DFMN: A Dual-feet Matching Network with Hybrid Transformer-based Feature Extractor for Unsupervised Deformable Medical Image Registration

Liwen Li[1], Xinrui Guo[1], Wentao Guo[2], Shunqi Yang[1], Fumin Guo[1]*

[1]MOE Key Laboratory for Biomedical Photonics, Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430074, China
[2]Department of Liberal arts and Sciences, University of Illinois Urbana-Champaign, Urbana IL 61801, United States
liliwen2002@hust.edu.cn, xinruig@hust.edu.cn, wentao5@illinois.edu, 840905872@qq.com, fguo@hust.edu.cn

Abstract

Deformable medical image registration is essential in medical image analyses. Recent transformer-based registration methods have achieved high registration accuracy. However, these methods often rely on patch embedding at the beginning of encoding, resulting in limited ability to capture detailed anatomical structural information in the images and explore local semantic relationships within individual patches. Here, we proposed a novel Dual-feet Encoder (DFEnc) to asynchronously model semantic information from moving and fixed images at various scales through two separate branches in three steps. For each step, features from adjacent resolution levels were processed by a Single Step Hybrid Extractor (SSHExt), which performed patch convolution to preserve local information, followed by several transformer blocks to capture global context. Dense connections were employed to enhance semantic awareness across adjacent feature resolution levels. Additionally, we introduced a Feature Fusion-based Decoder (FFDec) to progressively fuse features related to the fixed and moving images and to generate intermediate deformation fields at each stage, enabling accurate image alignment through stepwise warping and alignment refinement. Extensive ablation studies demonstrated the effectiveness of the proposed DFEnc, SSHExt, and FFDec. Compared to a state-of-the-art AutoFuse-Trans method, our approach yielded improvements in Dice of 1.14%, 1.77%, and 4.47% on the ACDC, OASIS, and Abdomen CT datasets, respectively, while maintaining relatively low computational cost. These results suggest the utility of the proposed approach for broad research and clinical applications.

Code — https://github.com/sfpl-code/DFMN

Introduction

Deformable medical image registration aims to align a moving image and a fixed image by establishing the spatial mapping relationships between the two. This technique plays a significant role in various medical image analysis tasks, including estimation and compensation of cardiac motion, construction of brain atlas, and evaluation of abdominal lesion development (Chen et al. 2025). Analytical methods (Rueckert et al. 1999; Periaswamy and Farid 2003; Beg et al. 2005; Ashburner 2007; Avants et al. 2008; Klein et al. 2010) formulate image registration tasks as optimization problems and typically employ iterative optimization to solve for the optimal parameters that are used to maximize the similarity between a fixed and a moving image subject to a regularization constraint. Although effective and widely used, these methods involve independent optimization for each image pair, resulting in significantly increased computational burden. In addition, these algorithms are generally formulated as high-order non-convex objective functions, which are difficult to optimize (Chen et al. 2025). These issues may have led to the limited clinical application of these methods.

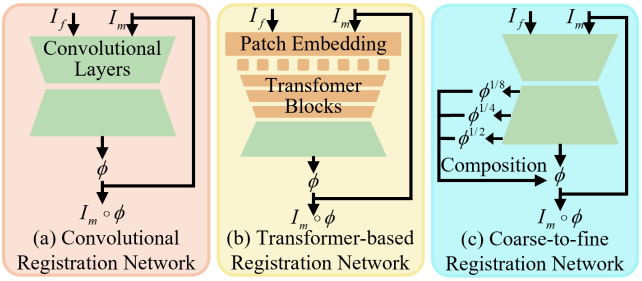

Figure 1: Schematic of different deep learning registration networks.

In the past few years, deep convolutional neural network (CNN) based methods have witnessed numerous successes in terms of registration accuracy and inference speed (Balakrishnan et al. 2019; De Vos et al. 2019; Dalca et al. 2019). Recently, the performance of CNN-based deformable registration methods was further enhanced with the application of attention mechanisms (Vaswani et al. 2017), which are intensively used in transformer architectures to enable long-range encoding (Dosovitskiy et al. 2021; Liu et al. 2021) that is beneficial for improved registration performance (Chen et al. 2022; Chen, Zheng, and Gee 2024). Existing transformer-based methods typically involve patch embedding at the beginning to transform an image into tokens (Dosovitskiy et al. 2021). However, artificially dividing a whole image into smaller patches impedes a comprehensive understanding of the anatomical structural information across the whole image, and the local semantic relevance

*Corresponding author.

