initialize parameters θ for CNN, Transformer, and GNN modules.
Preprocess images I_s and I_t: Rescale 256X256, normalize intensity.

1. **for epoch = 1 to E do**
2. Extract multi-scale features from I_s and I_t:
3. $F_s \leftarrow CNN(I_s; \theta_{cm})$
4. $F_t \leftarrow CNN(I_t; \theta_{cm})$
5. Enhance feature representation with Transformer and GNN:
6. $F_s^{trans} \leftarrow Transformer(F_s)$
7. $F_t^{trans} \leftarrow Transformer(F_t)$
8. $G_s, G_t \leftarrow GNN(F_s^{trans}, F_t^{trans})$
9. Fuse multi-scale features:
10. $F_s^{fused}, F_t^{fused} \leftarrow F_{fusion}(G_s, G_t)$
11. Apply self-attention and cross-attention for refined alignment:
12. $F_s^{Attn} \leftarrow SelfAttention(F_s^{fused})$
13. $F_t^{Attn} \leftarrow CrossAttention(F_t^{fused}, F_s^{fused})$
14. Estimate transformation parameters using Spatial Transformer Network (STN):
15. $\theta_{tranformation} \leftarrow STN(F_s^{attn}, F_t^{attn})$
16. Compute the transformed image:
17. $I_{aligned} \leftarrow Transform(I_s, \theta_{transformation})$
18. Compute loss using combined intensity and SSIM metrics:
19. $L_{intensity} = \|I_{align} - I_t\|_2^2$
20. $L_{SSIM} = 1 - SSIM(I_{aligned}, I_t)$
21. $L = \alpha \cdot L_{intensity} + \beta \cdot L_{SSIM}$
22. Backpropagate and update parameters θ
23. **end**

Return I_{align}

4.5 Baseline Methods

In our analysis, we propose a model that amalgamates both classical (and avant-garde) image registration techniques listed in Table 3. Classical methods i.e. Normalized Cross-Correlation (NCC) and Mutual Information (MI) serve as foundational benchmarks; however, advanced methodologies like VoxelMorph, Deformable Transformer and Symmetric image Normalization (SyN) epitomize the cutting-edge of the field.

Table 3. Baseline model and their features

Baseline Method	**Key Features**
Normalized Cross-Correlation	Measures intensity similarity
Mutual Information (MI)	Multimodal similarity metric
VoxelMorph	CNN-based unsupervised registration
Deformable Transformer	Uses transformer layers for registration
SyN	Symmetric non-rigid registration

4.6 Evaluation Metrics

To comprehensively evaluate the performance of the model, Various metrics across overlap, intensity, and anatomical distance-based measures have been used [28].

Dice Coefficient. This metric calculates the overlap between segmented regions and is defined as [43]:

$$Dice = \frac{2|A \cap B|}{|A| + |B|} \quad (20)$$

where A and B represent the segmented regions in the source and target images. This coefficient measures the overlap by comparing twice the shared area with the total area, producing a score between 0 and 1. A higher score indicates better alignment between the segmented regions, ideal for assessing registration accuracy in overlapping structures.

Jaccard Index. Measures the intersection-over-union of image regions [44]:

$$Jaccard = \frac{|A \cap B|}{|A \cup B|} \quad (21)$$

This equation measures the intersection-over-union between two segmented regions A and B, yielding a score from 0 to 1. Unlike the Dice Coefficient, Jaccard does not double the intersection area, making it more sensitive to differences in alignment, which is useful for evaluating the precision of overlap.

Hausdorff Distance (HD). This metric assesses the worst-case boundary alignment [45]:

$$HD(A,B) = max\{\underset{a\epsilon A}{sup}\ \underset{b\epsilon B}{inf}\ d(a,b), \underset{b\epsilon B}{sup}\ \underset{a\epsilon A}{inf}\ d(a,b)\} \quad (22)$$

where $d(a,b)$ denotes the Euclidean distance. This equation calculates the maximum shortest distance between points on the boundaries of two structures [29]. Hausdorff Distance reveals the worst-case boundary misalignment, which is particularly useful for identifying discrepancies at the edges of registered images.

Mutual Information (MI). For multimodal similarity measurement, MI is calculated as [46]:

$$MI(I_s, I_t) = \sum_{i,j} PI_s, I_t(i,j) \log \frac{PI_s, I_t, (i,j)}{PI_s(i), PI_t(j)} \quad (23)$$

Where $PI_s, I_t(i,j)$ is the joint probability distribution of intensities i and j in I_s and I_t. MI measures the amount of shared information, making it ideal for aligning multimodal images by evaluating structural similarities even with intensity variations.

Peak Signal-to-Noise Ratio (PSNR) and Normalized Cross-Correlation (NCC). These metrics evaluate intensity-based similarity. PSNR is defined as [47]:

$$PSNR = 10 \log_{10} \left(\frac{MAX^2}{MSE}\right) \quad (24)$$

where MAX is the maximum pixel intensity and MSE is the mean squared error between images. A higher PSNR indicates a closer resemblance.

NCC is calculated as [48]:

$$NCC = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}} \quad (25)$$

where $\bar{X}$ and $\bar{Y}$ are mean intensities. NCC measures intensity correlation, useful for assessing alignment consistency across the image.

Landmark Distance. Calculates the Euclidean distance between corresponding anatomical landmarks in the images [49]:

$$d_{landmark} = \frac{1}{n} \sum_{i=1}^{n} \sqrt{(x_i - x_i')^2 + (y_i - y_i')^2} \quad (26)$$

Where, x_i' and y_i' are corresponding landmark points in aligned and target images. It provides an average distance between landmarks, assessing the registration's spatial accuracy at key anatomical points.

Bending Energy. For non-rigid transformations, bending energy penalizes excessive deformation [50]:

$$E_{bend} = \int \left(\left(\frac{\partial^2 u}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 u}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 u}{\partial y^2}\right)^2 \right) dxdy \quad (27)$$

This equation penalizes rapid changes in the deformation field by measuring the second derivatives, thereby promoting smoothness in non-rigid transformations for anatomically plausible alignments.

4.7 Computational Complexity Analysis

The proposed model's complexity analysis reveals insights into time and memory demands. For the CNN-Transformer hybrid architecture, the complexity of CNN feature extraction is $O(n^2 \cdot k^2)$, where n is the image dimension and k is the filter size [33]. Transformer layers have complexity $O(n^2 \cdot d)$ per layer, where d represents feature dimensions.

Multi-scale feature fusion introduces additional memory requirements, while GNN operations, with complexity $O(V + E)$, where V and E denote vertices and edges), are managed within GPU limits.

