

FractMorph: A Fractional Fourier-Based Multi-Domain Transformer for Deformable Image Registration

Shayan Kebriti^a, Shahabedin Nabavi^{a,*} and Ali Gooya^{b,c}

^a*Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran*

^b*School of Computing Science, University of Glasgow, Glasgow, UK*

^c*Alan Turing Institute, London, UK*

Abstract

Deformable image registration (DIR) is a crucial and challenging technique for aligning anatomical structures in medical images and is widely applied in diverse clinical applications. However, existing approaches often struggle to capture fine-grained local deformations and large-scale global deformations simultaneously within a unified framework. We present FractMorph, a novel 3D dual-parallel transformer-based architecture that enhances cross-image feature matching through multi-domain fractional Fourier transform (FrFT) branches. Each Fractional Cross-Attention (FCA) block applies parallel FrFTs at fractional angles of 0° , 45° , 90° , along with a log-magnitude branch, to effectively extract local, semi-global, and global features at the same time. These features are fused via cross-attention between the fixed and moving image streams. A lightweight U-Net style network then predicts a dense deformation field from the transformer-enriched features. On the intra-patient ACDC cardiac MRI dataset, FractMorph achieves state-of-the-art performance with an overall Dice Similarity Coefficient (DSC) of 86.45%, an average per-structure DSC of 75.15%, and a 95th-percentile Hausdorff distance (HD95) of 1.54 mm on our data split. FractMorph-Light, a lightweight variant of our model with only 29.6M parameters, preserves high accuracy while halving model complexity. Furthermore, we demonstrate the generality of our approach with solid performance on a cerebral atlas-to-patient dataset. Our results demonstrate that multi-domain spectral–spatial attention in transformers can robustly and efficiently model complex non-rigid deformations in medical images using a single end-to-end network, without the need for scenario-specific tuning or hierarchical multi-scale networks. The source code is available at <https://github.com/shayankebriti/FractMorph>.

Keywords: Deep learning ; Vision Transformer ; Image registration ; Medical image analysis ; Fractional Fourier transform

1. Introduction

Image registration is the process of estimating spatial transformations to align the corresponding anatomical structures of multiple images (Zitová and Flusser, 2003). This technique is widely used in clinical applications, since aligning images taken at various times, orientations, or modalities is needed for accurate medical analysis and diagnosis. Registration methods are generally categorized according to the type of transformation they perform. Rigid registration is used for translation and rotation transformations, whereas affine registration extends this also to include scaling and shearing (Pluim and Fitzpatrick, 2003). A single 2D matrix is sufficient to represent these transformations. Although both rigid and affine approaches have been shown to be effective in practice, they become less accurate when the shape of anatomical structures changes significantly between images (Zou et al., 2022). Deformable image registration (DIR) overcomes this limitation by facilitating non-rigid deformations within the images (Pluim and Fitzpatrick, 2003). This flexibility allows DIR to model detailed changes in anatomical structures accurately. A deformable transformation is commonly represented as a dense deformation field,

formulated as a 3D matrix in 2D image registration tasks and a 4D matrix for 3D cases (Zou et al., 2022). Correctly estimating these complex, high-dimensional matrices is a challenging task because of the complexity of capturing large and irregular shape changes in anatomical structures, especially when the deformation is massive or highly irregular.

Deformable registration has shown significant value in a wide range of clinical uses, including diagnosis, surgical planning, and longitudinal monitoring. It forms the foundation of numerous critical tasks, such as identification of tumor progression, organ tracking, surgical navigation, and multi-organ mapping (Ramadan et al., 2024). It is especially valuable for integrating information from varied imaging modalities, such as combining the soft tissue contrast of MRI with the bone detail of CT, enabling a more comprehensive anatomical and functional understanding that enhances diagnostic accuracy and treatment planning (Huang et al., 2020). DIR also plays a crucial role in analyzing anatomical structures by allowing detailed motion tracking of tissues. In cardiac imaging, accurate assessment of cardiac strain, derived from DIR, is more sensitive than traditional measures such as left ventricular ejection fraction (LVEF), especially in detecting early signs of cardiac failure, including heart failure with preserved ejection fraction (HFpEF) (Pfeffer

*Corresponding author. E-mail address: s_nabavi@sbu.ac.ir

ORCID(s): 0009-0008-4446-0806 (S. Kebriti); 0000-0001-7240-0239 (S. Nabavi); 0000-0001-5135-4800 (A. Gooya)

et al., 2019). Although commonly used in clinical practice, cine MRI sequences such as steady-state free precession (SSFP) do not directly encode tissue motion (Bistroquet et al., 2008). DIR allows for the extraction of deformation-based metrics from cine SSFP MRI data, making advanced cardiac function analysis feasible without additional scans. These strain-based metrics are vital for the diagnosis of cardiomyopathies, the evaluation of valve diseases, the detection of regional dysfunctions, and the guidance of therapeutic interventions (Arratia López et al., 2023).

In this paper, we introduce a novel 3D transformer-based framework for deformable image registration. Unlike conventional CNNs, characterized by their emphasis on local receptive fields, transformers leverage the attention mechanism to learn long-range correspondences within input features. FractMorph processes fixed and moving volumes in parallel transformer streams, exchanging information via our Fractional Cross-Attention (FCA) blocks. Each FCA block enriches feature maps through multi-domain 3D fractional Fourier transform (FrFT) branches. These branches enable the model to capture spatial, mixed spatial-frequency, and frequency domains simultaneously. Finally, a lightweight U-Net style network generates the dense deformation field while preserving detail and smoothness. Compared to existing DIR methods, which struggle to identify both local and global characteristics efficiently, our proposed framework addresses this gap by understanding the local, semi-global, and global correspondences and deformations simultaneously. The key contributions of this study are listed below:

- We propose FractMorph, a novel 3D Transformer-based framework that fuses multi-scale local, semi-global, and global features at every stage for accurate deformable image registration. A lightweight encoder-decoder CNN follows the Transformer to generate high-resolution, detailed deformation fields.
- We introduce the FCA module, which combines multi-order 3D FrFT branches with convolutional and attention mechanisms to enrich feature representations across spatial, spectral, and fractional domains.
- We develop an efficient separable implementation of 3D FrFT and integrate it into our framework, demonstrating improved registration performance with feasible computational cost. To the best of our knowledge, this is the first integration of 3D FrFT into deep learning models for medical imaging.
- We achieve registration accuracy that outperforms current state-of-the-art and recently published approaches using both our main and lightweight variants on an intra-patient cardiac MRI benchmark dataset. Additionally, we demonstrate the generality and solid performance of our approach on a cerebral atlas-to-patient dataset.

The structure of this paper is as outlined below. In Section 2, an overview of related studies is provided. We

present the proposed framework in Section 3. Section 4 presents statistical outcomes alongside qualitative analyses. We discuss these findings in Section 5. Lastly, we present our conclusions and suggest directions for future studies in Section 6.

2. Related Work

2.1. Traditional Methods

Traditional deformation-based alignment techniques have been extensively studied and widely applied in medical imaging over the past several decades. These techniques enable the alignment of complex anatomical structures across modalities and have become standard tools in applications such as neuroimaging, cardiac imaging, radiation therapy planning, and beyond. Key examples include free-form deformation (FFD) algorithms utilizing B-spline models (Jiang et al., 2003), optical flow-based methods such as the Demons algorithm (Cahill et al., 2009), and diffeomorphic approaches like Symmetric Normalization (SyN) (Avants et al., 2008). B-spline FFD methods model deformation as a continuous spline function constructed on a sparse control point mesh. Registration is typically performed by optimizing a cost function that combines an intensity-based similarity metric with a regularization term to suppress irregular transformations (Rueckert et al., 2002). B-spline FFDs are frequently employed in medical imaging applications due to their flexibility in modeling local deformations and robust performance across varying anatomies. However, achieving high registration accuracy often requires a fine control-point grid and a large number of optimization iterations, making the method computationally expensive for large 3D volumes. The Demons algorithm represents a parameter-free approach to deformable image registration that draws an analogy to optical flow and diffusion processes. It iteratively updates a dense displacement field by computing force vectors based on intensity differences and applying a smoothing step at each iteration. In essence, the source image is gradually aligned with the target image by accumulating incremental displacements. While the Demons algorithm is efficient and effective for moderate deformations, it struggles with large anatomical differences. Unlike traditional asymmetric approaches, the SyN algorithm performs symmetric optimization by computing halfway deformations from each image to a shared intermediate space, rather than warping one image entirely to the other. This formulation reduces bias toward either the fixed or moving image and yields inverse-consistent results. SyN has demonstrated excellent performance and has been widely adopted in applications such as neuroimaging, cardiac MRI, and abdominal CT. However, its primary drawback is high computational cost, as it requires substantial memory and runtime to execute effectively. In practice, all these classical methods involve solving an iterative optimization for each new pair of images, and their performance can degrade or require manual tuning when faced with very large or highly localized deformations (Klein et al., 2009).

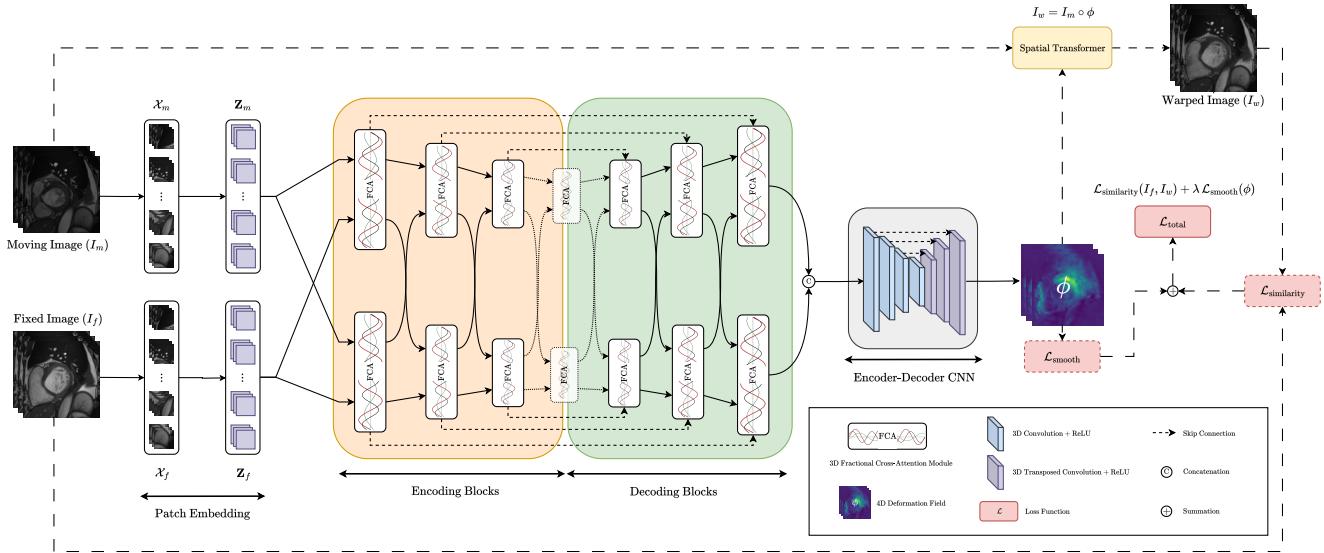


Figure 1: Overview of the proposed FractMorph framework. Given a fixed and a moving 3D image volume, both are first divided into non-overlapping patches and embedded as tokens. These token sequences are jointly processed by symmetric encoding and decoding transformer blocks. At each stage, FCA blocks enable bidirectional interaction between the fixed and moving streams, with skip connections linking encoding and decoding features. After decoding, the two streams are concatenated and passed to a lightweight 3D U-Net style network that outputs the displacement field ϕ . The Spatial Transformer uses this field to warp the moving image toward the fixed image. Ultimately, a similarity loss is computed on the warped-fixed pair, combined with a regularization term for smoothness on the deformation field.

2.2. CNN-Based and Other Non-Transformer Methods

Deep learning-based registration techniques have emerged in recent years to address limitations of traditional methods. These approaches accelerate the registration process by replacing per-case optimization with an end-to-end trainable network. A foundational contribution in this area is VoxelMorph (Balakrishnan et al., 2019), which introduced an unsupervised convolutional U-Net architecture to predict dense displacement fields between a moving and fixed image. VoxelMorph achieved registration accuracy comparable to traditional methods while reducing computational cost. However, its accuracy still leaves room for improvement. To address this limitation, CycleMorph (Kim et al., 2021) introduced a cyclically consistent model designed to register images from A to B and vice versa simultaneously, encouraging deformations that preserve the topological structure.