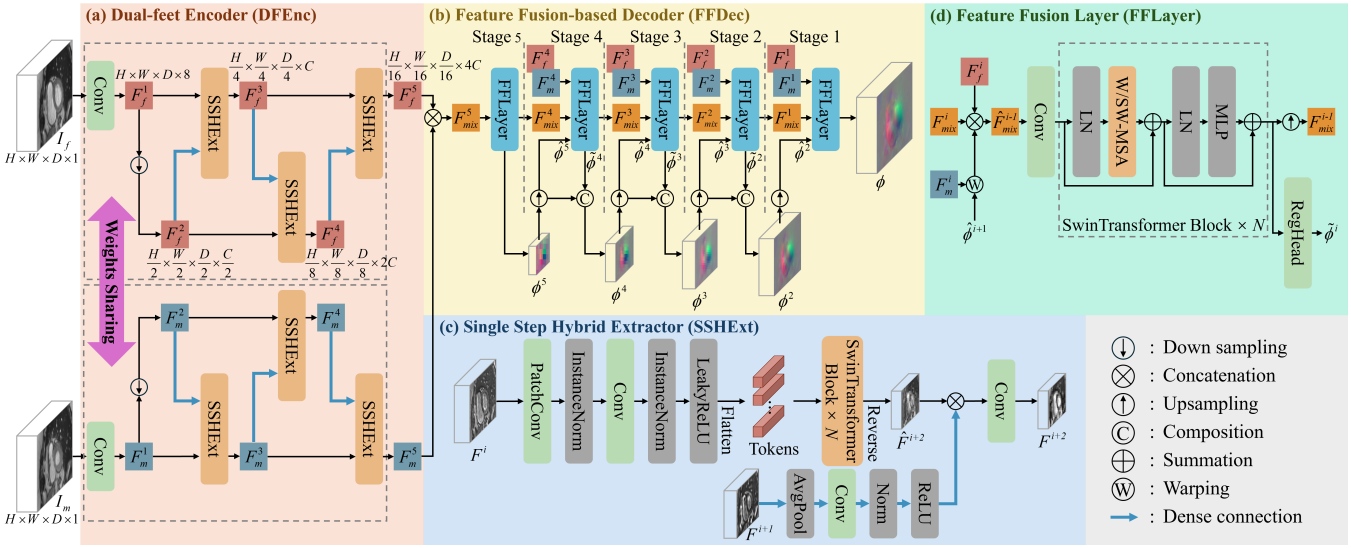

Figure 2: Overview of our Dual-feet Matching Network (DFMN) that encompasses (a) a Dual-feet Encoder, (b) a Feature Fusion Decoder, (c) a Single Step Hybrid Extractor, and (d) a Feature Fusion Layer. Input images I_f and I_m are processed by SSHExt in each layer of DFEnc with weights sharing for deep anatomical structural information extraction. The outputs of DFEnc are progressively fused in the FFLayer of FFDec for coarse-to-fine deformation estimation and composition.

within the individual patches has been largely ignored.

To alleviate this issue, we develop a novel registration network that encompasses a Dual-feet Encoder (DFEnc) and a Feature Fusion-based Decoder (FFDec). DFEnc enhances anatomical feature extraction through multi-step patch division and asynchronous dual-branch encoding, allowing the network to capture both fine-grained texture and high-level semantic information from multiple perspectives. By rethinking the fixed patch embedding strategy in conventional transformer-based methods, DFEnc adaptively models local semantic relevance and spatial dependencies across scales, leading to more comprehensive feature representation. The FFDec leverages these hierarchical representations to progressively estimate the deformation field in a coarse-to-fine manner for accurate image alignment. We note that the major contribution of this work is the development of the DFEnc module in recognition of the significant limitation of the fixed patch embedding in standard transformer framework. We summarize our main contributions as follows:

- We designed an asynchronous deep image feature encoder DFEnc to extract semantic information from each input image at multiple scales through two independent branches. This architecture enables the construction of a full image feature pyramid to enhance the encoding capability of the network.
- We proposed a novel hybrid image feature extractor SSHExt with dense connections to integrate and enhance the relevance of features from different branches for comprehensive image texture extraction. By integrating patch convolution with transformer blocks at each step, SSHExt effectively mitigates the issues of artificial patch division.
- We introduced FFDec to perform coarse-to-fine deformable image registration across multiple scales and progressively refine the alignment of the two images in a single pass, achieving significant improvements in registration accuracy.

Related Work

Deformable image registration aims to estimate spatial transformation parameters to minimize an objective (Chen et al. 2025) as follows:

$$\hat{\phi} = \arg\min_{\phi} S(I_f, I_m \circ \phi) + \lambda R(\phi) \tag{1}$$

where ϕ is the deformation field to warp the moving image I_m towards the fixed image I_f, S denotes the similarity metrics between I_f and the warped moving image $I_m \circ \phi$, and R represents a regularization term on the deformation field ϕ weighted by a hyperparameter λ. Common choices for similarity measurements S include mean squared error (MSE), normalized cross-correlation, and mutual information (Chen et al. 2022). The resulting deformation field ϕ is typically regularized by isotropic diffusion (Balakrishnan et al. 2019) and bending energy (Chen et al. 2022) on control points. While some early work performs supervised registration learning, these methods require synthesized deformation fields or those generated by traditional iterative methods (Dosovitskiy et al. 2015; Rohé et al. 2017; Yang et al. 2017). Recent emergence of spatial transformer networks (STN) (Jaderberg et al. 2015) enables developing unsupervised methods without the need for known deformation for algorithm training (De Vos et al. 2019; Balakrishnan et al. 2019; Dalca et al. 2019), providing immense flexibility in modeling various complex deformations and eliminating the reliance on ground truth deformation that is difficult to obtain (Chen et al. 2025).