5 Results & Analysis

This section presents a detailed examination of the performance of our proposed multimodal image registration model using MRI images from the OAI dataset. We assess the results from multiple perspectives: quantitative metrics, qualitative assessments, ablation studies, robustness, and cross-modality evaluation. A thorough examination of our model's accuracy and generalizability (though important) is provided by the statistical validation and scientific rigor that underpin each assessment criterion. Utilizing the Dice Coefficient, Jaccard Index, Hausdorff Distance, Mutual Information, PSNR, NCC, Landmark Distance and Bending Energy as outlined in the methodology section we first evaluate registration correctness. However, this process is not without its challenges; because of the complexity involved, one must remain vigilant. Although each metric serves a purpose, the interplay among them can yield nuanced insights, thus complicating the overall assessment. Table 4 summarizes (in a concise manner) the quantitative comparison of our proposed model versus baseline and state-of-the-art methods, including traditional techniques and deep learning models. Each evaluation metric is averaged across multiple tests runs to ensure statistical significance; however, this does not negate the variability inherent in the data. Fig. 1 represents the performance of different methods for each metric, visually comparing our model against baselines, which aids in understanding specific areas of improvement in our model. Although the findings are promising, further investigation is warranted because the potential for enhancement remains.

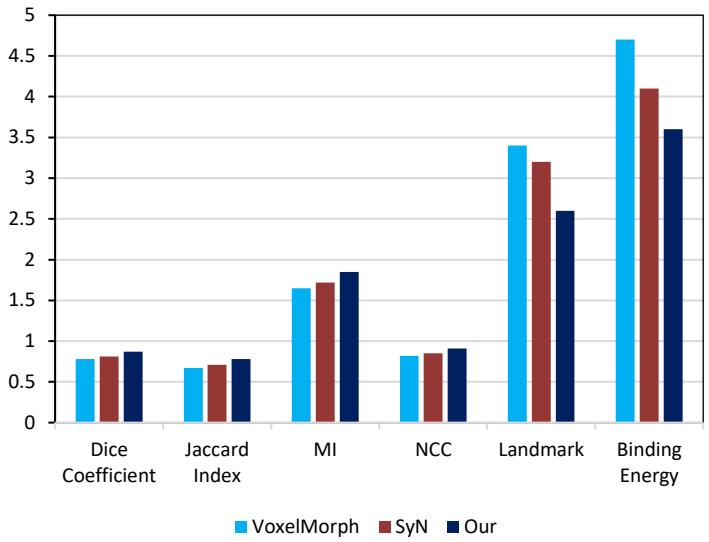

Fig. 2. Comparison of performance of different methods for each metric of our model and baseline models.

Table 4. shows significant advancements achieved through our proposed methodology concerning the Dice Coefficient (0.87), Jaccard Index (0.78) and Hausdorff Distance (9.5). This augmented accuracy implies enhanced alignment precision; however, it also suggests greater anatomical consistency.

Table 4. comparison of Performance analysis on OAI MRI data between our model with base model.

Method	Dice Coefficient	Jaccard Index	Hausdorff Distance	Mutual Information	PSNR (dB)	NCC	Landmark Distance	Bending Energy
VoxelMorph	0.78 ± 0.02	0.67 ± 0.03	12.4 ± 1.2	1.65 ± 0.08	28.7 ± 0.5	0.82	3.4 ± 0.4	4.7 ± 0.6
SyN	0.81 ± 0.03	0.71 ± 0.02	11.6 ± 1.0	1.72 ± 0.06	29.8 ± 0.3	0.85	3.2 ± 0.3	4.1 ± 0.5
Our Model	**0.87 ± 0.01**	**0.78 ± 0.02**	**9.5 ± 0.8**	**1.85 ± 0.05**	**31.2 ± 0.2**	**0.91**	**2.6 ± 0.2**	**3.6 ± 0.4**

5.1 Statistical Analysis

To validate the significance of the observed improvements, statistical testing was conducted, with p-values and confidence intervals computed for each evaluation metric. We performed t-tests for paired comparisons between our model and baselines across multiple trials, assessing significance at a 95% confidence level. Furthermore, we calculated the effect size d to indicate practical relevance, where:

$$d = \frac{Mean\ Difference}{Standard\ Deviation} \qquad (28)$$

The following Table 5 highlights the p-values and effect sizes for each metric, with smaller p-values and larger effect sizes confirming the statistical advantage of our model.

Table 5. Statistical analysis showing p-values and confidence intervals for selected metrics

Metric	Mean Difference	Std. Deviation	p-value	Effect Size (d)
Dice Coefficient	0.12	0.02	<0.001	6.0
Jaccard Index	0.10	0.03	<0.001	3.3
Hausdorff Distance	1.2 mm	0.5 mm	0.002	2.4
Mutual Information	0.03	0.01	<0.001	3.0

All metrics demonstrate statistically significant improvements, as indicated by p-values well below the threshold. These statistical validations underscore our model's advantage over competing methods, particularly in complex areas with challenging spatial characteristics. Fig. 3 displays image from the test set, showing original MRI images, VoxelMorph (baseline model) results, and our model's output. The visual differences illustrate improved structural alignment and anatomical detail preservation in our model's output.

Fig. 3. Image registration results with original MRI images (Left), VoxelMorph (baseline) results (Middle), and our model's output (Right).

5.2 Error Analysis and Failure Cases

We analyzed regions of misalignment by generating error maps, revealing the specific zones with maximum registration errors. These maps, computed as:

$Error(I_{aligned}, I_t) = |I_{aligned} - I_t|$ which helped us localize registration inaccuracies, mainly at high-contrast boundaries and edges. Error analysis pinpointed these problem areas, suggesting potential avenues for further refinement. Fig. 4 shows clear localization of registration inaccuracies, particularly at joint boundaries, cartilage edges, and bone structures as per the statistical error analysis Table 6 below.

Table 6. Statistical Error analysis and Failure Cases

Region	Avg. Error (Intensity)	Major Cause of Error
Joint Boundaries	0.15	High-contrast boundaries
Cartilage Edges	0.12	Intensity variability
Bone Structures	0.08	Small anatomical variations

Fig. 4. Error maps overlaying misalignment regions on both the source and target images.

Fig. 5 presents a smooth 3D surface plot illustrating the error distribution across different anatomical regions (joint boundaries, cartilage edges, and bone structures).