Beyond convolutional architectures, researchers have explored alternative frameworks that introduce new approaches to modeling the deformation field. For instance, a recent work involves diffusion probabilistic modeling as an alternative to single-pass forward prediction. DiffuseReg (Zhuo and Shen, 2024) employs a denoising diffusion model to iteratively learn the deformation field through diffusion steps. This approach improves interpretability as real-time observation of the registration process is possible. Furthermore, it allows users to inspect intermediate warped outputs and adjust the noise level, which offers more control over the registration behavior. Though DiffuseReg achieved higher accuracy than prior diffusion-based registration models, its

inference time remains considerably slower than single-pass methods. (Jia et al., 2023) explores frequency-domain representations for more efficient registration by proposing Fourier-Net. They replace the U-Net architecture’s expansive path with a parameter-free, model-driven decoder that operates in a band-limited Fourier domain. Also, rather than estimating a high-resolution displacement field within the spatial space, it learns a compact Fourier formulation. By adopting this approach, the design significantly reduces both the parameter count and the number of multiply-add operations.

Another example of alternative frameworks is WarpPINN (Arratia López et al., 2023), which incorporates physics-informed neural networks into DIR that formulates the task with biomechanical constraints. In WarpPINN, the network directly predicts the displacement field by enforcing the limited compressibility of heart tissue by penalizing the Jacobian determinant associated with the deformation. Additionally, the use of Fourier feature mapping of the input images helps WarpPINN overcome its spectral bias and better capture high-frequency deformations. A major challenge here arises from large non-rigid deformations and the complex structure of certain anatomies, such as the heart. To address this, Chang et al. (2024) presents a multi-scale registration model trained in isolation. In place of a singular end-to-end network, a hierarchical structure is implemented to sequentially capture large-scale deformations with increasing resolution. This framework consists of several separate V-Net architectures, each trained on different resolutions of the input images. At the inference

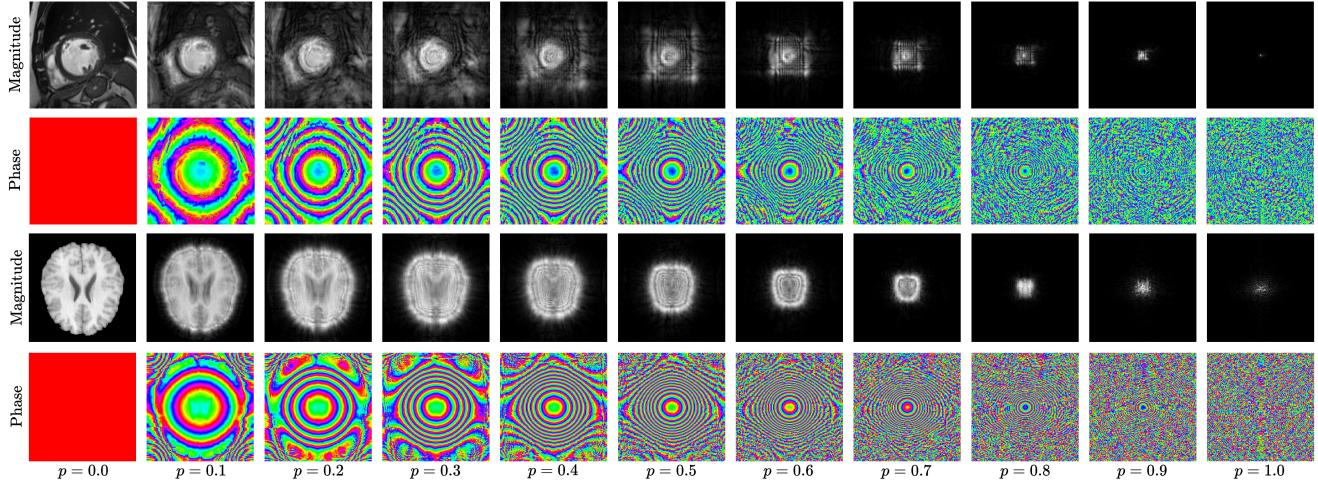


Figure 2: Magnitude and phase of our implemented 3D FrFT applied to cardiac (top two rows) and cerebral (bottom two rows) medical images at different fractional orders. For illustration, only a single 2D slice of each volume is shown.

stage, the moving image is initially processed by the low-resolution network. The predicted deformation field is then upsampled and passed through the next network, and so on. By decomposing the problem, this approach effectively addresses the challenge of very large motions. However, because each scale’s model is trained independently, there is no joint optimization across scales, which increases the risk of inconsistencies when fusing multi-scale deformations. In general, although previous methods perform well in capturing local details, they often struggle to model global context effectively. So far, it remains challenging to find a straightforward approach to capture both local and global context simultaneously.

2.3. Transformer-Based Methods

Transformers, as a new class of architectures, have recently gained significant attention in the research community. These models are distinguished by their proficiency in learning long-range and global dependencies through their attention mechanisms, enabling them to overcome the local receptive field limitations of previous network types. An early example of Transformer usage in image registration is ViT-V-Net (Chen et al., 2021), which integrates a Vision Transformer (ViT) encoder block into a V-Net architecture. This hybrid ConvNet–Transformer model achieved superior performance compared to VoxelMorph. Chen et al. (2022) later presented TransMorph, which leverages Swin Transformer-based encoders (Liu et al., 2021) to retrieve feature maps from the moving and fixed images, and a convolutional decoder to generate the corresponding displacement field. Notably, TransMorph was extended with diffeomorphic warping and a Bayesian variant for uncertainty estimation. TransMorph achieves higher alignment accuracy than previous pure CNN models and also outperforms ViT-V-Net.

More recent works have primarily focused on designing Transformers specialized for the DIR task. For example, Attention-Reg (Song et al., 2022) leverages a cross-modal attention mechanism that maps the features of one image

to those of another, thereby overcoming the feature correspondence gap of previous methods and better capturing the mutual information between images. Another example is XMorpher (Shi et al., 2022), which proposes a dual-parallel Transformer design that processes the moving and fixed images in parallel by keeping correspondence at each stage. It takes the moving and fixed images as separate inputs and learns their correspondence by sharing information via parallel cross-attention blocks. This approach achieves enhanced results over earlier methods, which either concatenated the images before passing them through a network or used separate networks for each image and fused the features afterward. TransMatch (Chen et al., 2023) takes a different strategy to improve feature correspondence. It combines self-attention and cross-attention mechanisms in a two-stream architecture, where each image is first passed through separate self-attention blocks. The resulting features are then combined using cross-attention and passed to a decoder convolution architecture to generate the final deformation field. This approach enables the model to capture correspondences between the moving and fixed images in a hierarchical manner. However, the field still lacks models that can fully and simultaneously address both local features and global context throughout the entire registration process. Moreover, there remain promising strategies that have not yet been explored in this area.

3. Methodology

FractMorph is a deep 3D deformable image registration framework that incorporates a novel parallel transformer to fuse multi-domain features from the input images, followed by an encoder-decoder U-Net style architecture to generate a detailed deformation field. This framework takes a fixed image I_f and a moving image I_m as input and generates a deformation field ϕ that warps I_m to align with I_f . Our proposed framework is designed to capture both local details

and global contextual information by integrating convolutional processing, FrFT branches, and a cross-attention mechanism between I_f and I_m . The overall pipeline of our work is illustrated in Fig. 1.

3.1. Patch Embedding

The Patch Embedding module is the first component of the network and is responsible for converting the input 3D image volumes into structured sequences of feature tokens. It converts the input volumes fixed image ($I_f \in \mathbb{R}^{H \times W \times D}$) and moving image ($I_m \in \mathbb{R}^{H \times W \times D}$) into a set of feature map tokens which are suitable for transformer processing. The 3D volume of each of the two input images (I_f and I_m) is partitioned into non-overlapping cubic blocks of dimensions $P_x \times P_y \times P_z$. This results in a total of $N = \frac{H \cdot W \cdot D}{P_x \cdot P_y \cdot P_z}$ patches per image. Let us denote the set of 3D patches extracted from an image I as:

$$\mathcal{X} = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^N\}, \quad \mathbf{x}^i \in \mathbb{R}^{P_x \times P_y \times P_z} \quad (1)$$

Each patch \mathbf{x}^i is then mapped to a d -dimensional token using a learnable linear projection matrix $\mathbf{E} \in \mathbb{R}^{P_x \times P_y \times P_z \times d}$. Specifically, each token is computed as:

$$\mathbf{z}^i = \mathbf{x}^i \mathbf{E}, \quad \mathbf{z}^i \in \mathbb{R}^d \quad (2)$$

Embedded token sequences are obtained by applying this to all patches:

$$\mathbf{Z}_f = \{\mathbf{z}_f^1, \mathbf{z}_f^2, \dots, \mathbf{z}_f^N\}, \quad \mathbf{Z}_m = \{\mathbf{z}_m^1, \mathbf{z}_m^2, \dots, \mathbf{z}_m^N\} \quad (3)$$

Here, \mathbf{Z}_f and \mathbf{Z}_m are the token representations of the fixed and moving volumes, respectively. This patch-based embedding provides two main benefits. First, it reduces the spatial resolution of the input, enabling more efficient processing and attention across the volume. Second, by mapping each patch to a higher-dimensional and semantically more meaningful feature space, the network gains more expressive capacity for modeling structures.

3.2. Dual-Parallel Transformer

The core of our framework is a dual-input Transformer that processes the moving and fixed image tokens in parallel and enables explicit cross-feature matching between them. This architecture integrates our novel FCA mechanism for feature matching. Our transformer utilizes a dual-stream feature extraction approach, with one stream dedicated to the moving image and the other to the fixed image. In contrast to conventional backbone designs that either process a single image or use a single stream to handle the concatenation of two images, this dual-stream approach enables the network to exchange information between the moving and fixed images at multiple levels of representation. By propagating mutual information across multiple resolutions, the network can uncover multi-level semantic correspondences while simultaneously extracting features. This design facilitates more effective alignment, as features from I_f and I_m are progressively brought into correspondence at different scales. Let \mathbf{F}_m^l and \mathbf{F}_f^l ($\mathbf{F}_m^l, \mathbf{F}_f^l \in \mathbb{R}^{H_l \times W_l \times D_l \times C_l}$)

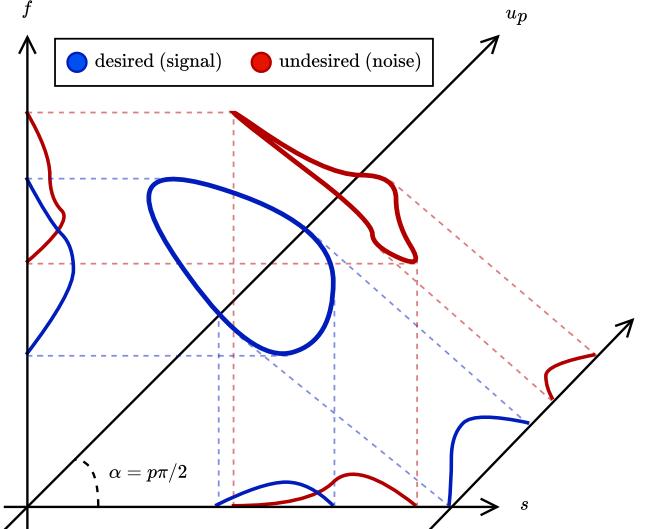


Figure 3: FrFT-based separation of desired from undesired components in the spatial–spectral domain, where the spatial (s), frequency (f), and fractional order (u_p) axes are shown.

denote the moving and fixed feature maps at level l of the encoder. Level 0 takes the output of the patch embedding layer as its input. The encoder of our transformer consists of L levels with increasing channel dimensions and decreasing spatial resolution. At each level l , a series of Transformer blocks is applied. n_l denotes the number of blocks at level l . Each Transformer block takes $(\mathbf{F}_m^l, \mathbf{F}_f^l)$ as input and produces refined feature maps $(\tilde{\mathbf{F}}_m^l, \tilde{\mathbf{F}}_f^l)$ through dual cross-attention fusion. After processing each level l , the feature maps are downsampled by a scaling ratio of 2 across all spatial dimensions to produce the inputs \mathbf{F}_m^{l+1} and \mathbf{F}_f^{l+1} for the next level:

$$H_{l+1} = \frac{H_l}{2}, \quad W_{l+1} = \frac{W_l}{2}, \quad D_{l+1} = \frac{D_l}{2}, \quad C_{l+1} = 2 C_l \quad (4)$$

This builds a low-resolution, high-semantic latent representation at the coarsest level $l = L - 1$.