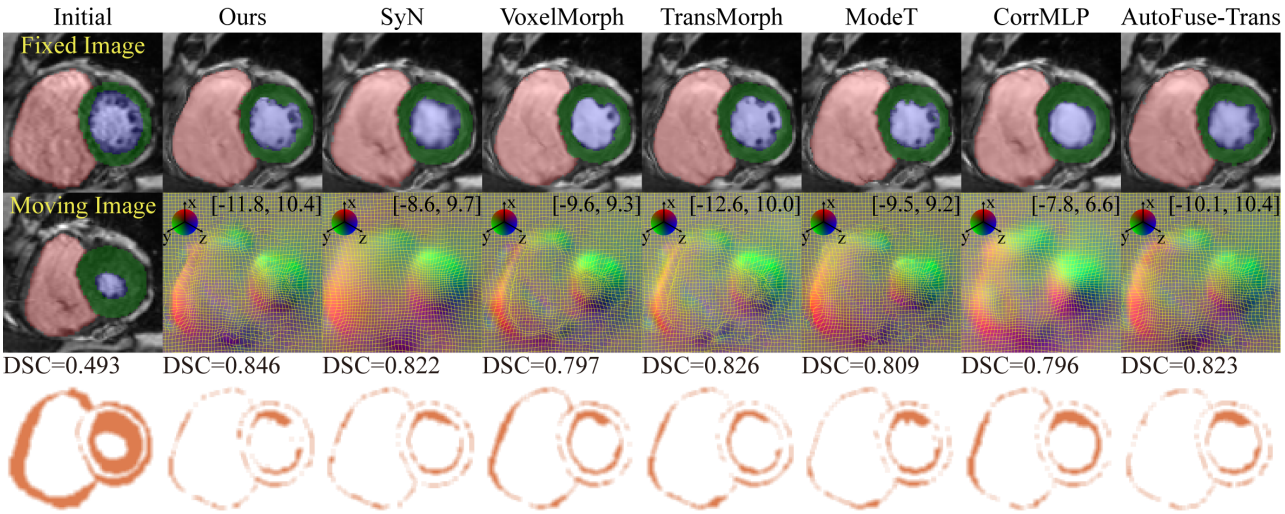

Figure 3: Representative intra-subject cardiac cine MRI registration for an ACDC subject. Warped moving images, the associated deformation field, and the differences between the fixed and warped moving image masks provided by various methods are shown in the 1^{st}, 2^{nd}, and 3^{rd} row, respectively. The x, y, and z dimensions of the deformation are mapped to the three RGB channels. $[m, n]$ indicates the minimal and maximal magnitude of the deformation field.

The demonstrated success of the U-Net network has led to the development and prosperity of CNN-based methods for effective medical image registration. Figure 1(a) shows the architecture of a typical CNN-based registration method, which involves an encoder-decoder structure consisting of multiple convolutional layers. For example, VoxelMorph (Balakrishnan et al. 2019) employed a five-layer U-Net with skip connections between the encoder and decoder to predict a dense deformation field ϕ. De Vos et al. (2019) utilized a CNN to predict the displacement field anchored on sparse control points with B-Spline regularization. Dalca et al. (2019) performed scaling-and-squaring on the derived stationary velocity field to ensure diffeomorphism of the deformation for topology preservation. Although effective and successful, these CNN-based methods are limited by the effective receptive fields (Luo et al. 2016), hindering the ability to capture long-range spatial correspondence and leading to sub-optimal local and global image alignment.

Leveraging the capability of capturing long-range dependencies, transformer-based registration networks provide a way to improve image alignment by capturing global and local spatial relationships (Vaswani et al. 2017). Figure 1(b) illustrates the architecture of transformer-based methods, which use patch embedding to convert whole images into individual tokens at the beginning of the network. ViT-V-Net (Chen et al. 2021) combined Vision Transformer (Dosovitskiy et al. 2021) and convolutions in the encoding process to predict a dense displacement field. Subsequently, TransMorph (Chen et al. 2022) was developed by integrating the SwinTransformer (Liu et al. 2021) in the encoder to extract image features. The significantly improved performance of TransMorph *vs* prior CNN-based registration methods suggests the effectiveness of transformer models for deformable image registration. XMorpher (Shi et al. 2022) further constructed a full transformer architecture, whereby fixed and moving images were processed in dual channels in parallel, and information across these two channels was exchanged through a cross-attention mechanism. TransMatch (Chen, Zheng, and Gee 2024) leveraged a cross-attention module integrated with a foundational convolutional backbone to enhance feature matching. However, these methods typically divide images into patches via patch embedding at the beginning of the network, limiting the ability to fully capture anatomical structures in medical images. Additionally, simplistic application of attention mechanisms often results in an overlook of the local semantic continuity within individual patches.

While the final full-resolution deformation field can be generated through a single pass prediction, e.g., at the end of the decoder, previous work also proposed to estimate the intermediate deformation at multiple scales in the decoder pyramid and generate the final deformation field through progressive warping and/or composition, as shown in Figure 1(c). NICE-Net (Meng et al. 2022) proposed a non-iterative coarse-to-fine deformable registration network by predicting the deformation field at multiple levels of the decoder. NICE-Trans (Meng et al. 2023) further employed SwinTransformer to improve long-range encoding capability and an additional affine transformation head to perform joint affine and deformable registration. PIViT (Ma et al. 2023) proposed a lightweight, pyramid, and iterative framework based on SwinTransformer for the application of the large deformation image registration. ModeT (Wang, Ni, and Wang 2023) introduced a motion decomposition strategy to explicitly model multiple deformation modes by exploiting the intrinsic capability of the transformer structure for deformation estimation. CorrMLP (Meng et al. 2024) exploited the strengths of correlation and MLP blocks in the field of registration by integrating them with a coarse-to-fine registration backbone.

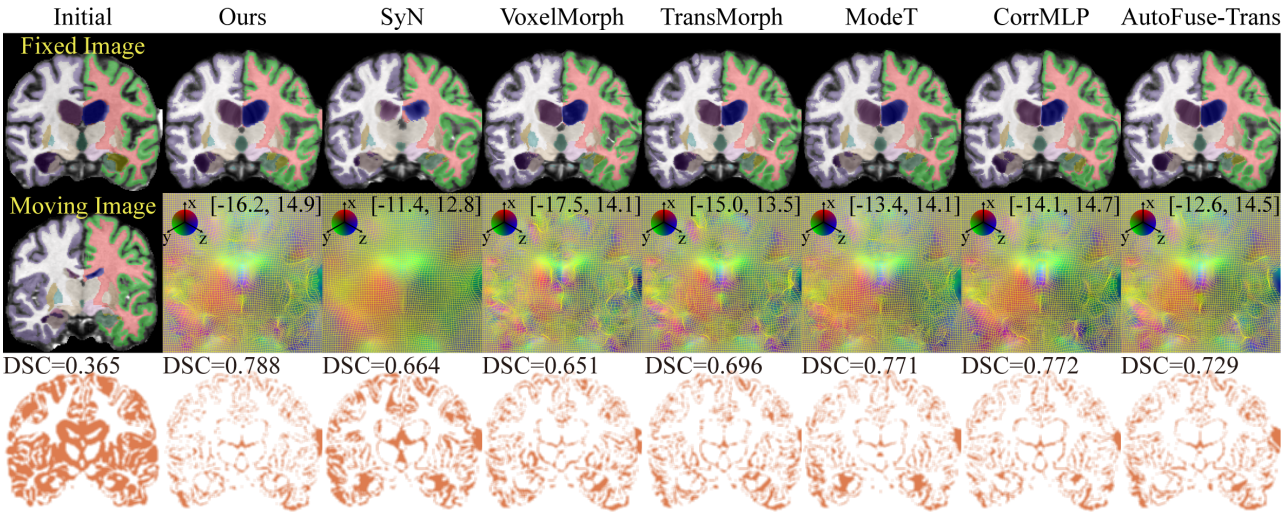

Figure 4: Representative registration of two brain MRI datasets. Results are shown in the same way as Figure 3.