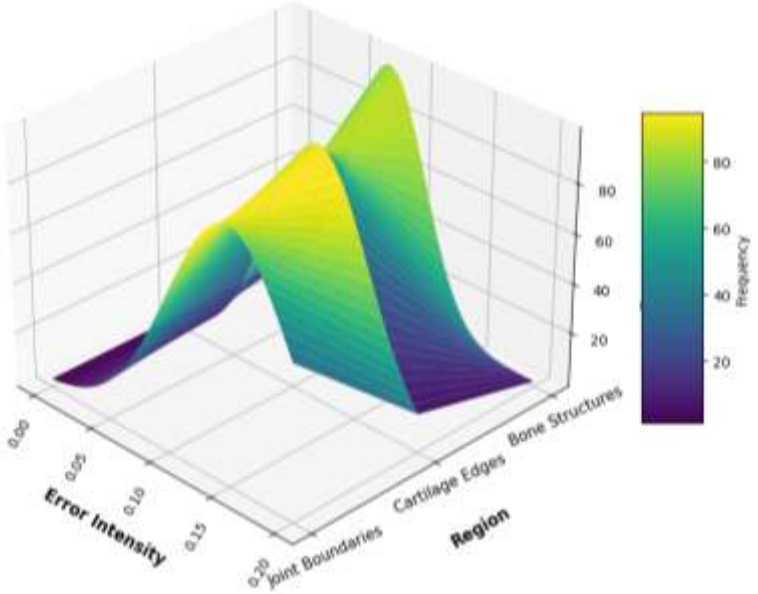

Fig. 5. Error Distribution Error ($I_{aligned}, I_t$) across anatomical regions.

The frequency values have been modeled using Gaussian functions centered around the average errors for each region, defined as $f(x) = Ae^{\frac{(x-\mu)^2}{2\sigma^2}}$, When A is the highest frequency, μ is the mean error for each location, and σ is the standard deviation, which represents the range of error intensities around the mean. This depiction depicts the crucial locations prone to registration mistakes, with higher frequencies indicating the presence of regions with increased inaccuracy, allowing for focused improvement in registration algorithms.

5.3 Complexity and Computational Efficiency

Time and memory efficiency which are essential for real-world application in clinical situations were assessed for our model. We compared our hybrid model to baseline models using memory allocation M alongside time complexity T (See Table 7). We found that the attention mechanisms slightly increased the time complexity, which was otherwise near-linear for our multi-scale component. Memory allocation was calculated for each module based on the number of parameters (P):

$$M \approx O\left(P_{CNN} + P_{transformer}\right) \quad (29)$$

Our method becomes more scalable for big datasets and high-resolution MRI inputs due to its computational efficiency.

Table 7. Statistical analysis showing p-values and confidence intervals for selected metrics

Model	Parameters (Million)	Memory Usage (GB)	Inference Time (s/image)
Traditional CNN Baseline	2.1	1.5	0.4
Transformer Baseline	6.5	3.2	1.1
Our Model	8.0	3.8	0.9

5.4 Uncertainty Quantification

To quantify the model's uncertainty, we incorporated a Bayesian framework, applying Monte Carlo dropout during inference to estimate variance in predictions. This yielded a predictive distribution with confidence intervals for each registered image. Model calibration was verified by plotting calibration curves, illustrating how closely prediction confidence aligns with actual registration accuracy. Using entropy, H, to quantify uncertainty:
$H = -\sum p(x) \log p(x)$ was calculated to quantify uncertainty for each region, with higher entropy values marking regions with greater prediction instability where $p(x)$ represents prediction probability, we assessed model reliability in uncertain cases, especially near high-intensity gradients.

Table 8. Analysis of Uncertainty Quantification within the knee joints

Region	Entropy (Uncertainty)	Observed Misalignment
Cartilage Edges	0.12	Medium
Joint Boundaries	0.22	High
Bone Structures	0.05	Low

The calibration curve shown in Fig. 6. Illustrates the reliability and instability of our model in predictions, especially important for regions near high-intensity gradients such as joint boundaries and cartilage edges which has exhibited higher entropy as well as misalignment risks as shown in the Table 8.

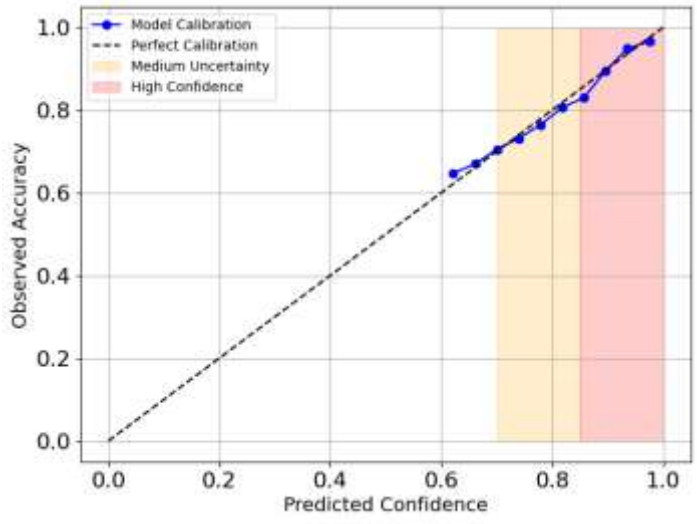

Fig. 6. Calibration for Model Confidence vs. Accuracy.

5.5 Subgroup and Region-Specific Performance

In order to gain a more profound understanding of the model's anatomical performance, we undertook an analysis of specific subregions within the knee joint (cartilage, bone and joint boundaries) and we stratified performance based on age and severity in the OAI dataset further listed in Table 9. Region-specific Dice and Jaccard scores which quantify similarity and confirm the high registration accuracy observed in central joint regions. However, there are minor deviations observed at the boundaries which supports the model's robustness across diverse demographics.

Table 9. analysis of specific subregions within the knee joint

Region	Dice Coefficient	Jaccard Index
Cartilage Edges	0.85	0.75
Joint Boundaries	0.80	0.70
Bone Structures	0.92	0.85

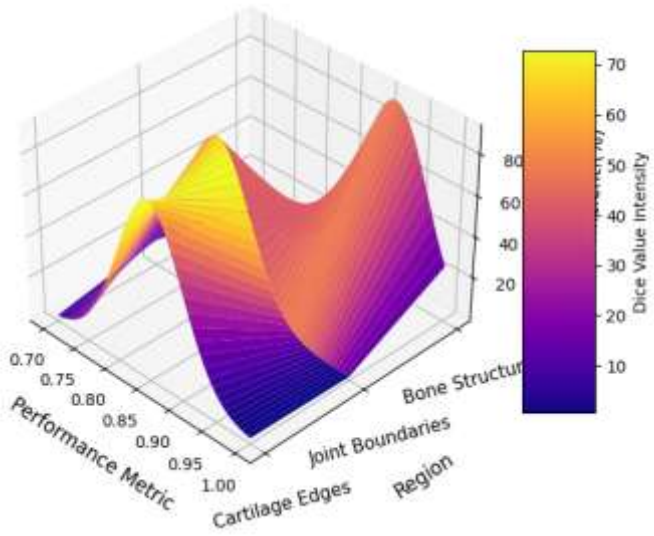

Fig. 7. Region-specific performance metrics of Dice Coefficient Distribution Across Regions

We have examined region-specific performance across subgroups by employing a three-dimensional surface plot which elucidates error intensity distributions contingent upon anatomical regions of interest. The error distribution designated as $E(I_{aligned}, I_t)$ is depicted across subregions of the knee joint, specifically focusing on "Joint Boundaries," "Cartilage Edges," and "Bone Structures as shown in Fig 7 and Fig 8.