The decoder mirrors this process in reverse and also has L levels, with the same n_l Transformer blocks per level as the encoder. Starting from the coarsest features, the feature maps are progressively upsampled and refined back to higher resolution. At each decoding level l , the upsampled moving and fixed feature maps are merged with the skip connections $(\tilde{\mathbf{F}}_m^l, \tilde{\mathbf{F}}_f^l)$ from the encoder. Then, a linear fusion is applied to halve the channel dimensionality, and n_l cross-attention Transformer blocks fuse and refine these features. By the end of the decoder, feature maps with the identical patch resolution as level 0 of the encoder are recovered. These two feature maps are finally concatenated along the channel dimension and passed to a U-Net style network for final flow prediction. This dual-parallel stream architecture, with repeated cross-attention fusion, ensures that the output features encapsulate enriched information from both images across multiple domains and

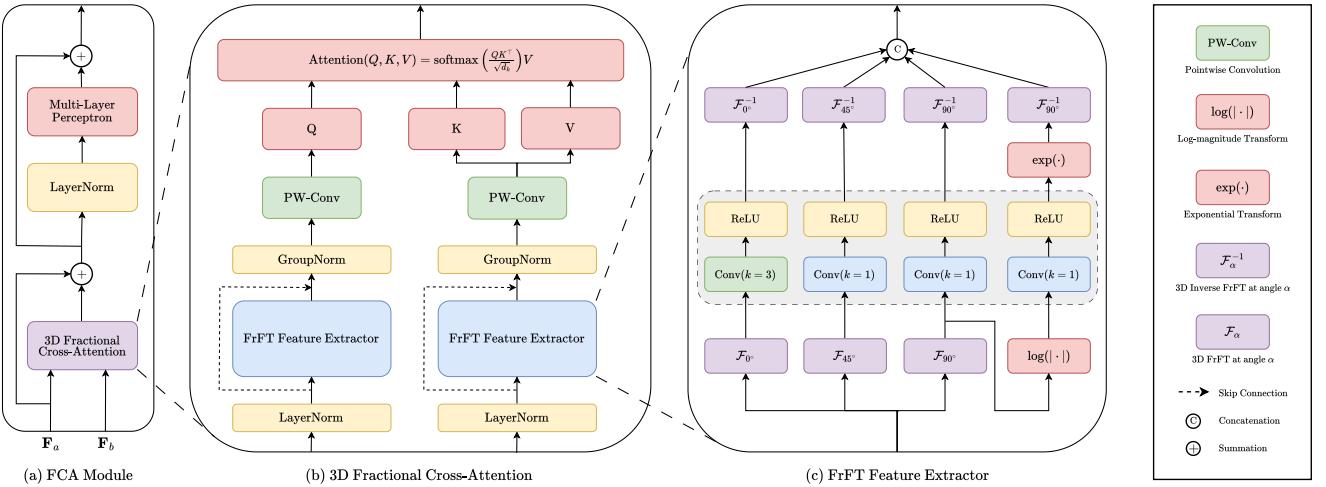


Figure 4: (a) Overview of the proposed FCA module. (b) Internal structure of the Fractional Cross-Attention mechanism within the FCA module. (c) Architecture of the FrFT Feature Extractor, showing its four FrFT branches.

scales, facilitating precise deformation prediction for the U-Net architecture.

3.3. Fractional Fourier Transform

3.3.1. Theory and Fundamental Properties

The FrFT generalizes the standard Fourier transform by introducing a fractional order parameter p , which corresponds to a rotation by angle $\alpha = p\pi/2$ in the combined spatial-frequency domain (Almeida, 1994). This makes the FrFT a powerful tool for observing different characteristics of a signal by varying its fractional degree α . For a 1D signal $x(t)$, the p th-order FrFT $X_p(u)$ is defined by the integral transform:

$$X_p(u) = \mathcal{F}_p\{x\}(u) = \int_{-\infty}^{\infty} K_p(t, u) x(t) dt \quad (5)$$

where $K_p(t, u)$ is the fractional kernel. For $\alpha \neq n\pi$, this kernel is given by:

$$K_p(t, u) = A_\alpha e^{j\left(\frac{t^2 \cot \alpha}{2} - ut \csc \alpha + \frac{u^2 \cot \alpha}{2}\right)} \quad (6)$$

with $A_\alpha = \sqrt{\frac{1-j \cot \alpha}{2\pi}}$ as a normalization factor. For the special cases where α is an integer multiple of π , the kernel reduces to a Dirac delta function. Specifically, $K_p(t, u) = \delta(t - u)$ when $\alpha = 2n\pi$, and $K_p(t, u) = \delta(t + u)$ when $\alpha = (2n+1)\pi$. Also notably $\alpha = \pi/2$ ($p = 1$) yields $\cot \alpha = 0$ and $\csc \alpha = 1$, so the kernel reduces to $K_{p=1}(t, u) = e^{-j t u}$, which reproduces the standard Fourier transform formula. The FrFT continuously interpolates between the spatial domain ($p = 0$, no rotation) and the frequency domain ($p = 1, 90^\circ$ rotation). In simple terms, a conventional image and its Fourier spectrum are just orthogonal extreme cases in the spatial-frequency domain. In contrast, the FrFT can represent the image at any intermediate rotation angle that simultaneously reflects both spatial and frequency characteristics of the image. This means applying a fractional transform by angle α is equivalent to taking the signal's joint time-frequency

energy distribution and rotating it by α in the plane (Ozaktas et al., 1996). For $p = 0.5$ (45° rotation), the representation is halfway between pure spatial and pure frequency view.

3.3.2. Applications in Image Processing

As previously noted, the fractional order p controls the balance between spatial and frequency information in the transformed signal. This behavior is illustrated in Fig. 2, which exhibits how the FrFT progressively transforms input images as the fractional order increases. At $p = 0$, the transformation fully preserves the original spatial structure. At $p = 1$, the FrFT becomes the conventional Fourier transform, concentrating the image energy in the frequency domain and suppressing spatial details. At intermediate values, such as $p = 0.5$, the FrFT produces representations that retain both spatial and spectral features, as shown in the figure. This property enables the capture of image characteristics ranging from coarse to fine across different fractional domains. In practice, extracting features at multiple fractional orders helps construct a rich, multi-scale feature set. It is well known that natural images, including medical images, are highly non-stationary, meaning that their local statistics, such as intensity distributions and textures, vary significantly across different regions. Here, the FrFT provides an advantage by capturing localized frequency components more effectively than the conventional Fourier transform. The FrFT in image processing has the ability to enhance feature extraction, especially for edges, oriented textures, and other direction-dependent features. Studies have shown that within the fractional Fourier domain, the phase component of an image carries rich texture and edge information (Zeng and Gao, 2015). This means that by tuning the FrFT order, different structural details of the image become pronounced. For instance, gradients and vessel-like structures in medical images can be better captured using FrFT-based filters or transforms.

Another practical advantage of using FrFT for image processing is its ability to suppress noise more effectively

Table 1

Parameter counts and FLOPs of convolution operations in each branch of the FCA feature extractor module with channel coefficient α , assuming an input tensor of size $C \times D \times H \times W$.

Branch	Kernel Size	Channels (In \rightarrow Out)	Parameters	FLOPs
FrFT0	3	$\alpha C \rightarrow \alpha C$	$27\alpha^2C^2$	$27\alpha^2C^2 D H W$
FrFT45	1	$2\alpha C \rightarrow 2\alpha C$	$4\alpha^2C^2$	$4\alpha^2C^2 D H W$
FrFT90	1	$2\alpha C \rightarrow 2\alpha C$	$4\alpha^2C^2$	$4\alpha^2C^2 D H W$
Log-magnitude	1	$\alpha C \rightarrow \alpha C$	α^2C^2	$\alpha^2C^2 D H W$
Total	-	-	$36\alpha^2C^2$	$36\alpha^2C^2 D H W$

than a conventional Fourier approach in many situations. The key is that FrFT can target noise that is not uniform across the image (i.e. non-stationary noise). A classic Fourier transform is effective if noise is concentrated at specific frequencies globally, but it struggles when noise characteristics vary over space. In contrast, the FrFT can adapt to such variations. Operating in a partial frequency domain, it can selectively isolate and suppress noise patterns that change with position (Subramaniam et al., 2010; Yang, 2012). Figure 3 illustrates this capability in the spatial-spectral representation of a non-stationary signal, where adjusting the fractional degree α enables effective separation of signal components from position-dependent noise. Such capability is highly beneficial for tasks like deformable image registration, where noise can otherwise disrupt the alignment of images. The above benefits make FrFT particularly useful in challenging image analysis domains like medical imaging. Medical images often contain structures with varying scales and orientations. For example, the heart muscle fibers have oriented textures, blood vessels form curvilinear patterns, and different tissues yield complex frequency content. FrFT-based methods have also shown promising results within the scope of image registration, as demonstrated in previous studies (Zhang et al., 2013).

3.3.3. Numerical Implementation and 3D Extension

In our work, inspired by the implementation of the 1D FrFT by Pei and Yeh (1997) and the 2D FrFT developed by Yu et al. (2023), we extend the FrFT to three-dimensional images. To implement the 3D FrFT, we take advantage of the fact that multidimensional transforms are separable (Pei and Yeh, 1998; Sahin et al., 1995). Thus, a 3D FrFT can be achieved by sequentially performing 1D fractional transforms along each axis (x, y, and z). We adopt the eigenvector-decomposition approach for the 1D discrete FrFT. A complete set of continuous FrFT eigenfunctions of order p is given by the Hermite-Gaussian functions $\psi_n(x)$, which satisfy:

$$\mathcal{F}_p\{\psi_n\}(x) = e^{-j p n \pi / 2} \psi_n(x) \quad (7)$$

and admit the closed-form:

$$\psi_n(x) = \frac{2^{1/4}}{\sqrt{2^n n!}} H_n(\sqrt{2\pi} x) \exp(-\pi x^2) \quad (8)$$

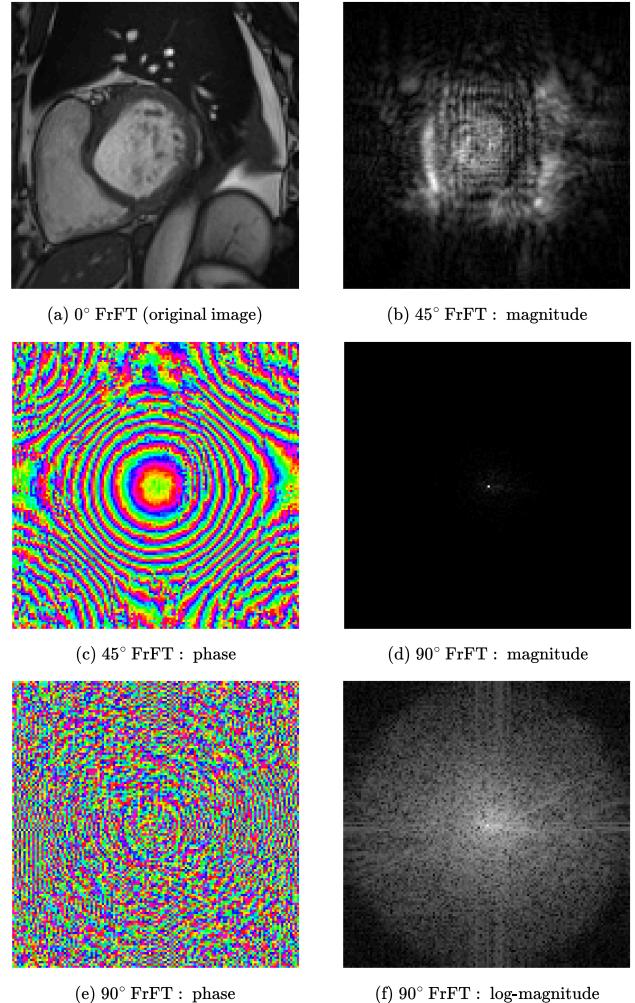


Figure 5: FrFT outputs from the FCA feature extractor branches, shown before the convolution layers. The operations are applied to a 3D medical image. For illustration, only a single 2D slice of the volume is displayed. (a) shows the result of \mathcal{F}_{0° operation. (b)-(c) show the magnitude and phase of \mathcal{F}_{45° operation, respectively. (d)-(e) illustrate the magnitude and phase of \mathcal{F}_{90° operation, respectively. (f) shows the logarithm of \mathcal{F}_{90° 's magnitude.

where $H_n(\cdot)$ is the n th Hermite polynomial. From Mehler's formula, the continuous FrFT kernel can be written spectrally as:

$$K_p(t, u) = \sum_{n=0}^{\infty} \psi_n(t) e^{-j p n \pi / 2} \psi_n(u) \quad (9)$$

To obtain a discrete implementation, we truncate and sample Equation (9) following Candan et al. (2000). The resulting discrete FrFT of a length- N vector $x[n]$ is:

$$X_p[m] = \mathcal{F}_p\{x\}[m] = \sum_{k=0}^{N-1} K_p[m, k] x[k] \quad (10)$$

Table 2

Quantitative results on performance, model complexity, runtime, and memory on the ACDC dataset. Metrics are reported as mean \pm std for overall DSC, average DSC across structures, HD95, fraction of non-positive Jacobian voxels, and Jacobian std. FractMorph-Light and FractMorph are our proposed lightweight and primary models. **Bold** and underline indicate best and second-best results, respectively.