Proposed Method

Overview

As shown in Figure 2, the proposed Dual-feet Matching Network (DFMN) consists of a DFEnc and an FFDec. DFEnc involves a dual-branch architecture to extract asynchronous deep semantic information from the input images across multiple scales. A fixed image I_f and a moving image I_m are processed in parallel by two DFEnc with shared weights, producing multi-resolution feature pyramids for each image. Within each encoder step, a Single Step Hybrid Extractor (SSHExt), which combines patch convolution with multiple transformer blocks, merges the features at adjacent resolution levels from the two branches to capture detailed anatomical structural information. Subsequently, FFDec integrates the multi-scale features extracted from the image pairs and progressively estimates the deformation field in a coarse-to-fine manner. At each stage of FFDec, a Feature Fusion Layer (FFLayer) fuses the corresponding features of I_f and I_m in the encoder to compute the deformation field. The final full-resolution deformation field ϕ is obtained by composing these incremental deformations from the lowest to the full image resolution level to warp I_m towards I_f.

Dual-feet Encoder

As shown in Figure 2(a), DFEnc constructs an asynchronous downsampling architecture composed of two independent branches, each of which extracts a subset of features from I_f and I_m. The outputs from the two branches are combined to form a comprehensive multi-resolution feature pyramid. This dual-branch design facilitates perceiving semantic information at multiple scales, thereby enhancing learning.

The fixed image $I_f \in \mathbb{R}^{H\times W\times D\times 1}$ is first passed through a convolutional layer (kernel size = 3, stride = 1, padding = 1) to generate feature $F_f^1 \in \mathbb{R}^{H\times W\times D\times 8}$, which is entered into two separate branches. In the first branch, F_f^1 goes through two consecutive SSHExt layers, each of which reduces the spatial resolution to $\frac{1}{4}$ and increases the number of channels to 4x of the input dimensions. This branch produces intermediate features F_f^3 (size = $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times C$) and F_f^5 (size = $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 4C$), where $C = 32$. Meanwhile, the second branch downsamples F_f^1 via average pooling (stride = 2) followed by two convolutional layers (kernel size = 3), yielding F_f^2 (size = $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times \frac{C}{2}$). Similarly, F_f^2 is processed by SSHExt to generate F_f^4 (size = $\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 2C$). Within SSHExt, these features are processed by a hybrid convolution and transformer architecture to further extract local and long-range image texture. The resulting feature maps $\{F_f^1, \ldots, F_f^5\}$ constitute a multi-resolution feature pyramid, which enables the model to comprehensively capture the semantic information at multiple scales. Unfortunately, the relevance of features at adjacent resolution levels remained sparse, and accordingly, we propose the dense connection to enhance the exchange of information between the two branches for robust modeling of the anatomical structural information for each image. The same weight-sharing DFEnc is used to encode the moving image I_m, producing the feature pyramid $\{F_m^1, \ldots, F_m^5\}$.

Single Step Hybrid Extractor

For each image, the extracted features with adjacent resolutions in two branches of DFEnc, e.g., F_m^i and F_m^{i+1}, $i = \{1, 2, 3\}$, were fused using the proposed SSHExt to generate the intermediate deformation field in the current stage. Inspired by the patch embedding operation in ViT (Dosovitskiy et al. 2021), we adopt a patch convolution (PatchConv) to convert the input feature F_m^i into two-dimensional tokens prior to applying attention mechanisms. In particular, a large convolution kernel is employed to generate patches with reduced spatial resolution and increased channel dimensions, differing from the standard patch-merging operation used in traditional ViT. The input feature is downsampled by a factor n, which is empirically set to 4 to balance detail retention ability and computational efficiency. A larger n would result

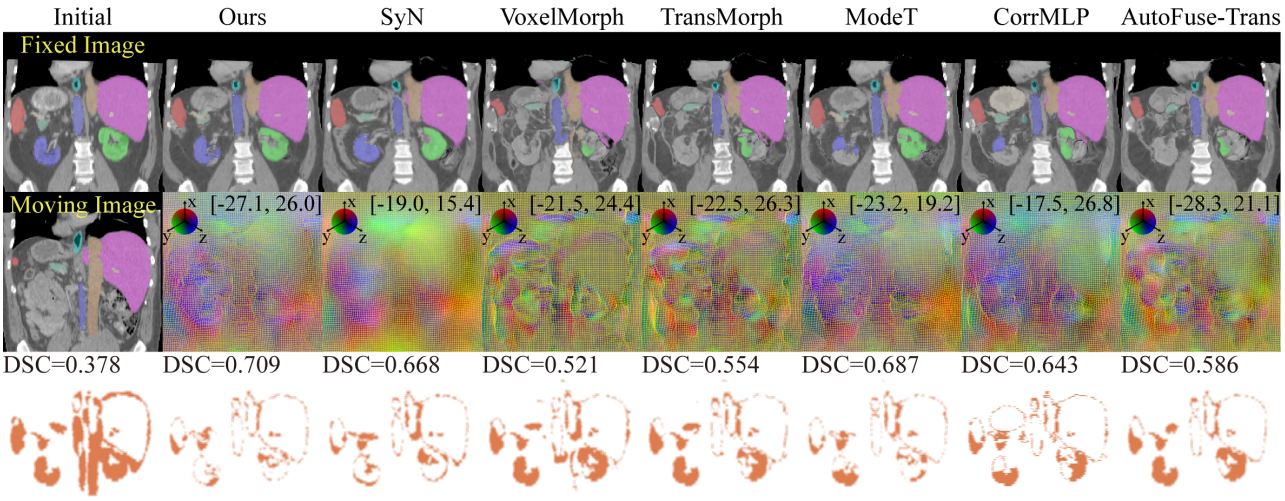

Figure 5: Example results Abdomen CT registrations. Results are shown in the same way as Figure 3.

in sparse features and loss of key textures, while a smaller value would limit patch-level abstraction. SSHExt generates a feature map with $\frac{1}{4}$ and 4x the sizes in the spatial and channel dimensions, respectively.