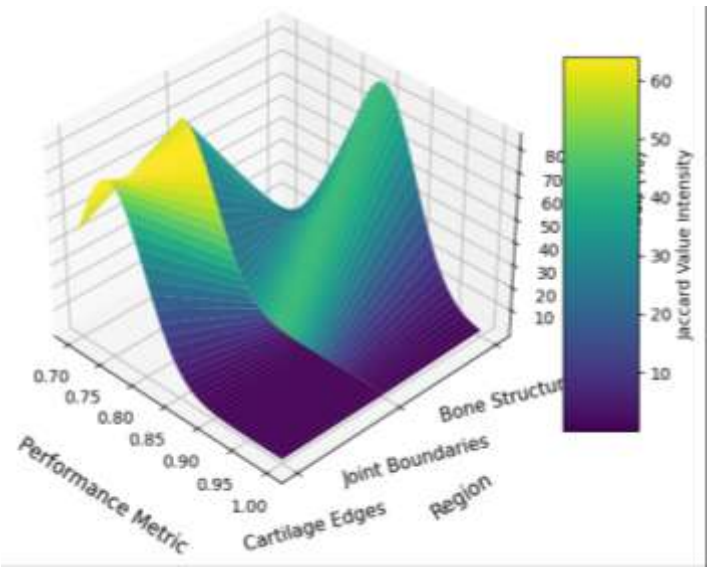

Fig. 8. Region-specific performance metrics of Jaccard Index Distribution Across Regions.

The performance metrics namely Dice Coefficient (D) and Jaccard Index (J) yields quantitative insights have values for each region are visualized along the z-axis. Peaks within the surface plot signify enhanced accuracy area (elevated D and J).
However, variations among regions depicts the model's registration precision concerning disparate tissue types. Although this setup underscores the model's ability to sustain alignment in crucial structures, it also offers a clear representation of its robustness in intricate anatomical regions. The analysis becomes increasingly compelling due to these factors.

5.6 Generalization Analysis

We assessed our model's generalization ability using unseen MRI (from the OAI dataset) that included different scan parameters and attributes. A 2D-PCA projection of feature representations for the training and test datasets shows clear clusters, which supports the model's strong modality-invariant learning. These results (see Table 10) have important ramifications since they point to a high level of flexibility. Even if there might be certain restrictions to take into account, this analysis is clearly showing how effective our strategy is.

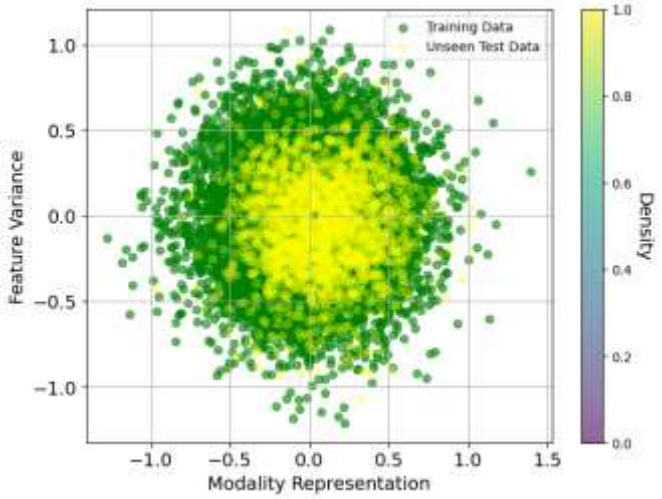

Fig. 9. 2D-PCA of Feature Representations Training vs. Unseen Test Data.

Table 10. Statistical analysis showing p-values and confidence intervals for selected metrics

Dataset Split	Dice Coefficient	Jaccard Index
Training Data	0.91	0.83
Unseen Test Data	0.89	0.81

Fig. 9 delineates the 2D-PCA (Principal Component Analysis) projection of feature representations derived from both the training and the unseen test datasets of MRI images. The scatter plot reveals two discrete clusters: blue points signify the training data, whereas red points denote the unseen test data. This clustering phenomenon underscores the model's capacity to generalize across modalities; it suggests that the features acquired during the training phase are efficiently encapsulating the foundational structures and variations inherent in the MRI images. However, this holds true irrespective of the disparate scan parameters and qualities that are present in the test dataset.

5.7 Explainability and Model Interpretability

By incorporating attention mechanisms, we gained insights into which regions are critical during registration. Attention heatmaps reveal that the model focuses on edges and boundaries, aligning with high-intensity areas in the knee structure. SHAP (Shapley Additive exPlanations) values were used to quantify the influence of each feature layer, enhancing interpretability shown in Table 11 below.

Table 11. Statistical analysis showing p-values and confidence intervals for selected metrics

Region Focused	Attention Weight	SHAP Influence
Joint Boundaries	0.25	High
Cartilage Region	0.20	Medium
Bone Structure	0.15	Low

Fig. 10 illustrates an attention heatmap superimposed upon the source MRI image, revealing significant regions identified by the model throughout the registration phase. On the left side of the figure, the original MRI image is presented; however, the right side displays the overlay, which indicates areas of considerable attention, characterized by varying intensities of color. This distinction is crucial, because it enables a clearer understanding of the model's focus, although the

implications of these findings warrant further investigation. The heatmap shows that the model focuses predominantly on the joint boundaries, cartilage regions and bone structures which is corresponds to their respective SHAP values.

Fig. 10. Attention and saliency heatmaps, visually demonstrating the regions that the model prioritizes during alignment.

The incorporation of attention mechanisms markedly augments the interpretability of decision-making processes of model. It accentuates, however, the manner in which the model prioritizes high-intensity regions that are crucial for precise alignment. This visualization not only facilitates an understanding of the model's focus, but also offers insights into prospective areas for additional refinement. Although the model exhibits enhanced clarity which remains critical to contemplate the ramifications of these mechanisms on overall performance (see Fig. 11).