Method	Overall DSC (%) \uparrow	Avg. DSC (%) \uparrow	HD95 (mm) \downarrow	$\% J_\phi \leq 0 \downarrow$	STD($ J_\phi $) \downarrow	Parameters	Time (s)	Memory (MB)
Initial	80.77 ± 7.05	58.41 ± 11.96	2.73 ± 1.09	—	—	—	—	—
ANTs (SyN)	83.70 ± 6.84	65.98 ± 12.25	2.22 ± 1.00	0.93 ± 3.01	0.08 ± 0.03	—	37.79	203
Demons	83.43 ± 6.16	71.23 ± 9.73	2.08 ± 0.90	0.77 ± 0.42	0.40 ± 0.07	—	18.86	27
VoxelMorph	86.08 ± 5.99	74.88 ± 8.69	1.58 ± 0.68	0.02 ± 0.03	0.14 ± 0.03	327,331	0.02	181
Fourier-Net	83.09 ± 6.01	60.68 ± 12.09	2.36 ± 0.97	0.00 ± 0.00	0.03 ± 0.01	879,677	0.02	38
TransMorph	82.87 ± 7.22	68.11 ± 11.70	2.05 ± 0.88	0.06 ± 0.01	0.15 ± 0.05	46,689,459	0.30	491
XMorpher	84.96 ± 4.04	70.09 ± 9.53	1.82 ± 0.77	<u>0.02 ± 0.04</u>	<u>0.13 ± 0.03</u>	15,093,891	0.46	207
TransMatch	86.19 ± 4.76	47.25 ± 8.54	3.21 ± 1.59	0.03 ± 0.04	0.14 ± 0.03	112,262,611	0.38	1,008
FractMorph-Light	<u>86.32 ± 4.86</u>	<u>74.97 ± 9.07</u>	<u>1.57 ± 0.82</u>	0.06 ± 0.06	0.15 ± 0.03	29,630,931	0.34	322
FractMorph	<u>86.45 ± 4.72</u>	<u>75.15 ± 8.95</u>	<u>1.54 ± 0.78</u>	0.05 ± 0.04	0.15 ± 0.03	63,910,483	0.36	461

with the spectral kernel:

$$\mathcal{K}_p[m, n] = \sum_{k=0}^{N-1} u_k[m] e^{-j \frac{\pi}{2} p k} u_k[n] \quad (11)$$

where $u_k[n]$ are the discrete Hermite–Gaussian eigenvectors.

To compute the 3D FrFT of order p on a volume $x[n_x, n_y, n_z]$, we apply three successive 1D FrFTs along the x -, y -, and z -axes:

$$X_p[m_x, m_y, m_z] = \mathcal{F}_p^{(z)} \left\{ \mathcal{F}_p^{(y)} \left\{ \mathcal{F}_p^{(x)} \{x\} \right\} \right\} [m_x, m_y, m_z] \quad (12)$$

where $\mathcal{F}_p^{(x)}$, $\mathcal{F}_p^{(y)}$, and $\mathcal{F}_p^{(z)}$ denote applying the 1D FrFT of order p along the x -, y -, and z -axes, respectively. The inverse 3D FrFT is obtained by applying the inverse fractional transforms of order $-p$ along each axis:

$$x[n_x, n_y, n_z] = \mathcal{F}_{-p}^{(x)} \left\{ \mathcal{F}_{-p}^{(y)} \left\{ \mathcal{F}_{-p}^{(z)} \{X_p\} \right\} \right\} [n_x, n_y, n_z] \quad (13)$$

with $\mathcal{F}_{-p} = (\mathcal{F}_p)^{-1}$. This separable formulation offers superior computational efficiency compared to naive multi-dimensional FrFT implementations (Yu et al., 2023).

3.4. Fractional Cross-Attention Module

At the core of each transformer block is our novel FCA module. The architecture of this module is illustrated in Fig. 4. This module enhances the cross-attention mechanism by incorporating multi-domain feature extraction through the FrFT. Each block processes a pair of input feature tensors \mathbf{F}_m^l and \mathbf{F}_f^l of shape $\mathbb{R}^{H_l \times W_l \times D_l \times C_l}$. As implied by Fig. 4b, the FCA block consists of two main stages. The first stage is a multi-branch FrFT feature extractor that operates separately on each input. The second stage involves cross-attention, which matches the features between the two inputs. In the following sections, explanations of each stage are provided.

3.4.1. Fractional Fourier Feature Extraction

To enrich the features with both spatial and spectral representations, each normalized input is passed through four parallel branches, as illustrated in Fig. 4c. In each

branch, a FrFT with a specific fractional order is applied to the input, followed by a convolution operation and a ReLU activation. The convolution kernel size is adapted to the domain. Specifically, in the spectral domain, a kernel size of 1 is sufficient since, according to the convolution theorem (Bracewell and Kahn, 1966):

$$\mathcal{F}\{f * g\} = \mathcal{F}\{f\} \cdot \mathcal{F}\{g\} \quad (14)$$

where $\mathcal{F}\{\cdot\}$ denotes the Fourier transform, $*$ represents convolution in the spatial domain, and \cdot denotes element-wise multiplication in the spectral domain. This property implies that each spectral coefficient already encodes global spatial information, making larger kernels unnecessary. In contrast, in the spatial domain, a kernel size of 3 is used to effectively aggregate local neighborhood information. Finally, each branch passes through its corresponding inverse FrFT to transform the features back to the spatial domain, enabling the model to capture and enhance features in several FrFT domains. According to the experiments by Yu et al. (2023), fractional orders of $\alpha = 0^\circ$, $\alpha = 45^\circ$, and $\alpha = 90^\circ$ are chosen for capturing local, semi-global, and global features, respectively. By selecting only these three fractional domains, our model facilitates effective feature enrichment while maintaining an appropriate level of computational complexity. Additionally, we adopt a log-magnitude branch that operates on the magnitude of the FrFT at $\alpha = 90^\circ$. As demonstrated by Gonzalez (2009), the log-magnitude representation can reveal finer details within a transform, thereby enhancing the model's ability to capture spectral information. Figure 5 illustrates this effect, along with representative outputs from the different fractional Fourier transform branches. This transform is defined as $A = \log(1 + |X_{90^\circ}(u)|)$. After applying the convolution operation, the logarithm is inverted as $|X_{90^\circ}(u)| = \exp(A) - 1$. After the four branch transformations, four distinct feature maps are obtained, each with the same shape as the input ($D_l \times H_l \times W_l \times C_l$). The outputs from the branches and the skip connection are concatenated along the channel dimension. Each branch is normalized, and after that, a pointwise convolution is applied to fuse them into a single unified representation. This fused feature

Table 3

Quantitative results on performance, runtime, and memory on the LPBA40 dataset.

Method	Overall DSC (%) ↑	Avg. DSC (%) ↑	HD95 (mm) ↓	% J _ϕ ≤ 0 ↓	STD(J _ϕ) ↓	Time (s)	Memory (MB)
Initial	82.06 ± 2.82	54.79 ± 4.32	6.28 ± 1.74	—	—	—	—
ANTs (SyN)	92.38 ± 0.43	71.83 ± 0.98	2.32 ± 0.22	1.63 ± 1.28	0.16 ± 0.01	303.87	214
Demons	86.23 ± 0.19	61.68 ± 3.51	2.70 ± 1.07	0.04 ± 0.01	0.22 ± 0.01	63.11	244
VoxelMorph	91.51 ± 0.80	68.45 ± 2.73	1.20 ± 0.35	0.51 ± 0.05	0.29 ± 0.02	0.26	3,608
Fourier-Net	87.56 ± 1.21	61.60 ± 2.50	1.93 ± 0.44	0.28 ± 0.09	0.20 ± 0.02	0.05	819
TransMorph	91.79 ± 0.99	<u>68.68 ± 2.75</u>	1.23 ± 0.42	0.78 ± 0.06	0.33 ± 0.02	0.61	4,356
XMorpher	91.08 ± 0.90	67.78 ± 1.96	1.24 ± 0.39	0.46 ± 0.05	0.30 ± 0.02	3.44	2,309
TransMatch	91.36 ± 0.65	49.84 ± 1.98	1.32 ± 0.40	0.57 ± 0.04	0.30 ± 0.01	0.45	3,945
FractMorph-Light	91.76 ± 0.69	68.26 ± 2.81	1.12 ± 0.26	0.70 ± 0.06	0.32 ± 0.01	5.23	3,894
FractMorph	91.79 ± 0.74	68.52 ± 2.58	1.09 ± 0.18	0.53 ± 0.03	0.29 ± 0.01	5.26	4,037

map contains information from multiple complementary perspectives: spatial localized details, semi-global fractional domain, global spectral context, and detailed spectral magnitude. By combining these, the network fully utilizes the advantages of both the spatial and frequency domains while maintaining feasible computational complexity, as shown in Table 1. This table summarises the parameter counts and floating-point operations (FLOPs) for the convolution operations in each branch of the FCA feature extractor, assuming an input tensor of size $C \times D \times H \times W$ and channel coefficient α . FLOPs measure the total number of arithmetic operations required to process data. Lower FLOPs generally indicate higher computational efficiency. The Parameters and FLOPs columns are computed from the kernel size and the number of input and output channels in each branch. We also propose two versions of our model: one in which each branch operates on the entire set of input channels, and another in which the input channels are split so that each branch processes only a portion of the channels, thereby reducing the model’s complexity. Formally, regarding Table 1, for our main model we set the channel coefficient to $\alpha = 1$, which yields a total of $36 C^2$ parameters and $36 C^2 D H W$ convolutional FLOPs. For our lightweight model, we set $\alpha = \frac{1}{3}$, which reduces the overall branch parameters and convolutional FLOPs to $\frac{1}{9}$ of the main model.

3.4.2. Cross-Attention Mechanism

After enriching each input feature map, cross-attention is performed to exchange information between the enriched moving feature map O_m and the fixed feature map O_f . The spatial dimensions of each feature map are first flattened into a sequence of length $N_l = D_l \cdot H_l \cdot W_l$. The query matrix $Q_m \in \mathbb{R}^{N_l \times C}$ is computed from O_m , while the key and value matrices $K_f, V_f \in \mathbb{R}^{N_l \times C}$ are computed from O_f , all using learned linear projections. The moving-to-fixed attention output is then computed as:

$$\text{Cross-Attention}(m \rightarrow f) = \text{softmax} \left(\frac{Q_m K_f^T}{\sqrt{d_k}} \right) V_f \quad (15)$$

where $d_k = C/h$ is the dimension of the key vectors, with h denoting the number of attention heads. Similarly, in parallel, the reverse fixed-to-moving Cross-Attention($f \rightarrow m$) is

computed. This cross-attention mechanism explicitly links features in I_m and I_f that are mutually relevant, in contrast to standard self-attention, which only considers intra-image relationships. After the attended outputs are reshaped back to $D_l \times H_l \times W_l \times C$, the original input corresponding to the query is added. This is followed by layer normalization, a feed-forward multi-layer perceptron, and a skip connection for each cross-attention, as illustrated in Fig. 4a. This design encourages effective communication between the features of the two images at each level, while the fractional branches ensure that the attention mechanism has access to enriched feature representations spanning the spatial, mixed spatial-frequency, and frequency domains.

3.5. Encoder-Decoder CNN for Deformation Field Generation

The final stage of FractMorph is a lightweight U-Net style architecture (Ronneberger et al., 2015) that generates the dense 3D deformation field ϕ from the fused Transformer features. After the dual Transformer decoder, feature maps for both the moving and fixed images are obtained. These feature maps are first concatenated and normalized. To restore the original spatial resolution, a reverse patch embedding is applied before the features are passed into the U-Net-style network. This network then processes these high-resolution features to generate the deformation field. Our U-Net architecture consists of three convolutional down-sampling layers, a bottleneck, and three deconvolutional up-sampling layers. It is a suitable choice because the Transformer generates multi-domain features that integrate local, semi-global, and global information. The U-Net architecture further refines these features to produce an accurate deformation field with well-preserved localized details.

3.6. Spatial Transformation Function

To calculate the similarity loss, the moving image I_m is warped using the predicted deformation field ϕ to obtain the resulting warped image I_w , which is then compared with the fixed image I_f . To achieve this, we use a Spatial Transformation Function as proposed by Jaderberg et al. (2015) and has also been widely adopted in DIR works such as (Balakrishnan et al., 2019; Chen et al., 2022).