Without loss of generality, we denote $H_0 \times W_0 \times D_0 \times C_0$ as the size of F^i that is entered into SSHExt. PatchConv with stride $s = 4$ is used to ensure the sizes of the outputs are reduced to $\frac{1}{4}$ of F^i. The kernel size is set to $k = 2s - 1$ such that the convolution adequately covers the entire patch, and padding $p = \lfloor \frac{k}{2} \rfloor$ in each dimension is performed to match the sizes of features in the corresponding layers in the encoder and decoder. The output of PatchConv is further convolved to obtain intermediate feature sized $\frac{H_0}{4} \times \frac{W_0}{4} \times \frac{D_0}{4} \times 4C_0$, which is subsequently flattened to generate tokens as input to N SwinTransformer blocks (dotted box in Figure 2(b)) to conduct window-based self-attention for efficient global and contextual feature modeling. The SwinTransformer block consists of layer normalization (LN), window-based multi-head self-attention (W-MSA), shifted window-based multi-head self-attention (SW-MSA), and multilayer perceptron (MLP). This process is formulated as follows:

$$\begin{aligned} \hat{z}_i &= W\text{-}MSA(LN(z_{i-1})) + z_{i-1} \\ z_i &= MLP(LN(\hat{z}_i)) + \hat{z}_i \\ \hat{z}_{i+1} &= SW\text{-}MSA(LN(z_i)) + z_i \\ z_{i+1} &= MLP(LN(\hat{z}_{i+1})) + \hat{z}_{i+1} \end{aligned} \tag{2}$$

where z_{i-1} indicates the input to the SwinTransformer module, z_i denotes the intermediate output produced by the W-MSA and MLP modules, and z_{i+1} is the final output of the Swin-Transformer block. The MSA operation can be formulated as follows:

$$\begin{aligned} attn_i &= Softmax(\frac{Q_i K_i^T}{\sqrt{d_h}} + B_i)V_i \\ MSA &= Concat_i^h(attn_i)W^o \end{aligned} \tag{3}$$

where Q_i, K_i, and V_i are the query, key, and value generated from the tokens using the i^{th} linear projection head in the MSA procedure, $i = 1, \ldots, h$; B_i represents the relative position bias, and $attn_i$ represents outputs of the i^{th} attention head. $d_h = \frac{d}{h}$, in which d denotes the number of dimensions of the tokens and h represents the number of the attention heads. The results of all the attention heads are concatenated ($Concat_i^h()$) and linearly projected by a projection matrix W^o. The processed tokens are reversed to the shape of the original features to obtain $\hat{F}^{i+2}$.

As shown in Figure 2(a), the semantic information transfers only within each branch. For instance, F_f^{i+2} receives information solely from F_f^i, $i = \{1, 3\}$ for the first branch and $i = 2$ for the second branch, and the differences in the sizes of F_f^{i+2} and F_f^i are 4x. This introduces potential sampling error during encoding. To alleviate this issue, dense connection is leveraged to enhance the generation of F_f^{i+2} from F_f^i. As shown in Figure 2(c), F^{i+1} from the other branch is processed by average pooling (AvgPool), convolution (Conv), a normalization (Norm), and a rectified linear unit (ReLU) in sequence. The output is concatenated with $\hat{F}^{i+2}$ followed by convolution to obtain F^{i+2} with a matrix size of $\frac{H_0}{4} \times \frac{W_0}{4} \times \frac{D_0}{4} \times 4C_0$.

Feature Fusion Decoder

We adopt the coarse-to-fine scheme and design an FFDec consisting of multiple feature fusion layers to estimate the deformation field in a progressive manner. As shown in Figure 2(b), FFDec consists of five stages. For the i^{th} stage, $i = \{4, 3, 2, 1\}$, the moving image feature F_m^i is warped by STN using the upsampled deformation $\hat{\phi}^{i+1}$ from the previous stage, as illustrated in Figure 2(d). The warped moving image feature $\hat{F}_m^i$ is then concatenated with the fixed image feature F_f^i and the mixed feature F_{mix}^i from the previous stage to generate the updated mixed feature $\hat{F}_{mix}^{i-1}$ as follows:

$$\begin{aligned} \hat{F}_m^i &= STN(F_m^i, \hat{\phi}^{i+1}) \\ \hat{F}_{mix}^{i-1} &= Concat(F_f^i, F_{mix}^i, \hat{F}_m^i) \end{aligned} \tag{4}$$