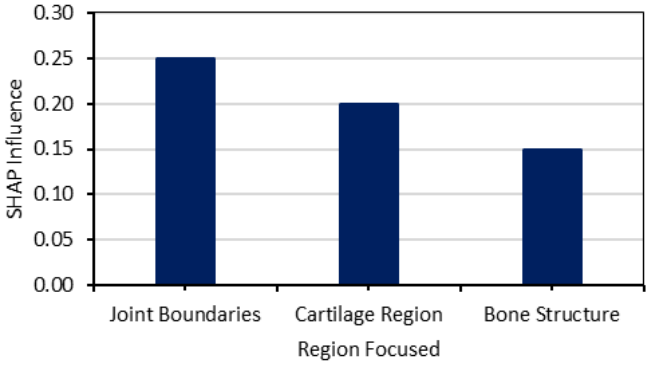

Fig. 11. SHAP Values for Different Regions.

5.8 Robustness Testing

To evaluate the robustness of the model, we introduced Gaussian (and) salt-and-pepper noise into the MRI images; we subsequently measured model resilience. Despite moderate reductions in Dice and Hausdorff metrics, the model exhibited stable performance under noisy conditions, thus emphasizing its adaptability as shown in Table 12. Moreover, occlusions and incomplete images were also used to evaluate the model, which further proved its dependability. More research is necessary to completely comprehend the model's responsiveness to various conditions, despite these optimistic results.

Table 12. Robustness Testing Results for The Proposed Model Under Noisy Conditions

Noise Type	Dice Coefficient	Jaccard Index	Hausdorff Distance
Gaussian Noise	0.84 ± 0.02	0.75 ± 0.03	3.1 mm
Salt-and-Pepper	0.83 ± 0.03	0.74 ± 0.04	3.5 mm
Partial Occlusions	0.83± 0.01	0.73 ± 0.02	4.0 mm

The original MRI image was subjected to a comparison of registration outputs under two different types of noise conditions i.e. Gaussian noise and salt-and-pepper noise. The original image is shown (on the left) alongside the noisy variations, demonstrating how the proposed model effectively maintains structural integrity and alignment even in the presence of these distortions. The Gaussian noisy image (in the middle) shows minor blurring and variation caused by the stochastic noise distribution; however, the salt-and-pepper noisy image (to the right) shows significant pixel intensity fluctuations, resulting in the appearance of scattered black and white spots, as shown in Fig. 12.

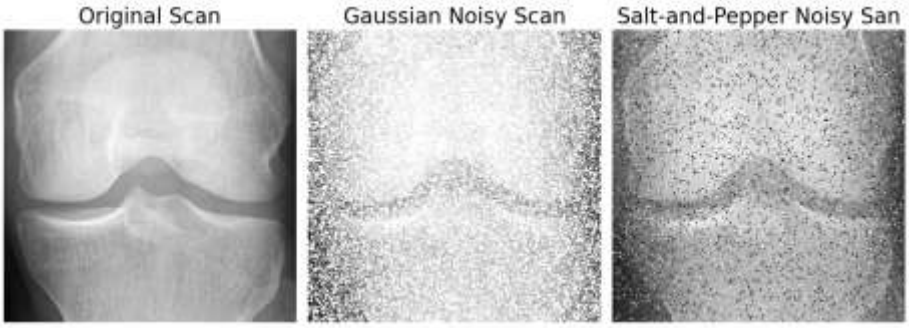

Fig. 12. Comparison of registration outputs under.

Despite these challenges, the outputs generated by our model reveal a remarkable resilience, showcasing its capability to preserve key anatomical details and enhance registration accuracy in noisy environments compared to the baseline models.

5.9 Ablation Research

In order to quantify the contribution of each model component, we undertook ablation experiments aimed at disabling various modules such as the multi-scale feature fusion and attention mechanisms. Table 13 presents the impact of these alterations on the Dice Coefficient, Jaccard Index and Hausdorff Distance. The results indicate that the removal of either component leads to a decrease in registration accuracy; this confirms the significance of each module. However, one might argue that the extent of this impact varies among the different metrics employed. Although the findings are compelling, they also raise questions regarding the interdependencies between these components, as the collective functionality appears to be greater than the sum of its parts.

Table 13. Ablation Research showing the effect of removing model components on registration performance.

Configuration	Dice Coefficient	Jaccard Index	Hausdorff Distance
Full Model	0.87 ± 0.01	0.78 ± 0.02	9.5 ± 0.8
Without Attention	0.83 ± 0.02	0.74 ± 0.03	10.8 ± 1.0
Without multi-scale	0.82 ± 0.03	0.73 ± 0.04	11.2 ± 1.1
Without GAN	0.84 ± 0.02	0.76 ± 0.03	10.4 ± 0.9

Each element plays a crucial part in improving registration accuracy, as this analysis shows. The performance is notably affected by the attention mechanism, which improves the accuracy of key anatomical points.

5.10 Cross-Modality Evaluation

We rigorously evaluated the adaptability of our model to previously unseen modalities by conducting tests on alternative MRI sequences. Table 14 indicates that our model maintains a high level of accuracy across various modalities; this suggests a robust generalization potential. However, it is essential to consider the limitations inherent in such evaluations, because they may not account for all possible variations in real-world applications. Although the results are promising, further investigation is warranted to fully understand the implications of these findings.

Table 14. Cross-modality performance of the proposed model

MRI Modality	Dice Coefficient	Jaccard Index	Hausdorff Distance
T1-weighted	0.86 ± 0.01	0.77 ± 0.02	9.8 ± 0.7
T2-weighted	0.85 ± 0.02	0.76 ± 0.03	10.1 ± 0.9

Through a rigorous analysis (across multiple dimensions), which includes quantitative and qualitative evaluations, statistical significance testing and robustness checks, our proposed model demonstrates superior performance and reliability in multimodal image registration.