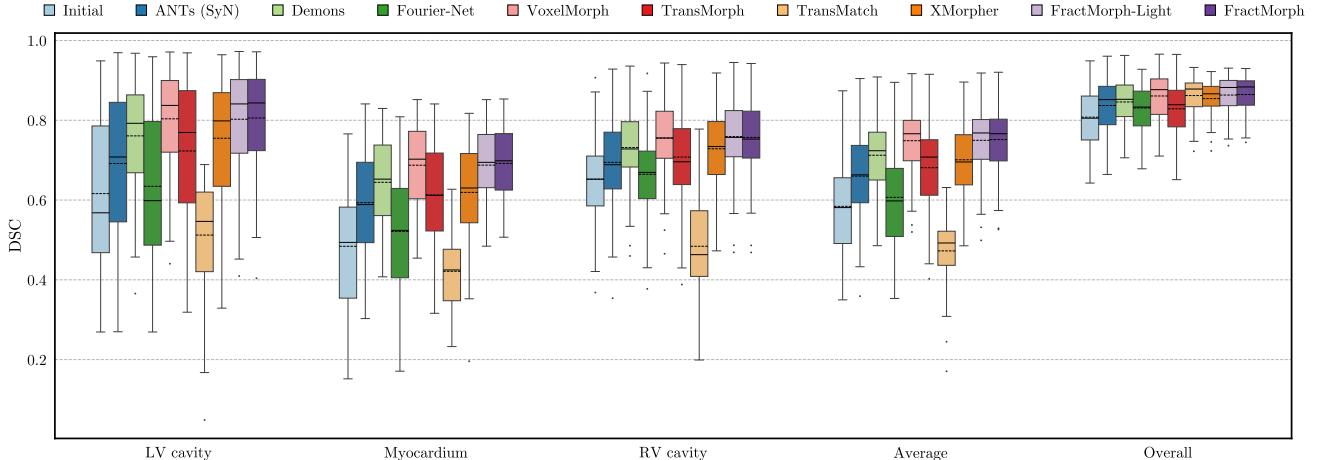


Figure 6: Boxplots of Dice similarity coefficient for left ventricular cavity, myocardium, and right ventricular cavity segmentations, as well as the per-structure average and overall scores across nine registration methods on the ACDC dataset. FractMorph-Light and FractMorph are our proposed lightweight and primary models, respectively.

For each voxel p , the deformation field ϕ maps it to a new location given by $p' = p + \phi(p)$. Since image intensities are defined only at integer voxel locations, trilinear interpolation is applied to estimate the intensity at p' using the values of its eight neighboring voxels:

$$I_w(p) = I_m \circ \phi(p) \quad (16)$$

$$I_m \circ \phi(p) = \sum_{q \in \mathcal{N}(p')} I_m(q) \prod_{d \in \{x,y,z\}} (1 - |p'_d - q_d|) \quad (17)$$

where $\mathcal{N}(p')$ denotes the set of voxel neighbors of p' , and d iterates over the three spatial dimensions. This differentiable interpolation allows gradients to flow through the warping operation, which enables the model to learn the optimal deformation field ϕ end-to-end.

3.7. Loss Function

We train FractMorph in an unsupervised manner by combining an image similarity loss with a deformation regularization term. Adopting an unsupervised approach allows the model to be applicable to a wider range of datasets without requiring ground-truth deformation fields or anatomical segmentations. The total loss is defined as:

$$\mathcal{L}_{\text{total}}(I_f, I_m, \phi) = \mathcal{L}_{\text{similarity}}(I_f, I_m \circ \phi) + \lambda \mathcal{L}_{\text{smooth}}(\phi) \quad (18)$$

where $(I_m \circ \phi)$ denotes the moving image I_m warped by the deformation field ϕ , producing the resulting image I_w . The function $\mathcal{L}_{\text{similarity}}(\cdot, \cdot)$ measures the similarity between the fixed image I_f and the warped moving image $(I_m \circ \phi)$, while $\mathcal{L}_{\text{smooth}}(\cdot)$ encourages smoothness in the deformation field ϕ . The parameter λ controls the trade-off between similarity and smoothness. In the following subsections, we explain each term in detail.

3.7.1. Image Similarity Loss

We use the local cross-correlation (CC) between the fixed image I_f and the warped moving image $(I_m \circ \phi)$ as the similarity measure. The CC is robust to intensity variations across scans and datasets. Specifically, the local mean intensities $\hat{I}_f(p)$ and $\widehat{I_m \circ \phi}(p)$ are computed over a local n^3 neighborhood centered at voxel p :

$$\hat{I}_f(p) = \frac{1}{n^3} \sum_{p_i} I_f(p_i) \quad (19)$$

$$\widehat{I_m \circ \phi}(p) = \frac{1}{n^3} \sum_{p_i} (I_m \circ \phi)(p_i) \quad (20)$$

where p_i iterates over the neighborhood. The local cross-correlation is then given by:

$$CC(I_f, I_m \circ \phi) = \sum_{p \in \Omega} \frac{(\sum_{p_i} [I_f(p_i) - \hat{I}_f(p)] [\widehat{I_m \circ \phi}(p_i) - \widehat{I_m \circ \phi}(p)]^2)}{(\sum_{p_i} [I_f(p_i) - \hat{I}_f(p)]^2) (\sum_{p_i} [\widehat{I_m \circ \phi}(p_i) - \widehat{I_m \circ \phi}(p)]^2)} \quad (21)$$

A higher CC indicates better alignment, so the similarity loss is defined as:

$$\mathcal{L}_{\text{similarity}}(I_f, I_m, \phi) = -CC(I_f, I_m \circ \phi) \quad (22)$$

3.7.2. Smoothing Regularization Term

Minimizing only the similarity loss can lead to deformation fields that are non-smooth and unrealistic. To ensure that the estimated deformation field ϕ is spatially smooth, we add a diffusion regularizer on its spatial gradients. The smoothness loss is defined as:

$$\mathcal{L}_{\text{smooth}}(\phi) = \sum_{p \in \Omega} \|\nabla \phi(p)\|^2 \quad (23)$$

where $\nabla \phi(p)$ denotes the spatial gradient at voxel p :

$$\nabla \phi(p) = \left(\frac{\partial \phi(p)}{\partial x}, \frac{\partial \phi(p)}{\partial y}, \frac{\partial \phi(p)}{\partial z} \right) \quad (24)$$

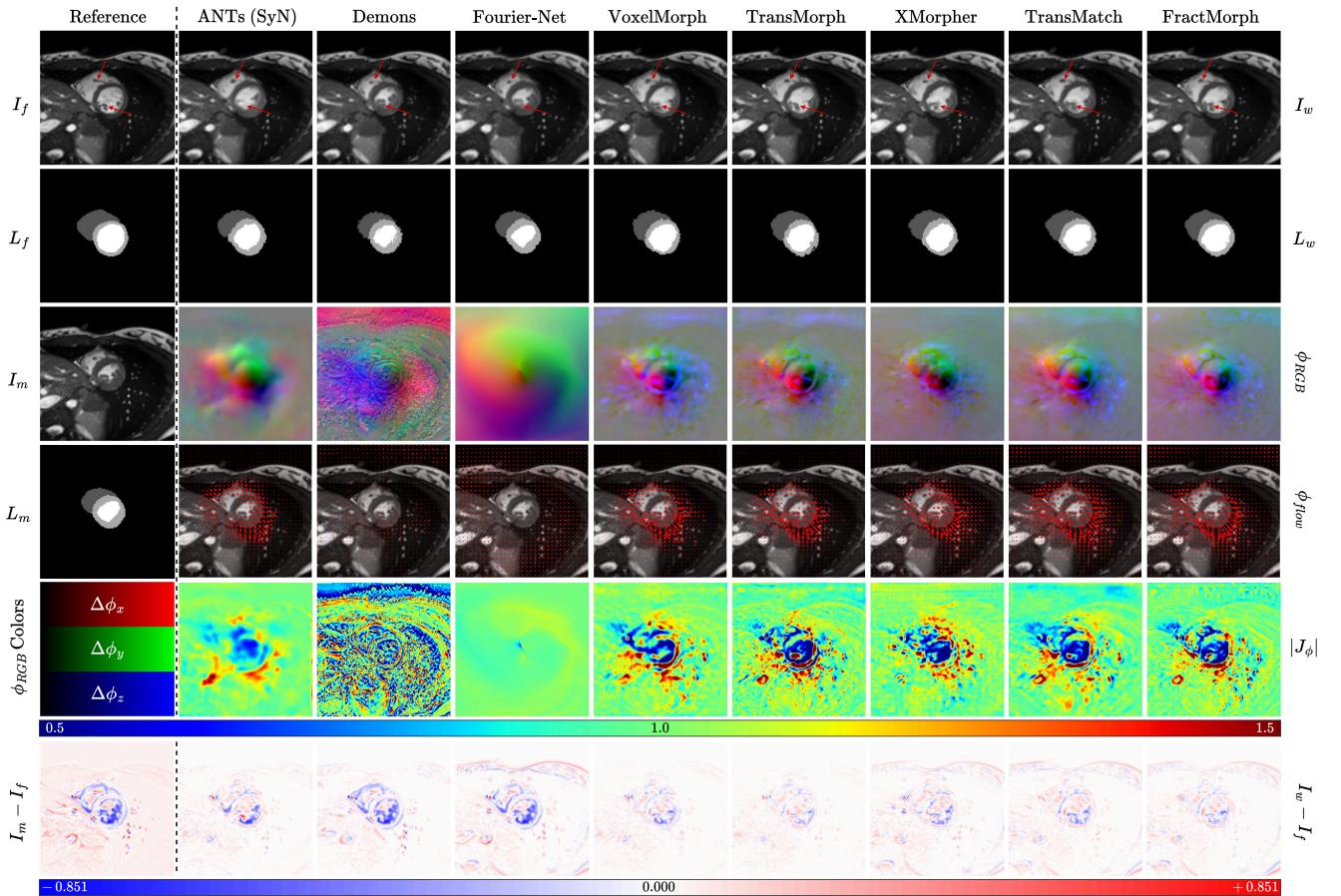


Figure 7: Qualitative comparison of registration methods applied to a fixed image (I_f) and a moving image (I_m), along with their corresponding anatomical segmentation labels (L_f and L_m , respectively). From top to bottom: the warped image ($I_w = I_m \circ \phi$) and warped label ($L_w = L_m \circ \phi$), the generated deformation field (visualized in RGB), the 2D deformation flow along the x and y axes for the shown slice, the Jacobian determinant of the deformation field, and the difference between the warped and fixed images are shown. This example highlights the superiority of our model in capturing fine-grained local deformations, as indicated by the red arrows in the first row.

The partial derivatives are approximated by finite differences between neighboring voxels, for example:

$$\frac{\partial \phi(p)}{\partial x} \approx \phi(p_x + 1, p_y, p_z) - \phi(p_x, p_y, p_z) \quad (25)$$

and similarly for $\frac{\partial \phi(p)}{\partial y}$ and $\frac{\partial \phi(p)}{\partial z}$

4. Experiments and Results

4.1. Datasets

We evaluated our approach on the cardiac MRI ACDC (Bernard et al., 2018) and cerebral MRI LPBA40 (Shattuck et al., 2008) datasets. For ACDC, we performed intra-patient registration, while for LPBA40, we applied atlas-to-patient registration. The ACDC dataset contains cardiac cine MRI scans acquired at end-diastole (ED) and end-systole (ES). It includes 150 subjects, both healthy and with various cardiac pathologies. Expert annotations are provided for key cardiac structures: the left ventricular (LV) cavity, myocardium, and right ventricular (RV) cavity in both ED and ES phases. The

dataset is divided into five groups of 30 cases each, categorized by physiological characteristics such as ventricular volumes, ejection fraction, local LV contraction, LV mass, and myocardial thickness. From each group, we selected 18 cases for training, 2 for validation, and 10 for testing, yielding 90 training cases, 10 validation cases, and 50 test cases. The test set corresponds to the official split proposed in (Bernard et al., 2018), while training and validation cases were randomly sampled from the remainder. All scans were normalized to $[0, 1]$ and resized to $16 \times 128 \times 128$, corresponding to a voxel size of $10 \times 1.8 \times 1.8$ mm. Ground-truth segmentations were used only for evaluation, as training was conducted in an unsupervised manner. The LPBA40 dataset comprises 40 T1-weighted brain MRI scans, each with 56 anatomical structures annotated by experts. All scans are skull-stripped, intensity-normalized, and resized to $160 \times 192 \times 160$. We split the dataset into 30 cases for training and 9 for testing, with one case designated as the atlas.