Method	ACDC		OASIS		AbdomenCT	
	DSC (↑)	Jac (↓)	DSC (↑)	Jac (↓)	DSC (↑)	Jac (↓)
initial	59.10±16.98	-	57.22±7.17	-	38.63±5.09	-
SyN	78.78±9.89*	**0.02±0.03**	76.60±3.35*	**0.02±0.02**	51.97±10.22*	**0.34±0.28**
VoxelMorph	78.99±9.96*	0.27±0.26	78.20±3.55*	0.82±0.15	50.75±5.01*	7.41±1.99
TransMorph	80.26±9.61*	0.44±0.41	79.83±2.84*	0.71±0.11	54.16±5.48*	7.35±1.90
TransMatch	79.64±10.10*	0.35±0.34	79.73±2.81*	0.73±0.10	54.85±5.33*	7.02±1.78
NICE-Trans	78.70±10.60*	0.43±0.29	80.85±2.51*	0.75±0.10	52.87±9.01*	6.11±1.61
ModeT	80.45±9.28*	0.26±0.16	80.56±2.28*	0.46±0.08	59.16±6.07*	3.62±1.09
RDP	80.53±9.72*	0.13±0.07	80.96±2.42*	0.48±0.07	57.40±6.91*	3.38±1.16
CorrMLP	79.93±10.01*	0.02±0.03	80.48±2.33*	0.57±0.09	59.02±6.54*	4.23±1.40
IIRP-Net	80.64±10.08*	0.46±0.37	79.84±2.84*	1.17±0.22	52.40±7.05*	6.70±1.83
AutoFuse-Trans	80.82±9.23*	0.18±0.16	80.23±2.44*	0.46±0.07	57.28±5.55*	3.60±1.08
Ours	**81.96±8.78**	0.26±0.19	**82.00±2.21**	0.54±0.09	**61.75±5.46**	3.89±1.27

Table 1: DSC and Jac provided by various registration methods on the ACDC, OASIS, and Abdomen CT datasets. Bold indicates the higher performance. *: p<0.05 for comparison to our approach.

A convolutional layer is leveraged to reduce the sizes of $\hat{F}_{mix}^{i-1}$ to minimize redundant information, followed by N SwinTransformer blocks to improve the capability of estimating the mapping relationship between image features. The outputs of the SwinTransformer blocks are used to estimate the deformation $\widetilde{\phi}^i$ at the current stage using a registration head (RegHead) following previous work (Balakrishnan et al. 2019). In addition, the same outputs are upsampled to generate the mixed feature F_{mix}^{i-1} for the current stage using transposed convolutions. Of note, for the 5^{th} stage, F_{mix}^5 is obtained by directly concatenating F_m^5 and F_m^5.

The full-resolution deformation field ϕ is generated by iteratively compositing the intermediate deformation field for all the stages:

$$\phi^i = STN(\hat{\phi}^{i+1}, \tilde{\phi}^i) + \tilde{\phi}^i, \quad i = 4, 3, 2, 1 \tag{5}$$

where $STN(\hat{\phi}^{i+1}, \tilde{\phi}^i)$ indicates warping the upsampled deformation from the previous stage $\hat{\phi}^{i+1}$ using the estimated deformation field for the current stage $\tilde{\phi}^i$ by STN. ϕ^1 is used as the final full-resolution deformation ϕ, i.e., $\phi = \phi^1$.

Loss Functions for Unsupervised Learning

The proposed registration algorithm is optimized in an unsupervised manner. The loss function comprises a similarity term based on MSE and a regularization term based on L2 norm of the spatial gradients of the deformation field:

$$\mathcal{L}(I_f, I_m, \phi) = \|I_m \circ \phi - I_f\|_2^2 + \lambda \|\nabla\phi\|_2^2 \tag{6}$$

where λ is the weight of the regularization term.

Experiments

Datasets and Implementation Details

Datasets: We investigated the ACDC (Bernard et al. 2018), OASIS (Marcus et al. 2007; Hoopes et al. 2021), and Abdomen CT (Xu et al. 2016) datasets. The ACDC dataset comprises cardiac cine MRI of the whole heart from 150 patients. Expert manual segmentations of the right ventricle (RV), myocardium (Myo), and left ventricle (LV) were provided for the end-systolic (ES) and end-diastolic (ED) frames only. Following the same procedure in (Meng et al. 2025), all images were resampled to a voxel size of $1.5 \times 1.5 \times 3.15$ mm^3 and cropped to a matrix size of $128 \times 128 \times 32$ for network input. Random 90, 10, and 50 subjects were used for training, validation, and testing, respectively. During training, two random frames were paired for each subject, resulting in a total of 2,697 image pairs. For validation and testing, the ES frame was aligned to the ED frame for intra-subject registration.

Following previous work (Hoopes et al. 2021), the same 414 T1-weighted brain MRI images from different subjects in the OASIS dataset were used for inter-subject registration. The dataset included 35 anatomical structural labels for each subject. Bias correction, skull stripping, rigid alignment, resampling, and cropping were performed using FreeSurfer to generate pre-processed images with a uniform matrix size of $160 \times 192 \times 224$ and a voxel spacing of $1 \times 1 \times 1$ mm^3. The dataset was split into 294, 40, and 80 subjects for training, validation, and testing, respectively. During training, two subjects were randomly sampled from the 294 training cases, whereas validation and testing were performed on subject pairs in a fixed order.

The Abdomen CT dataset contains 30 abdominal CT scans, and a total of 13 anatomical structures were annotated. Due to missing annotations in some subjects, only the 9 labels available for all subjects were used for evaluation, consistent with (Zheng et al. 2024). All scans were resampled, padded, and cropped to a uniform size of $192 \times 160 \times 224$ with a voxel spacing of $2 \times 2 \times 2$ mm^3. The dataset was divided into 23 cases for training and 7 for testing. Random two scans from the two subsets were used for both training (506 pairs) and testing (42 pairs).

Implementation Details: Our approach was implemented using Pytorch 2.1.2 on NVIDIA GeForce RTX 4090/4090D GPUs (24 GB memory) with the following parameters: optimizer = Adam, learning rate = 1×10^{-4}, number of epochs = 200, batch size = 8, 1 and 1 for ACDC, OASIS, and Abdomen CT, respectively. The weight λ in Eq. (6) was set to 0.01.