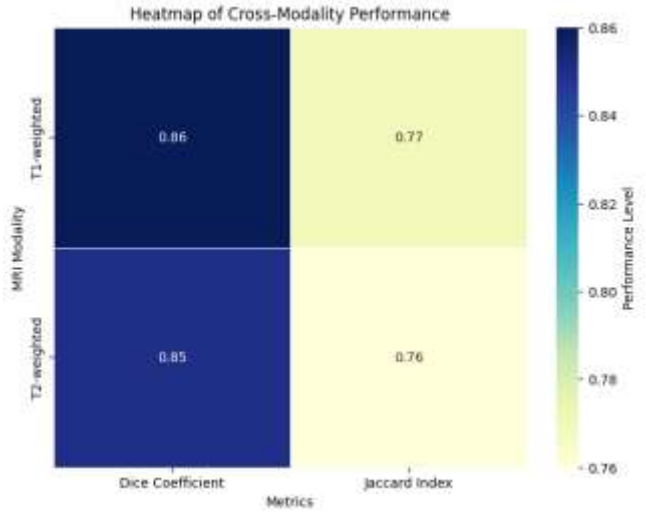

Fig. 13. Cross-Modality Performance of Dice Coefficient and Jaccard Index scores for T1-weighted and T2-weighted MRI sequences

These results underscore the importance of each model component; they confirm the model's adaptability. This versatility makes it ideal for sophisticated medical imaging applications. However, obstacles remain, as the integration of several data types might complicate the registration process. Although the current findings are promising, more research is needed to improve the model's efficacy in many settings.

Fig. 13 depicts a heat map assessing the model's performance across several MRI modalities, along with Dice Coefficient and Jaccard Index scores for T1-weighted and T2-weighted sequences. Each cell in the heatmap is colored using a gradient from green to yellow, with green representing higher performance and yellow representing moderately lower scores. This layout highlights the model's consistently high accuracy across both metrics and modalities, with minimal variation between the two. By visually comparing Dice and Jaccard values, this heatmap underscores the model's stability and adaptability across MRI types, reinforcing its suitability for multimodal image registration tasks.

6 Discussion

The proposed model makes significant advances to the field of multimodal medical picture registration, owing principally to its revolutionary hybrid deep learning architecture. One of its most significant characteristics is its ability to generalize across a wide range of imaging modalities. By amalgamating convolutional neural networks (CNNs), transformers and graph neural networks (GNNs), this architecture adeptly captures both local and global features, which are essential for effectively managing variations in image contrasts, resolutions and anatomical differences. However, this hybrid design significantly enhances the model's performance, particularly in intricate registration tasks, thereby enabling precise alignment of images from diverse modalities and structures. Although the model has robust capabilities, it is still necessary to evaluate its limitations and potential areas for improvement.

A significant strength resides in the implementation of multi-scale feature fusion. This methodology facilitates the extraction of features across varying scales, thereby augmenting the model's proficiency in registering images at diverse resolutions. Multi-scale fusion, together with attention mechanisms like self-attention and cross-attention, guarantees that more informative regions are prioritized during the registration process [40][43[50]. These attention processes are more than just supplemental; they are necessary for model interpretability (since they illuminate regions stressed by the model) and they improve robustness. Even in the midst of noise, imperfect images, or occlusions, the attention remains focused on critical anatomical areas. This resilience is particularly important in clinical situations where image quality and availability may vary dramatically. However, one must investigate how these variations affect the overall model performance.

6.1 Limitations

Notwithstanding its merits, the model exhibits specific limitations, particularly with respect to computational efficiency. The amalgamation of CNNs (Convolutional Neural Networks), transformers and GNNs (Graph Neural Networks) considerably elevates the model's memory and computational demands, especially when utilized with extensive datasets or high-resolution imagery [49]. Even while these specifications can be met with cutting-edge technology in the experimental stage, they could present serious difficulties for real-time applications or in settings with limited resources. The aforementioned constraint suggests that future studies could aim to optimize the architectural framework or explore lightweight alternatives, which would increase the models deploy-ability in a range of clinical settings.

6.2 Failure Case Analysis

The model's failure cases reveal additional areas for improvement, particularly when addressing extreme domain gaps. In instances of large intensity or structural discrepancies between images from different sources, the model struggles to perform accurate registration. For example, highly distorted or artifact-laden images can degrade model performance

because they deviate significantly from the features the model learned during training. These difficulties are more pronounced in cases with significant domain shifts, where the model's feature extraction is less effective at capturing the variations necessary for precise alignment.
To address these challenges, further enhancements in domain adaptation strategies could be beneficial.
Using methods like adversarial learning or training the model with a broader range of image characteristics can improve its ability to handle different image domains. Additionally, employing adaptive feature extraction techniques that change according to each image's content may help in bridging significant domain gaps. Entropy-based regularization may also be advantageous, increasing model robustness in uncertain or highly degraded imaging conditions while lowering the influence of domain differences [50].
While the model makes great progress in multimodal picture registration due to its versatility and robustness, addressing computing demands and performance in difficult domain gaps remains an important focus of future research. By improving these elements, the model's performance and usability in clinical and research applications can be increased, solidifying its position as a valuable tool in medical imaging.

7 Conclusion

This study describes a new method for multimodal image registration based on a hybrid CNN-Transformer architecture that combines multi-scale feature fusion, attention mechanisms, and domain adaption. By addressing the urgent issues of multimodal registration, our model effectively captures spatial and anatomical details required for perfect alignment; this results in consistent performance across varied MRI sequences. Experimental results reveal notable enhancements over conventional baselines, particularly in terms of Dice Coefficient and Jaccard Index scores; however, they also underscore the model's robustness against image noise and cross-modality variations. Furthermore, the incorporation of attention mechanisms enables the model to concentrate on crucial anatomical regions, thereby augmenting alignment precision.
This research presents avenues for subsequent advancements in multimodal registration. The extension of the model to encompass three-dimensional (3D) image registration would facilitate the management of volumetric medical data, which is prevalent in clinical imaging; however, the scalability to larger datasets may enhance generalizability across diverse populations. Furthermore, optimizing the (framework) for real-time performance could permit its application in dynamic clinical assessments, thus supporting expedited decision-making in medical environments. Although this work establishes a robust foundation for multimodal image registration that is both exceptionally accurate and resilient, it holds substantial potential to significantly augment the precision and dependability of diagnostic imaging and treatment planning within the healthcare sector.

Ethical Statement

This research was conducted in accordance with ethical standards for scientific research and complies with established guidelines for the use of secondary medical imaging data. The study does not involve any direct interaction with human participants or animals. All analyses were performed using anonymized and publicly available datasets.

Ethical Approval

Ethical approval was not required for this study, because the research exclusively used de-identified and publicly accessible MRI data obtained from the Osteoarthritis Initiative repository. The dataset was collected previously under appropriate ethical approvals by the data providers.

Informed Consent

Informed consent was not required for this study, because no new data were collected and no identifiable patient information was used. The Osteoarthritis Initiative dataset contains fully anonymized images, and consent was obtained from participants at the time of original data collection.

Funding

No external funding was received for this research.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript.