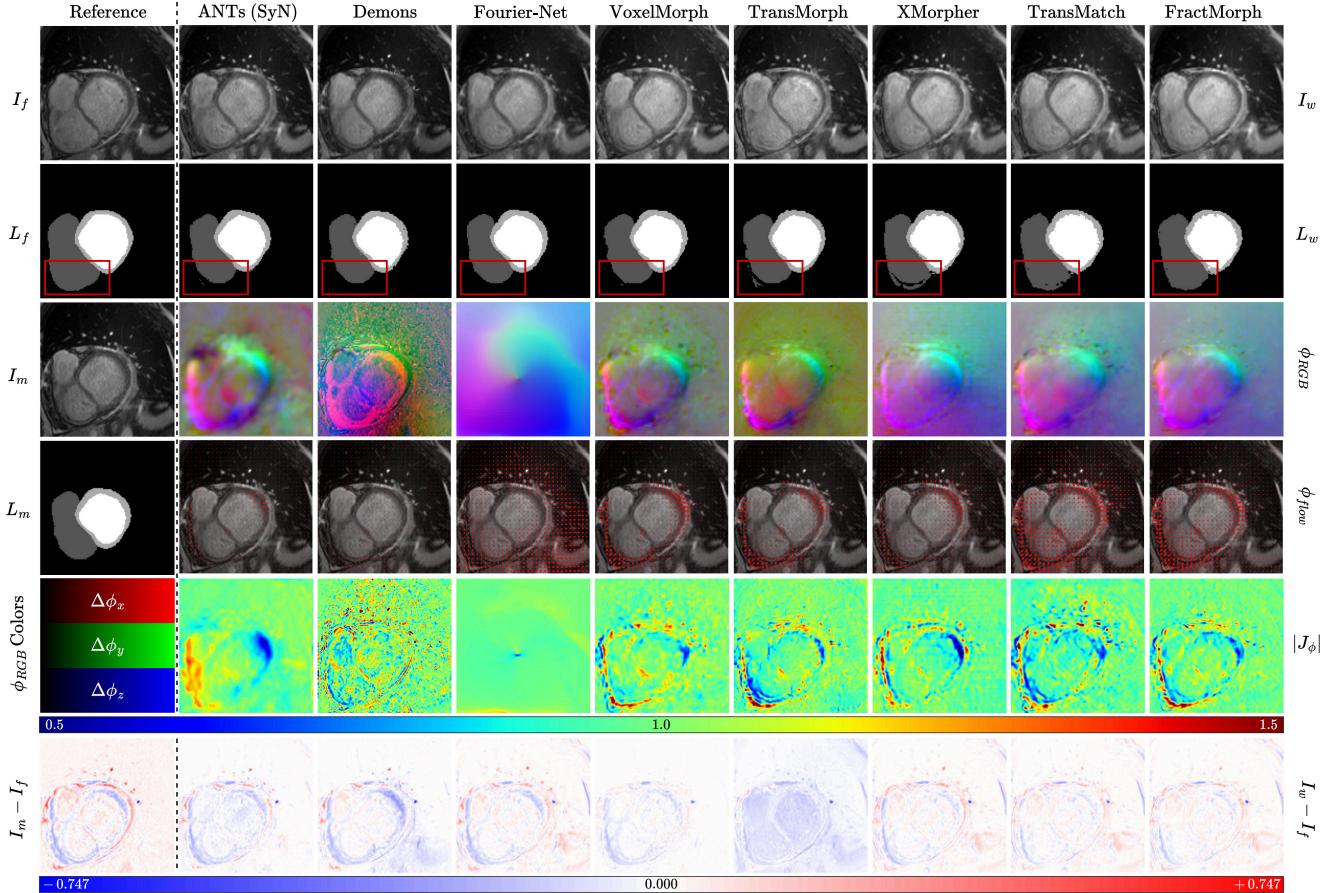


Figure 8: Qualitative comparison of registration methods on a case with semi-global deformations. This example illustrates the superior performance of our model in handling such deformations, as indicated by the red rectangles in the second row.

4.2. Evaluation Metrics

We evaluate the models using the Dice Similarity Coefficient (DSC) and the 95th-percentile Hausdorff Distance (HD95) to assess registration accuracy, as well as two additional metrics to measure the topological correctness of the deformation field. The standard deviation of each metric is also reported to demonstrate the stability of the models.

4.2.1. Dice Similarity Coefficient

The DSC is a region overlap measure used to quantify the similarity between two sets. Mathematically, for two binary sets X and Y , the Dice coefficient is defined as:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \quad (26)$$

where $|X|$ and $|Y|$ denote the number of voxels in the annotated segmentations X and Y , respectively. A higher DSC indicates greater overlap between the warped and fixed segmentations, and thus reflects more accurate registration.

4.2.2. Hausdorff Distance

The Hausdorff Distance (HD) is a boundary-based distance measure that evaluates the worst-case distance between two shapes. Given two sets of points, the HD is defined as the maximum of all nearest-neighbor distances from one

set to the other. Formally, for two point sets X and Y , the one-sided HD from X to Y is defined as:

$$hd(X, Y) = \max_{x \in X} \min_{y \in Y} d(x, y) \quad (27)$$

where $d(x, y) = \|x - y\|_2$. To ensure symmetry, we use the bidirectional HD, as the one-sided HD is not commutative:

$$HD(X, Y) = \max\{hd(X, Y), hd(Y, X)\} \quad (28)$$

The HD95 is defined as the 95th percentile of all point-to-point distances between two surfaces. By excluding the most extreme 5% of distances, HD95 reduces sensitivity to outliers. Therefore, we report HD95 as a robust boundary distance measure.

4.2.3. Percentage of Non-positive Jacobian

This metric evaluates the topological correctness of a deformation field by measuring the fraction of the spatial mapping that has a non-positive Jacobian determinant. The Jacobian matrix, defined as $J_\phi(p) = \nabla\phi(p)$, represents the local measurement of ϕ at voxel p . The Jacobian determinant describes the local volume change induced by the transformation at that point. To ensure a physically plausible deformation, it is required that $|J_\phi(p)| > 0$ everywhere, thereby preventing local inversions or foldings. If $|J_\phi(p)| \leq$

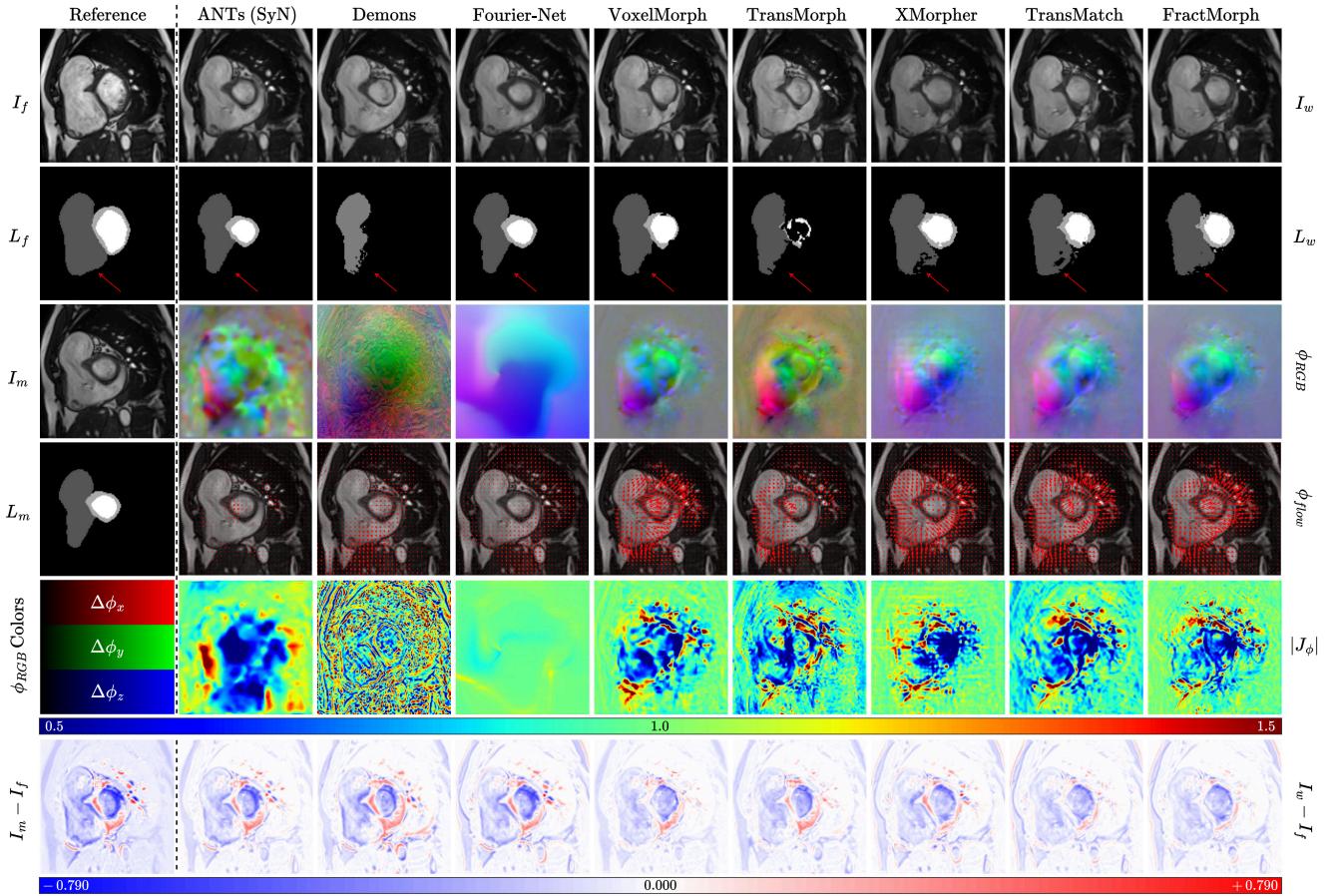


Figure 9: Qualitative comparison of registration methods on a case involving large-scale and global deformations. This example demonstrates the superior performance of our model in handling such deformations, which is clearly visible in the second row.

0 at any location, it indicates a fold at that voxel, meaning the mapping is not one-to-one in that neighborhood. Thus, a lower percentage reflects a deformation that better preserves topological consistency.

4.2.4. Standard Deviation of Jacobian Values

This metric measures the global variability of local volume changes caused by the deformation. It is computed as the standard deviation of the Jacobian determinant values across the entire image domain:

$$\sigma_J = \sqrt{\frac{1}{N} \sum_{p \in \Omega} (J_\phi(p) - \mu_J)^2} \quad (29)$$

A lower σ_J indicates a more uniformly smooth deformation, whereas a higher σ_J suggests that the deformation contains regions with more extreme local volume changes, resulting in a less smooth deformation field.

4.2.5. Model Size (Parameter Count)

This metric measures the total number of trainable parameters in the model. Parameter count serves as an indicator of model capacity and the on-disk storage required.

4.2.6. Inference Time

This metric measures the average time elapsed to complete a single forward pass on a pair of fixed and moving images.

4.2.7. Memory Usage

This metric measures the peak memory allocated during inference. We track RAM consumption for CPU-based methods and VRAM usage for GPU-based methods.

4.3. Implementation Details

We compare our framework with seven state-of-the-art registration methods, selected from traditional, non-transformer-based, and transformer-based approaches, to ensure a comprehensive evaluation. The learning-based models for the ACDC dataset were trained on an NVIDIA RTX 3080 GPU (12 GB VRAM), while those for the LPBA40 dataset, which required higher memory capacity, were trained on an RTX 4090 GPU (24 GB VRAM). To ensure fairness, evaluations of all methods were performed on the RTX 3080 GPU with identical software settings (CUDA 12.2, PyTorch 2.2.1). Traditional methods, for which GPU-accelerated versions are not yet fully developed, were executed on an 11th Gen Intel Core i7-1165G7 CPU with 16 GB of 3200 MHz RAM. For our framework, we use

Table 4Ablation study of the FCA module and FrFT branches on the ACDC dataset, with mean \pm std reported.

Branches				Results				
FrFT _{0°}	FrFT _{45°}	FrFT _{90°}	log(FrFT _{90°})	Overall DSC (%) ↑	Avg. DSC (%) ↑	HD95 (mm) ↓	% J _φ ≤ 0 ↓	STD(J _φ) ↓
✗	✓	✓	✓	86.24 ± 4.96	75.15 ± 9.09	1.56 ± 0.83	0.05 ± 0.06	0.15 ± 0.03
✓	✗	✓	✓	86.12 ± 4.97	74.66 ± 9.08	1.65 ± 0.88	0.06 ± 0.07	0.14 ± 0.03
✓	✓	✗	✓	86.11 ± 4.87	74.05 ± 9.26	1.68 ± 0.79	0.06 ± 0.06	0.15 ± 0.03
✓	✓	✗	✗	86.07 ± 5.10	74.66 ± 9.04	1.60 ± 0.81	0.05 ± 0.05	0.15 ± 0.03
✓	✓	✓	✗	86.37 ± 4.94	74.89 ± 9.35	1.57 ± 0.83	0.06 ± 0.06	0.14 ± 0.03
✓	✓	✓	✓	86.45 ± 4.72	75.15 ± 8.95	1.54 ± 0.78	0.05 ± 0.04	0.15 ± 0.03

a patch size of $P_x = P_y = P_z = 4$ and a patch embedding dimension of $d = 48$. All methods were trained for 400 epochs using the Adam optimizer with a batch size of 1. We employed the loss function described in Section 3.7, using $\lambda = 1$ across all learning models to ensure a fair comparison. By setting $\lambda = 1$, we equally weight the similarity term $L_{\text{similarity}}$ and the smoothness regularization L_{smooth} , balancing alignment accuracy against deformation smoothness and preventing either term’s scale from dominating. Keeping λ fixed also streamlines hyperparameter tuning and guarantees an unbiased, reproducible evaluation across all architectures.

4.4. Baseline Methods

In the following subsections, we briefly describe the seven state-of-the-art registration methods selected for comparison and the software packages used to run the traditional algorithms.