Method	Para.(e6)	FLOPs(e9)	TPI(s)
SyN	-	-	20.12
VoxelMorph	0.40	17.55	0.014
TransMorph	46.70	50.14	0.049
TransMatch	86.84	37.09	0.127
NICE-Trans	5.71	20.71	0.058
ModeT	1.03	6.15	0.083
RDP	8.92	308.00	0.164
CorrMLP	4.19	71.08	0.082
IIRP-Net	0.42	21.62	0.024
AutoFuse-Trans	8.75	126.04	0.097
Ours	9.81	25.61	0.057

Table 2: Computational complexity of various algorithms.

Comparison with Existing Methods

We compared the proposed DFMN with SyN (Avants et al. 2008), a CNN-based registration model VoxelMorph (Balakrishnan et al. 2019), and transformer-based registration methods TransMorph (Chen et al. 2022), TransMatch (Chen, Zheng, and Gee 2024) and AutoFuse-Trans (Meng et al. 2025), as well as five popular coarse-to-fine registration approaches NiCE-Trans (Meng et al. 2023), ModeT (Wang, Ni, and Wang 2023), RDP (Wang, Ni, and Wang 2024), CorrMLP (Meng et al. 2024) and IIRP-Net (Ma et al. 2024). All the deep learning-based methods were trained, validated, and tested on the same subjects as ours. Registration accuracy was evaluated by the mean of Dice Similarity Coefficient (DSC, %) between the fixed and warped moving image label for all anatomical structures. The percentage of negative Jacobian determinants (Jac, %) was calculated to assess the smoothness and invertibility of the deformation. Computational complexity was determined using FLOPs, inference time per image (TPI), and the number of network parameters (Para.). A paired t-test was conducted to compare the DSC obtained by our approach and the other methods.

Figure 3 shows that DFMN produced more accurate alignment of ES and ED frames for an ACDC subject. The contours of the RV, Myo, and LV in the warped image closely matched those in the fixed image, exhibiting cleaner error maps and higher DSC scores compared to other methods. Figure 4 and 5 show representative registration results for brain MRI (OASIS) and abdominal CT scans (Abdomen CT). For brain MRI, DFMN achieved better alignment of the left and right lateral ventricles, although some small structures, like the hippocampus, remained challenging to align. For abdominal CT, where deformations are more complex, our approach still produced promising results, especially in the left and right kidneys (Figure 5).

Table 1 reveals that DFMN yielded significantly greater DSC than all the comparative methods ($p<0.05$), and these were 1.14% (on ACDC), 1.77% (on OASIS), and 4.47% (on Abdomen CT) higher than a recently developed state-of-the-art transformer-based AutoFuse-Trans method. In addition, DFMN generated relatively smooth deformation as evidenced by the comparable Jacobian determinants to other learning-based methods. Table 2 shows that our approach required fewer network parameters and FLOPs compared to other transformer-based methods, while maintaining a fast inference time of 57 ms to align two frames of the ACDC dataset, demonstrating its practical applicability.

Model	Mean	RV	Myo	LV
CNN baseline	78.6±9.8	78.6±8.0	71.6±8.4	85.6±7.6
Trans. baseline	80.2±9.5	79.9±8.2	73.4±7.7	87.1±7.2
w/o SSHExt&DC	80.2±9.3	79.4±7.7	73.9±8.1	87.1±6.9
w/o SSHExt	80.9±9.2	80.4±7.9	74.7±7.6	87.7±7.0
w/o DC	80.9±9.4	80.5±7.7	74.0±8.1	88.1±6.1
Ours	81.8±8.7	80.9±7.6	75.7±7.1	88.8±5.8

Table 3: DSC for ablation studies of DFEnc on ACDC.

Model	Mean	RV	Myo	LV
w/o FFDec	75.9±11.8	77.8±8.0	67.1±11.2	82.9±9.9
w/o C2FDef	79.8±10.0	79.5±8.0	72.6±9.0	87.3±6.9
w/o SwinBlock	79.1±10.1	79.0±7.9	71.9±8.9	86.2±8.0
w/o FFLayer	76.1±11.5	77.2±8.4	67.9±11.0	83.3±9.4
Ours	81.8±8.7	80.9±7.6	75.7±7.1	88.8±5.8

Table 4: DSC for ablation studies of FFDec on ACDC.

Ablation Studies

We conducted ablation studies on the ACDC dataset to assess the effectiveness of DFEnc, FFDec, and their respective sub-modules. Table 3 shows that replacing DFEnc with a convolutional encoder similar to that in VoxelMorph (CNN baseline) resulted in a substantial decrease in DSC. Our approach achieved 1.6% higher DSC than uniform transformer variant (Trans. baseline) using a standard patch embedding process, clearly demonstrating the effectiveness of DFEnc for multi-step patch division. Additionally, we evaluated the effects of removing the dense connection (w/o DC), replacing SSHExt with regular convolutions (w/o SSHExt), and retaining only the dual-branch architecture (w/o DC&SSHExt) from DFMN, and observed exclusively degraded registration accuracy for each case. For FFDec, a convolutional architecture (w/o FFDec) yielded significantly lower DSC, as illustrated in Table 4. Single pass estimation using convolutions (w/o C2FDef), removal of SwinTransformer blocks (w/o SwinBlock), and replacement of FFLayer with convolutions (w/o FFLayer) all led to suboptimal DSC compared to the full FFDec implementation.

Conclusions

We propose an unsupervised method that performs asynchronous feature encoding and fusion via a hybrid convolution–transformer module. By employing patch convolution to alleviate the limitations of artificial patch division, our approach enables more comprehensive extraction of anatomical structural information, achieving improved deformable medical image registration performance. For three cardiac, brain, and abdominal MRI and CT datasets, the proposed approach demonstrated the effectiveness of the entire algorithm framework and the individual modules, outperforming several state-of-the-art registration methods.

References

Ashburner, J. 2007. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1): 95–113.

Avants, B. B.; Epstein, C. L.; Grossman, M.; and Gee, J. C. 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1): 26–41.

Balakrishnan, G.; Zhao, A.; Sabuncu, M. R.; Guttag, J.; and Dalca, A. V. 2019. VoxelMorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8): 1788–1800.