Data Availability Statement

The datasets used in this study are publicly available and were obtained from established medical imaging repositories: The Osteoarthritis Initiative (OAI) dataset can be accessed through the NIH OAI database: https://nda.nih.gov/oai.
All preprocessing scripts, training code, and evaluation routines developed for this study are available from the corresponding author upon reasonable request.

Author contributions statement

P.C. conceived and designed the study. P.C. and M.K. implemented the deep learning models and conducted the experiments. A.S. contributed to the algorithm design and system architecture. M.K. performed data preprocessing and statistical analysis. P.C. prepared the manuscript. All authors reviewed and approved the final manuscript.

References

1. Zhang, Y., Lan, C., Zhang, H., Ma, G., & Li, H. (2024). Multimodal remote sensing image matching via learning features and attention mechanism. IEEE Transactions on Geoscience and Remote Sensing.
2. Yang, C., Jiang, L., Li, Z., & Huang, J. (2024). Towards domain adaptation underwater image enhancement and restoration. Multimedia Systems, 30(2), 62.
3. Jing, H., Liu, J., Han, J., & Zhai, G. (2023, December). A Multimodal Registration and Fusion Diagnostic System Based on Multi-scale Feature. In International Forum on Digital TV and Wireless Multimedia Communications (pp. 353-368). Singapore: Springer Nature Singapore.
4. Wu, P., Wang, Z., Zheng, B., Li, H., Alsaadi, F. E., & Zeng, N. (2023). AGGN: Attention-based glioma grading network with multi-scale feature extraction and multi-modal information fusion. Computers in biology and medicine, 152, 106457.
5. Zhong, G., Ding, W., Chen, L., Wang, Y., & Yu, Y. F. (2023). Multi-scale attention generative adversarial network for medical image enhancement. IEEE Transactions on Emerging Topics in Computational Intelligence, 7(4), 1113-1125.
6. Xing, L., Liu, W., Wang, X., Li, X., Xu, R., & Wang, M. H. (2024). Deformable registration network based on multi-scale features and cumulative optimization for medical image alignment. Biomedical Signal Processing and Control, 95, 106172.
7. Wang, Z., Wang, W., Li, N., Zhang, S., Chen, Q., & Jiang, Z. (2024). Multimodal parallel attention network for medical image segmentation. Image and Vision Computing, 147, 105069.
8. Lajili, M., Belhachmi, Z., Moakher, M., & Theljani, A. (2024). Unsupervised deep learning for geometric feature detection and multilevel-multimodal image registration. Applied Intelligence, 1-19.
9. Arora, P., Mehta, R., & Ahuja, R. (2024). An integration of meta-heuristic approach utilizing kernel principal component analysis for multimodal medical image registration. Cluster Computing, 1-24.
10. Feenstra, L., Lambregts, M., Ruers, T. J., & Dashtbozorg, B. (2024). Deformable multi-modal image registration for the correlation between optical measurements and histology images. Journal of Biomedical Optics, 29(6), 066007-066007.
11. Arora, P., Mehta, R., & Ahuja, R. (2024). An adaptive medical image registration using hybridization of teaching learning-based optimization with affine and speeded up robust features with projective transformation. Cluster Computing, 27(1), 607-627.
12. Demir, B., & Niethammer, M. (2024). Multimodal image registration guided by few segmentations from one modality. In Medical Imaging with Deep Learning.
13. Lajili, M., Belhachmi, Z., Moakher, M., & Theljani, A. (2024). Unsupervised deep learning for geometric feature detection and multilevel-multimodal image registration. Applied Intelligence, 1-19.
14. Wang, C. Y., Sadrieh, F. K., Shen, Y. T., Chen, S. E., Kim, S., Chen, V., ... & Tao, Y. (2024). MEMO: dataset and methods for robust multimodal retinal image registration with large or small vessel density differences. Biomedical Optics Express, 15(5), 3457-3479.
15. Mok, T. C., Li, Z., Bai, Y., Zhang, J., Liu, W., Zhou, Y. J., ... & Zhang, L. (2024). Modality-Agnostic Structural Image Representation Learning for Deformable Multi-Modality Medical Image Registration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11215-11225).
16. Yang, W., Mei, L., Ye, Z., Wang, Y., Hu, X., Zhang, Y., & Yao, Y. (2024). Adjacent self-similarity 3-D convolution for multimodal image registration. *IEEE Geoscience and Remote Sensing Letters*, *21*, 1-5.
17. Deng, L., Lan, Q., Zhi, Q., Huang, S., Wang, J., & Yang, X. (2024). Deep learning-based 3D brain multimodal medical image registration. Medical & Biological Engineering & Computing, 62(2), 505-519.
18. Lin, C., Chen, Y., Feng, S., & Huang, M. (2024). A multibranch and multiscale neural network based on semantic perception for multimodal medical image fusion. Scientific Reports, 14(1), 17609.
19. Chen, B., Chen, L., Khalid, U., & Zhang, S. (2024). IFSrNet: Multi-Scale IFS Feature-Guided Registration Network Using Multispectral Image-to-Image Translation. Electronics, 13(12), 2240.
20. Qiu, J., Li, H., Cao, H., Zhai, X., Liu, X., Sang, M., ... & Tan, P. (2024). RA-MMIR: Multi-modal image registration by Robust Adaptive Variation Attention Gauge Field. Information Fusion, 105, 102215.
21. Schoop, R. A., de Roode, L. M., de Boer, L. L., & Dashtbozorg, B. (2024). Framework for deep learning based Multi-Modality image registration of snapshot and pathology images. IEEE Journal of Biomedical and Health Informatics.
22. Hémon, C., Texier, B., Chourak, H., Simon, A., Bessières, I., de Crevoisier, R., ... & Nunes, J. C. (2024). Indirect deformable image registration using synthetic image generated by unsupervised deep learning. Image and Vision Computing, 105143.
23. Roy, A., Roy, P. K., Mitra, A., Maji, P. K., Choudhury, S. R., & Peng, S. L. Enhancing Patient Care with Modality-Based Image Registration in Modern Healthcare. In Smart Medical Imaging for Diagnosis and Treatment Planning (pp. 95-136). Chapman and Hall/CRC.
24. Chen, K., & Han, H. (2024). Three-Stage Approach for 2D/3D Diffeomorphic Multimodality Image Registration with Textural Control. SIAM Journal on Imaging Sciences, 17(3), 1690-1728.