4.4.1. ANTs (SyN)

For the Symmetric Normalization (SyN) algorithm (Avants et al., 2008) with mutual information as optimization metric, we used its Python wrapper package, ANTsPy¹. We applied the default recommended hyperparameters of the SyN registration function, but increased the number of iterations to (100, 80, 60) across the multi-resolution levels to achieve improved alignment accuracy.

4.4.2. Demons

For the Demons algorithm (Thirion, 1998), we used the Python implementation provided by SimpleITK². The Demons Registration Filter was configured with its default hyperparameters, except that we increased the Gaussian smoothing standard deviation of the displacement field to 5.0 and the number of iterations to 200, to improve convergence and reduce spurious deformations.

4.4.3. VoxelMorph

For VoxelMorph (Balakrishnan et al., 2019), we used the official VoxelMorph-1 implementation³ and trained it using the suggested learning rate of 1e–4.

4.4.4. Fourier-Net

For Fourier-Net (Jia et al., 2023), we used the official 3D Fourier-Net implementation⁴ and trained it using the suggested learning rate of 1e–4.

4.4.5. TransMorph

For TransMorph (Chen et al., 2022), we employed the official implementation⁵. We trained the model with the recommended learning rate of 1e–4.

4.4.6. XMorpher

For XMorpher (Shi et al., 2022), we used the official implementation⁶ and trained it using the suggested learning rate of 1e–4.

4.4.7. TransMatch

For TransMatch (Chen et al., 2023), we employed the official implementation⁷. The model was trained using the recommended learning rate of 4e – 4.

4.5. Comparison to Baseline Methods

4.5.1. Registration Accuracy

As shown in Table 2, our proposed FractMorph framework achieved the highest performance across all registration accuracy metrics among the compared methods on the ACDC test set. FractMorph attained an overall DSC of 86.45% and an average per-structure DSC of 75.15%, which are marginally higher than the best baseline approaches. In particular, FractMorph improved the overlap of the cardiac structures compared to both traditional algorithms and prior learning-based models. The lightweight variant of our model (FractMorph-Light) performed nearly on par with the full model, achieving an overall DSC of 86.32% and an average DSC of 74.97%. This indicates that even with reduced channels, our multi-domain attention approach retains a clear accuracy advantage over previous methods. FractMorph and its lightweight variant also demonstrated superior boundary alignment accuracy, achieving the lowest HD95 values of 1.54 mm and 1.57 mm, respectively, thereby outperforming all baselines. In contrast, some competing methods showed larger HD95 values. For instance, TransMatch’s HD95 was

¹<https://github.com/ANTsX/ANTsPy>

²<https://github.com/SimpleITK/SimpleITK>

³<https://github.com/voxelmorph/voxelmorph>

⁴<https://github.com/xi-jia/Fourier-Net>

⁵https://github.com/junyuchen245/TransMorph_Transformer_for_Medical_Image_Registration

⁶<https://github.com/Soleemoon/XMorpher>

⁷https://github.com/tzayuan/TransMatch_TMI

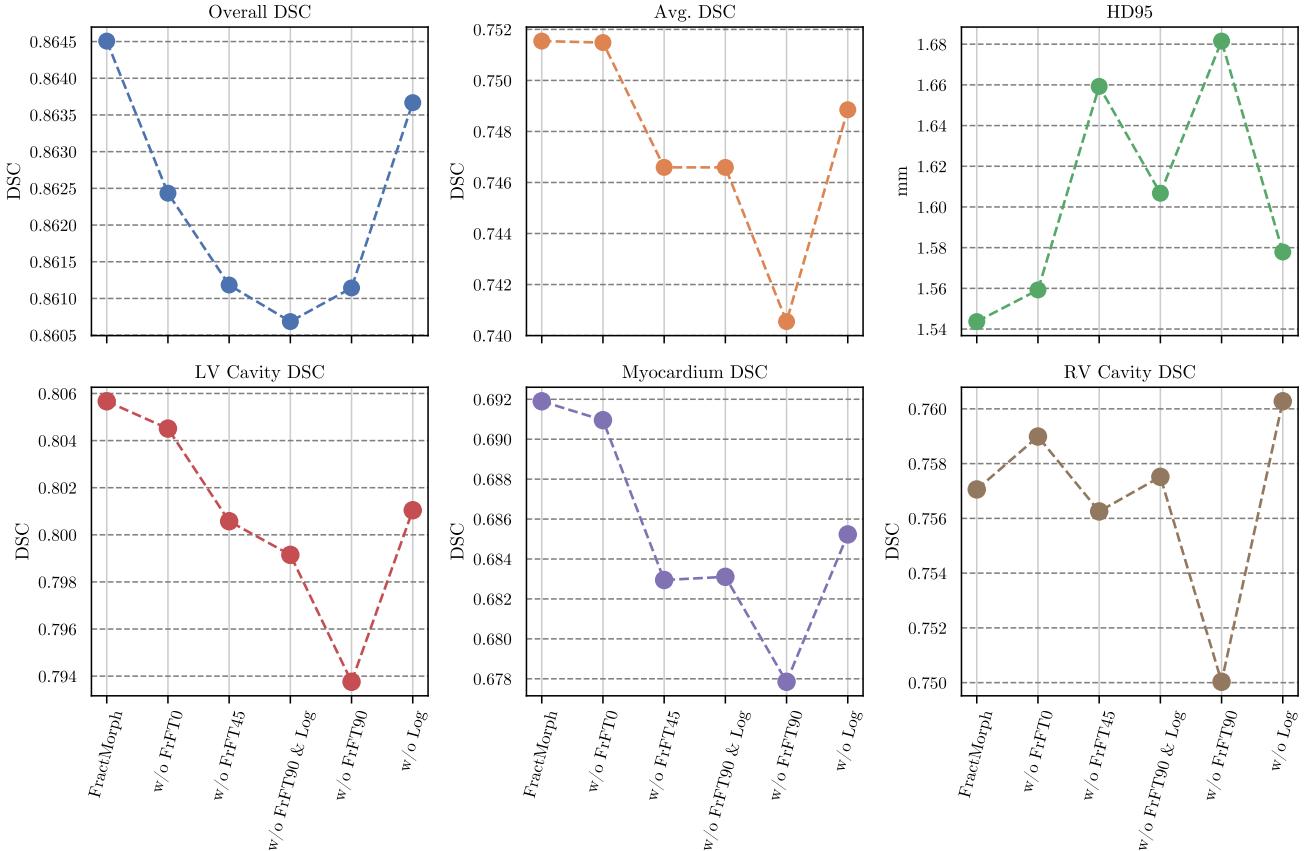


Figure 10: Overall and per-structure registration accuracy in the ablation study showing the contribution of each FrFT branch on the ACDC dataset.

3.21 mm, indicating less consistent alignment at the boundaries despite its high overall DSC.

The superior precision of our method is evident across all anatomical structures. Fig. 6 presents boxplots of DSC for the LV cavity, myocardium, and RV cavity across all methods. Both FractMorph and its lightweight variant (FractMorph-Light) achieve higher mean DSC in every structure compared to the baselines. Moreover, FractMorph exhibits slightly higher means and lower variances than FractMorph-Light, indicating greater stability. Our approach also preserves the topological correctness of deformations. The fraction of voxels with non-positive Jacobian determinants for FractMorph is 0.05%, which is comparable to the best deep learning-based baselines and significantly lower than that of traditional iterative methods. The Jacobian standard deviation for our method is 0.15, on par with other learning-based models. Notably, Fourier-Net achieved the lowest Jacobian variability (0.03) and zero foldings, but this came at the cost of substantially lower registration accuracy (83.09% DSC). In summary, FractMorph provides high alignment accuracy and physiologically plausible deformations, improving overlap across all structures while maintaining smooth and invertible transformations. Qualitative results further support these findings. Fig. 7 shows a representative case that reflects typical localized anatomical variability, demonstrating the overall registration quality achieved by our method. Fig. 8

includes a case with semi-global deformations, selected to illustrate the model’s ability to handle extended non-localized transformations that challenge baseline approaches. Fig. 9 highlights a case with large-scale and global deformations, representing scenarios where conventional models tend to break down. Together, these examples demonstrate the robustness and generalization capability of FractMorph across diverse deformation patterns without the need for tuning to specific deformation scenarios.

The results above were obtained on an intra-patient registration task. In contrast, the cerebral dataset used for atlas-to-patient registration features more localized deformations compared to cardiac datasets, where detailed and large-scale deformations occur simultaneously. As shown in Table 3, FractMorph achieved the best HD95 performance while maintaining substantially lower runtime compared to iterative models such as ATNs (SyN). Although SyN obtained the highest DSC scores, it required approximately 5 minutes per case, whereas our model achieved comparable Dice scores and superior HD95 in just 5 seconds. Notably, SyN also exhibited the highest number of non-positive Jacobians, indicating that its high DSC scores come at the cost of substantial folding and a lack of one-to-one transformations.

Table 5Ablation study of the FCA module skip connection on the ACDC dataset, with mean \pm std reported.

Skip Connection	Overall DSC (%) \uparrow	Avg. DSC (%) \uparrow	HD95 (mm) \downarrow	$\% J_\phi \leq 0 \downarrow$	$STD(J_\phi) \downarrow$
\times	85.45 ± 4.54	74.42 ± 9.14	1.65 ± 0.80	0.05 ± 0.07	0.14 ± 0.03
\checkmark	86.45 ± 4.72	75.15 ± 8.95	1.54 ± 0.78	0.05 ± 0.04	0.15 ± 0.03

4.5.2. Computational Efficiency

The gains in accuracy with FractMorph come with a moderate increase in model complexity and runtime (see Table 2). Our primary model contains approximately 63.9M parameters, which is larger than most CNN-based methods yet comparable to transformer-based models. The proposed FractMorph-Light variant significantly reduces the size to 29.7M parameters by using channel splitting (channel coefficient of $\alpha = 1/3$) in the FrFT feature extractor. This nearly halves the model size while maintaining superior accuracy. In terms of memory footprint, FractMorph requires approximately 461 MB during inference on a 3D volume pair. This is higher than simpler architectures such as Fourier-Net and VoxelMorph, but remains manageable on modern GPUs. Moreover, it is substantially lower than TransMatch, which requires around 1 GB, and also lower than larger models such as TransMorph. FractMorph-Light further lowers this consumption to 322 MB, which is also less than TransMorph’s and comparable to SyN, VoxelMorph, and XMorpher. Inference speed is also slightly affected by the richer multi-domain approach of our model. FractMorph registers a pair of 3D images in approximately 0.36 seconds. Although this is slightly slower than lightweight models such as VoxelMorph or Fourier-Net, it remains substantially faster than traditional iterative methods such as ANTs and Demons. Transformer-based baselines fall in between, with TransMatch requiring 0.38 seconds and XMorpher 0.46 seconds per case. Our lightweight variant, FractMorph-Light, provides a modest speed-up, completing registration in 0.34 seconds. In the higher-resolution dataset of LPBA40, our models maintain superior runtime performance compared to iterative methods and also consume less memory than VoxelMorph, TransMorph, and TransMatch.

4.6. Ablation Study

We conducted ablation studies to assess how removing key components affects the performance of our proposed model. First, we examine the contribution of each FrFT branch within the FCA module. Then, we demonstrate the necessity of the skip connection in the FCA module to achieve more effective registration.

4.6.1. FrFT Branches

To quantify each FrFT branch’s contribution, we disabled each branch in turn while keeping all other components unchanged (see Table 4). Removing any branch degrades accuracy, and the full model performs best overall. The branch FrFT0° has the smallest effect with a 0.08% drop in overall DSC and a 0.02 mm increase in HD95, while the branches FrFT45° and FrFT90° each cause over

0.33% declines in overall DSC, over 0.11 mm increases in HD95, and average DSC losses of 0.49% and 1.10%, respectively. Fig. 10 visualizes these trends across global metrics and per-structure DSCs, showing minimal impact from the FrFT0° and log branches and maximal impact from FrFT45° and FrFT90°. Notably, the log branch still improves every metric. Crucially, all variants preserve diffeomorphic regularity, as the percentage of non-positive Jacobian determinants remains between 0.05% and 0.06%, and the standard deviation of $|J_\phi|$ ranges only from 0.14 to 0.15. The full model achieves the lowest folding rate ($0.05\% \pm 0.04$) and a Jacobian-magnitude SD of 0.15 ± 0.03 , demonstrating that adding branches boosts accuracy without compromising deformation smoothness.

4.6.2. Skip Connection in FCA

To assess the impact of the skip connection in our FrFT feature extractor, we compared models with and without it (see Table 5). Adding the skip connection increases overall DSC by 1%, average DSC by 0.73%, and reduces HD95 by 0.11 mm. The rate of non-positive Jacobian determinants remains at 0.05%, while the standard deviation of the Jacobian rises only marginally. These results show that skip connections significantly boost accuracy while maintaining diffeomorphic regularity.