Beg, M. F.; Miller, M. I.; Trouvé, A.; and Younes, L. 2005. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision*, 61(2): 139–157.

Bernard, O.; Lalande, A.; Zotti, C.; Cervenansky, F.; Yang, X.; Heng, P.-A.; Cetin, I.; Lekadir, K.; Camara, O.; Gonzalez Ballester, M. A.; et al. 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11): 2514–2525.

Chen, J.; Frey, E. C.; He, Y.; Segars, W. P.; Li, Y.; and Du, Y. 2022. TransMorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82: 102615.

Chen, J.; He, Y.; Frey, E. C.; Li, Y.; and Du, Y. 2021. ViT-V-Net: Vision transformer for unsupervised volumetric medical image registration. *arXiv Preprint arXiv:2104.06468*.

Chen, J.; Liu, Y.; Wei, S.; Bian, Z.; Subramanian, S.; Carass, A.; Prince, J. L.; and Du, Y. 2025. A survey on deep learning in medical image registration: New technologies, uncertainty, evaluation metrics, and beyond. *Medical Image Analysis*, 100: 103385.

Chen, Z.; Zheng, Y.; and Gee, J. C. 2024. TransMatch: A Transformer-Based Multilevel Dual-Stream Feature Matching Network for Unsupervised Deformable Image Registration. *IEEE Transactions on Medical Imaging*, 43(1): 15–27.

Dalca, A. V.; Balakrishnan, G.; Guttag, J.; and Sabuncu, M. R. 2019. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical Image Analysis*, 57: 226–236.

De Vos, B. D.; Berendsen, F. F.; Viergever, M. A.; Sokooti, H.; Staring, M.; and Isgum, I. 2019. A deep learning framework for unsupervised affine and deformable image registration. *Medical Image Analysis*, 52: 128–143.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*.

Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Van Der Smagt, P.; Cremers, D.; and Brox, T. 2015. FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2758–2766.

Hoopes, A.; Hoffmann, M.; Fischl, B.; Guttag, J.; and Dalca, A. V. 2021. HyperMorph: Amortized hyperparameter learning for image registration. In *International Conference on Information Processing in Medical Imaging (IPMI)*, 3–17.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; and Kavukcuoglu, K. 2015. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2017–2025.

Klein, S.; Staring, M.; Murphy, K.; Viergever, M. A.; and Pluim, J. P. W. 2010. Elastix: a toolbox for intensity-based medical image registration. *IEEE Transactions on Medical Imaging*, 29(1): 196–205.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Luo, W.; Li, Y.; Urtasun, R.; and Zemel, R. 2016. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 4905–4913.

Ma, T.; Dai, X.; Zhang, S.; and Wen, Y. 2023. PiViT: Large deformation image registration with pyramid-iterative vision transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 602–612.

Ma, T.; Zhang, S.; Li, J.; and Wen, Y. 2024. IIRP-Net: Iterative inference residual pyramid network for enhanced image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11546–11555.

Marcus, D. S.; Wang, T. H.; Parker, J.; Csernansky, J. G.; Morris, J. C.; and Buckner, R. L. 2007. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9): 1498–1507.

Meng, M.; Bi, L.; Feng, D.; and Kim, J. 2022. Non-iterative coarse-to-fine registration based on single-pass deep cumulative learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 88–97.

Meng, M.; Bi, L.; Fulham, M.; Feng, D.; and Kim, J. 2023. Non-iterative coarse-to-fine transformer networks for joint affine and deformable image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 750–760.

Meng, M.; Feng, D.; Bi, L.; and Kim, J. 2024. Correlation-aware coarse-to-fine MLPs for deformable medical image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9645–9654.

Meng, M.; Fulham, M.; Feng, D.; Bi, L.; and Kim, J. 2025. AutoFuse: Automatic fusion networks for deformable medical image registration. *Pattern Recognition*, 161: 111338.

Periaswamy, S.; and Farid, H. 2003. Elastic registration in the presence of intensity variations. *IEEE Transactions on Medical Imaging*, 22(7): 865–874.

Rohé, M.-M.; Datar, M.; Heimann, T.; Sermesant, M.; and Pennec, X. 2017. SVF-Net: Learning deformable image registration using shape matching. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 266–274.

Rueckert, D.; Sonoda, L. I.; Hayes, C.; Hill, D. L. G.; Leach, M. O.; and Hawkes, D. J. 1999. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Transactions on Medical Imaging*, 18(8): 712–721.

Shi, J.; He, Y.; Kong, Y.; Coatrieux, J.-L.; Shu, H.; Yang, G.; and Li, S. 2022. XMorpher: Full transformer for deformable medical image registration via cross attention. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 217–226.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 6000–6010.

Wang, H.; Ni, D.; and Wang, Y. 2023. ModeT: Learning deformable image registration via motion decomposition transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 740–749.

Wang, H.; Ni, D.; and Wang, Y. 2024. Recursive deformable pyramid network for unsupervised medical image registration. *IEEE Transactions on Medical Imaging*, 43(6): 2229–2240.

Xu, Z.; Lee, C. P.; Heinrich, M. P.; Modat, M.; Rueckert, D.; Ourselin, S.; Abramson, R. G.; and Landman, B. A. 2016. Evaluation of six registration methods for the human abdomen on clinically acquired CT. *IEEE Transactions on Biomedical Engineering*, 63(8): 1563–1572.

Yang, X.; Kwitt, R.; Styner, M.; and Niethammer, M. 2017. Quicksilver: Fast predictive image registration–a deep learning approach. *NeuroImage*, 158: 378–396.

Zheng, J.-Q.; Wang, Z.; Huang, B.; Lim, N. H.; and Papież, B. W. 2024. Residual aligner-based network (RAN): Motion-separable structure for coarse-to-fine discontinuous deformable registration. *Medical Image Analysis*, 91: 103038.