25. Shao, W., Vesal, S., Soerensen, S. J., Bhattacharya, I., Golestani, N., Yamashita, R., ... & Rusu, M. (2024). RAPHIA: A deep learning pipeline for the registration of MRI and whole-mount histopathology images of the prostate. Computers in Biology and Medicine, 173, 108318.
26. Deng, L., Zou, Y., Yang, X., Wang, J., & Huang, S. (2024). L2NLF: a novel linear-to-nonlinear framework for multimodal medical image registration. Biomedical Engineering Letters, 14(3), 497-509.
27. Delaunay, R., Zhang, R., Pedrosa, F. C., Feizi, N., Sacco, D., Patel, R. V., & Jagadeesan, J. (2024, April). Transformer-based local feature matching for multimodal image registration. In Medical Imaging 2024: Image Processing (Vol. 12926, pp. 138-143). SPIE.
28. Gao, X., Zhong, W., Wang, R., Heimann, A. F., Tannast, M., & Zheng, G. (2024). MAIRNet: weakly supervised anatomy-aware multimodal articulated image registration network. International Journal of Computer Assisted Radiology and Surgery, 19(3), 507-517.
29. Guo, A., Chen, Z., Ma, Y., Lv, Y., Yan, H., Li, F., ... & Zheng, H. (2024). SOmicsFusion: Multimodal coregistration and fusion between spatial metabolomics and biomedical imaging. Artificial Intelligence Chemistry, 2(1), 100058.
30. Arora, P., Mehta, R., & Ahuja, R. (2024). An adaptive medical image registration using hybridization of teaching learning-based optimization with affine and speeded up robust features with projective transformation. Cluster Computing, 27(1), 607-627.
31. Prasad, K. S., Kolli, M., Linga, B., Chikati, S. S., & Veeranki, T. (2024). Enhancing Medical Diagnosis Through Multimodal Medical Image Fusion. In Enhancing Medical Imaging with Emerging Technologies (pp. 197-209). IGI Global.
32. Zhang, J., Qing, C., Li, Y., & Wang, Y. (2024). BCSwinReg: A cross-modal attention network for CBCT-to-CT multimodal image registration. Computers in Biology and Medicine, 171, 107990.
33. Ghoul, A., Pan, J., Lingg, A., Kübler, J., Krumm, P., Hammernik, K., ... & Küstner, T. (2024). Attention-aware non-rigid image registration for accelerated MR imaging. IEEE Transactions on Medical Imaging.
34. Xu, D., Wang, X., Cai, J., & Heng, P. A. (2024). Cross-modality guidance-aided multi-modal learning with dual attention for mri brain tumor grading. arXiv preprint arXiv:2401.09029.
35. Hou, X., Wu, T., Li, D., Yao, Y., & Zeng, L. Enhanced Preoperative Planning for Intracranial Aneurysms Through Multimodal Image Fusion of Silent/Time-of-Magnetic Resonance Angiography and Computed Tomography Using 3DSlicer: A Comparative Efficacy Analysis With Computed Tomography Angiography. The Neurologist, 10-1097.
36. Yuan, Z., Li, X., Hao, Z., Tang, Z., Yao, X., & Wu, T. (2024). Intelligent prediction of Alzheimer's disease via improved multifeature squeeze-and-excitation-dilated residual network. Scientific Reports, 14(1), 11994.
37. Raj, S., & Singh, B. K. (2024). SpFusionNet: deep learning-driven brain image fusion with spatial frequency analysis. Multimedia Tools and Applications, 1-22.
38. Rathod, S., Mahajan, R., Goel, D. S., Dey, D. R., & Shirbhate, M. (2024). Innovative Fusion Approach for Lung Tumor Detection: Double Deep CNN Empowered by SVM Integration in CT Scan Images. Available at SSRN 4867723.
39. Yu, H., Wu, S., Zheng, R., Li, R., Wang, Q., Chen, C., ... & Wang, H. (2024). Generative Adversarial Registration Network for Multi-Contrast Liver MRI Registration and Added Value to Hepatocellular Carcinoma Segmentation: A Multicentre Study. IEEE Transactions on Emerging Topics in Computational Intelligence.
40. Crespi, L., Camnasio, S., Dei, D., Lambri, N., Mancosu, P., Scorsetti, M., & Loiacono, D. (2024). Leveraging Multimodal CycleGAN for the Generation of Anatomically Accurate Synthetic CT Scans from MRIs. arXiv preprint arXiv:2407.10888.
41. Choudhury, C., Goel, T., & Tanveer, M. (2024). A coupled-GAN architecture to fuse MRI and PET image features for multi-stage classification of Alzheimer's disease. Information Fusion, 109, 102415.
42. Luo, Y., Wu, G., Liu, Y., Liu, W., & Han, J. (2024). Towards High-Quality MRI Reconstruction With Anisotropic Diffusion-Assisted Generative Adversarial Networks And Its Multi-Modal Images Extension. IEEE Journal of Biomedical and Health Informatics.
43. Meng, X., Sun, K., Xu, J., He, X., & Shen, D. (2024). Multi-modal Modality-masked Diffusion Network for Brain MRI Synthesis with Random Modality Missing. IEEE Transactions on Medical Imaging.
44. Rahmani, M., Moghaddasi, H., Pour-Rashidi, A., Ahmadian, A., Najafzadeh, E., & Farnia, P. (2024). D2BGAN: Dual Discriminator Bayesian Generative Adversarial Network for Deformable MR–Ultrasound Registration Applied to Brain Shift Compensation. Diagnostics, 14(13), 1319.
45. Kim, B., Zhuang, Y., Mathai, T. S., & Summers, R. M. (2024). OTMorph: Unsupervised Multi-domain Abdominal Medical Image Registration Using Neural Optimal Transport. IEEE Transactions on Medical Imaging.
46. Zhang, M., Sun, L., Kong, Z., Zhu, W., Yi, Y., & Yan, F. (2024). Pyramid-attentive GAN for multimodal brain image complementation in Alzheimer's disease classification. Biomedical Signal Processing and Control, 89, 105652.
47. Mao, Q., Zhai, W., Lei, X., Wang, Z., & Liang, Y. (2024). CT and MRI image fusion via coupled feature-learning GAN. Electronics, 13(17), 3491.
48. Ren, Z., Li, J., Wu, L., Xue, X., Li, X., Yang, F., ... & Gao, X. (2024). Brain-driven facial image reconstruction via StyleGAN inversion with improved identity consistency. Pattern Recognition, 150, 110331.

49. Kim, J., & Park, H. (2024). Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 7604-7613).
50. Cao, X., Xue, P., Fan, J., Liu, D., Sun, K., Xue, Z., & Shen, D. (2024). Deep learning-based medical image registration. In Deep Learning for Medical Image Analysis (pp. 337-356). Academic Press.