5. Discussion

FractMorph addresses a long-standing gap in deformable image registration: convolutional networks excel at modeling fine local details but lack global context, whereas standard transformers capture global structure at the expense of local precision and require large amounts of training data. To address these challenges, FractMorph brings together three key elements in a single end-to-end framework. First, we introduce a novel multi-domain FCA module that applies 3D FrFT branches at 0°, 45°, 90°, and a log-magnitude stream, enabling the network to extract local, semi-global, and global features in parallel. Second, we introduce a lightweight encoder-decoder CNN that transforms these transformer-enriched features into a high-resolution deformation field, preserving fine local details. Third, we adopt a dual-parallel transformer architecture to maintain continuous interaction between the fixed and moving image streams throughout feature extraction. Together, these components leverage their complementary strengths to achieve accurate and efficient deformable image registration.

Quantitatively (see Table 2), FractMorph achieves state-of-the-art accuracy on the ACDC cardiac cine-MRI benchmark, outperforming both classical algorithms and recent

deep learning baselines on every registration metric. This gain does come with a modest runtime cost, as the FrFT operations introduce a small computational overhead that slightly increases inference time. However, our model still runs significantly faster than traditional optimization-based methods. Furthermore, the total parameter count (63.9 M for FractMorph, 29.6 M for FractMorph-Light) and memory consumption remain comparable to or even lower than those of leading transformer architectures. The memory footprint is also halved in the lightweight variant, making both models practical for deployment on modern and resource-constrained GPUs. Qualitative examples in Figs. 7, 8, and 9 further illustrate FractMorph’s versatility. Whether the deformation is subtle and local or large and global, our single end-to-end network consistently produces smooth, topology-preserving warps without the need for scenario-specific tuning. This robustness stems directly from the multi-order FrFT branches, as each fractional angle captures features at a different spatial-frequency scale, as also explained in Section 3.

Our ablation study (see Table 4) confirms the importance of each FrFT branch. Removing either the 45° or the 90° branch reduces overall DSC by up to 0.34%, average per-structure DSC by up to 1.10%, and increases HD95 by up to 0.14 mm. In Fig. 10, we visualize each branch’s contribution and see that all branches improve performance. However, the log-magnitude branch (without the $\alpha = 90^\circ$ branch) can slightly degrade results, as the FrFT phase carries valuable information that the magnitude-only representation omits (see Section 3). Notably, adding the log-magnitude branch to the $\alpha = 90^\circ$ branch yields better performance than the configuration with the $\alpha = 90^\circ$ branch but without the log-magnitude branch, since the log transformation enables the model to better capture subtle details in the spectral magnitude domain. We also evaluated the residual skip connections in the FCA module and found that they further enhance feature enrichment and promote more stable gradient flow. Another notable exception is the RV cavity, where omitting the $\alpha = 0^\circ$ branch slightly increases DSC. In both the general population (Kawel-Boehm et al., 2025) and in our ACDC dataset, the RV cavity undergoes a large volume changes between end-diastole and end-systole and has a greater overall volume than the LV and myocardium on average. Omitting the $\alpha = 0^\circ$ branch slightly reduces DSC for the LV and myocardium, whereas it slightly increases DSC for the RV cavity. We hypothesize that deprioritizing pure spatial features enables the network to focus more on semi-global and global patterns, improving alignment for larger structures like the RV cavity while compromising fine-detail registration in smaller structures. Apart from the cardiac and intra-patient datasets, we also evaluated our method on cerebral and atlas-to-patient datasets to demonstrate its versatility across different modalities and registration tasks. As shown in the Experiments section, our model achieves significantly better performance on cases involving local-to-global and multi-scale deformations. Nonetheless, it also

performs well on other modalities with more localized deformations, making it a generally superior choice. Table 3 presents quantitative results supporting this claim. We note that the LPBA40 dataset contains very few training cases, which poses a challenge for learning-based methods to accurately model deformations. Additionally, iterative methods like SyN are typically optimized for cerebral cases, potentially introducing a bias toward cerebral tissue, as our cross-modality experiments also suggest.

In summary, we demonstrated and analyzed the experiments and ablation results, highlighting the effectiveness and generality of our approach and FCA modules, as well as the necessity of each FrFT branch in building an end-to-end model capable of handling various types and scales of deformation.

6. Conclusion

In this work, we have presented FractMorph, the first 3D transformer-based architecture that integrates multi-domain FrFT branches into cross-attention to capture local, semi-global, and global deformations simultaneously. This framework addresses the challenge of modeling deformations at multiple scales within a fully end-to-end deformable image registration network. Our comprehensive evaluation on the cardiac intra-patient ACDC dataset shows state-of-the-art registration accuracy and anatomically plausible and smooth deformations, with feasible inference time and memory consumption. To further demonstrate the versatility of our method across different modalities and registration tasks, we also achieved high-performing results on the cerebral atlas-to-patient LPBA40 dataset. Ablation studies confirm the indispensable roles of each FrFT branch and skip connections. The lightweight FractMorph-Light variant highlights a practical trade-off between resource use and performance. Future work will focus on accelerating the FrFT operation to further improve runtime efficiency. In addition, the proposed FCA module holds promise for broader application across diverse tasks in both medical and non-medical imaging.

Declaration of Competing Interest

The authors state that they have no financial or personal conflicts of interest that could have impacted the research presented in this manuscript.

Data Availability

The ACDC dataset used in this study is publicly available at <https://www.creatis.insa-lyon.fr/Challenge/acdc/>, and the LPBA40 dataset can be accessed at https://www.loni.usc.edu/research/atlas_downloads.

References

- Almeida, L.B., 1994. The fractional fourier transform and time-frequency representations. *IEEE Transactions on signal processing* 42, 3084–3091.
- Arratia López, P., Mella, H., Uribe, S., Hurtado, D.E., Sahli Costabal, F., 2023. Warppinn: Cine-mr image registration with physics-informed

- neural networks. *Medical Image Analysis* 89, 102925. URL: <https://www.sciencedirect.com/science/article/pii/S1361841523001858>, doi:<https://doi.org/10.1016/j.media.2023.102925>.
- Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis* 12, 26–41.
- Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging* 38, 1788–1800.
- Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Gonzalez Ballester, M.A., Sanroma, G., Napel, S., Petersen, S., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V.A., Krishnamurthi, G., Rohé, M.M., Pennec, X., Sermesant, M., Isensee, F., Jäger, P., Maier-Hein, K.H., Full, P.M., Wolf, I., Engelhardt, S., Baumgartner, C.F., Koch, L.M., Wolterink, J.M., Isgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O., Jodoin, P.M., 2018. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging* 37, 2514–2525. doi:[10.1109/TMI.2018.2837502](https://doi.org/10.1109/TMI.2018.2837502).
- Bistoque, A., Oshinski, J., Škrinjar, O., 2008. Myocardial deformation recovery from cine mri using a nearly incompressible biventricular model. *Medical image analysis* 12, 69–85.
- Bracewell, R., Kahn, P.B., 1966. The fourier transform and its applications. *American Journal of Physics* 34, 712–712.
- Cahill, N.D., Noble, J.A., Hawkes, D.J., 2009. A demons algorithm for image registration with locally adaptive regularization, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 574–581.
- Candan, C., Kutay, M.A., Ozaktas, H.M., 2000. The discrete fractional fourier transform. *IEEE Transactions on signal processing* 48, 1329–1337.
- Chang, Q., Wang, Y., Zhang, J., 2024. Independently trained multi-scale registration network based on image pyramid. *Journal of Imaging Informatics in Medicine* 37, 1557–1566.
- Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y., 2022. Transmorph: Transformer for unsupervised medical image registration. *Medical image analysis* 82, 102615.
- Chen, J., He, Y., Frey, E.C., Li, Y., Du, Y., 2021. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. *arXiv preprint arXiv:2104.06468*.
- Chen, Z., Zheng, Y., Gee, J.C., 2023. Transmatch: a transformer-based multilevel dual-stream feature matching network for unsupervised deformable image registration. *IEEE transactions on medical imaging* 43, 15–27.
- Gonzalez, R.C., 2009. Digital image processing. Pearson education india.
- Huang, B., Yang, F., Yin, M., Mo, X., Zhong, C., 2020. A review of multimodal medical image fusion techniques. *Computational and mathematical methods in medicine* 2020, 8279342.
- Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks. *Advances in neural information processing systems* 28.
- Jia, X., Bartlett, J., Chen, W., Song, S., Zhang, T., Cheng, X., Lu, W., Qiu, Z., Duan, J., 2023. Fourier-net: Fast image registration with band-limited deformation, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1015–1023.
- Jiang, J., Luk, W., Rueckert, D., 2003. Fpga-based computation of free-form deformations in medical image registration, in: Proceedings. 2003 IEEE International Conference on Field-Programmable Technology (FPT)(IEEE Cat. No. 03EX798), IEEE. pp. 234–241.
- Kawel-Boehm, N., Hetzel, S.J., Ambale-Venkatesh, B., Captur, G., Chin, C.W., Francois, C.J., Jerosch-Herold, M., Luu, J.M., Raisi-Estabragh, Z., Starekova, J., et al., 2025. Society for cardiovascular magnetic resonance reference values (“normal values”) in cardiovascular magnetic resonance: 2025 update. *Journal of Cardiovascular Magnetic Resonance* 27, 101853.
- Kim, B., Kim, D.H., Park, S.H., Kim, J., Lee, J.G., Ye, J.C., 2021. Cy-clemorph: cycle consistent unsupervised deformable image registration. *Medical image analysis* 71, 102036.
- Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., et al., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration. *Neuroimage* 46, 786–802.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012–10022.
- Ozaktas, H.M., Arikan, O., Kutay, M.A., Bozdag, G., 1996. Digital computation of the fractional fourier transform. *IEEE Transactions on signal processing* 44, 2141–2150.
- Pei, S.C., Yeh, M.H., 1997. Improved discrete fractional fourier transform. *Optics letters* 22, 1047–1049.
- Pei, S.C., Yeh, M.H., 1998. Two dimensional discrete fractional fourier transform. *Signal Processing* 67, 99–108.
- Pfeffer, M.A., Shah, A.M., Borlaug, B.A., 2019. Heart failure with preserved ejection fraction in perspective. *Circulation research* 124, 1598–1617.
- Pluim, J., Fitzpatrick, J., 2003. Image registration. *IEEE Transactions on Medical Imaging* 22, 1341–1343. doi:[10.1109/TMI.2003.819272](https://doi.org/10.1109/TMI.2003.819272).
- Ramadan, H., El Bourakadi, D., Yahyaouy, A., Tairi, H., 2024. Medical image registration in the era of transformers: a recent review. *Informatics in Medicine Unlocked*, 101540.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18, Springer. pp. 234–241.
- Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J., 2002. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging* 18, 712–721.
- Sahin, A., Ozaktas, H.M., Mendlovic, D., 1995. Optical implementation of the two-dimensional fractional fourier transform with different orders in the two dimensions. *Optics Communications* 120, 134–138.
- Shattuck, D.W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salomon, G., Narr, K.L., Poldrack, R.A., Bilder, R.M., Toga, A.W., 2008. Construction of a 3d probabilistic atlas of human cortical structures. *Neuroimage* 39, 1064–1080.
- Shi, J., He, Y., Kong, Y., Coatrieux, J.L., Shu, H., Yang, G., 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, Springer. pp. 217–226.
- Song, X., Chao, H., Xu, X., Guo, H., Xu, S., Turkbey, B., Wood, B.J., Sanford, T., Wang, G., Yan, P., 2022. Cross-modal attention for multimodal image registration. *Medical Image Analysis* 82, 102612.
- Subramanian, S.R., Hon, T.K., Georgakis, A., Papadakis, G., 2010. Fractional fourier-based filter for denoising elastograms, in: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, IEEE. pp. 4028–4031.
- Thirion, J.P., 1998. Image matching as a diffusion process: an analogy with maxwell’s demons. *Medical image analysis* 2, 243–260.
- Yang, C., 2012. Applications of the fractional fourier transform.
- Yu, H., Huang, J., Li, L., Zhao, F., et al., 2023. Deep fractional fourier transform. *Advances in Neural Information Processing Systems* 36, 72761–72773.
- Zeng, W., Gao, H., 2015. Image processing research based on fractional fourier transform. *Journal of Software Engineering* 9, 318–327.
- Zhang, X., Shen, Y., Li, S., Zhang, H., 2013. Medical image registration in fractional fourier transform domain. *Optik* 124, 1239–1242.
- Zhuo, Y., Shen, Y., 2024. Diffusereg: Denoising diffusion model for obtaining deformation fields in unsupervised deformable image registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 597–607.

- Zitová, B., Flusser, J., 2003. Image registration methods: a survey. *Image and Vision Computing* 21, 977–1000. URL: <https://www.sciencedirect.com/science/article/pii/S0262885603001379>, doi:[https://doi.org/10.1016/S0262-8856\(03\)00137-9](https://doi.org/10.1016/S0262-8856(03)00137-9).
- Zou, J., Gao, B., Song, Y., Qin, J., 2022. A review of deep learning-based deformable medical image registration. *Frontiers in Oncology* 12, 1047215.