IET Image Processing

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

Medical Image Registration via Spatial Feature Extraction Mamba and Substrate Iterative Refinement

Zilong Xue | Kangjian He | Dan Xu | Jian Gong

School of Information Science and Engineering, Yunnan University, Kunming, Yunnan, China

Correspondence: Kangjian He (hekj@ynu.edu.cn)

Received: 3 January 2025 | **Revised:** 7 May 2025 | **Accepted:** 19 May 2025

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 62202416, 62462066, 62162068, in part by the Yunnan Fundamental Research Projects under Grant 202401AU070204, 202501AT070228, in part by the Yunnan Provincial Science and Technology Department-Yunnan University Double First Class Construction Joint Fund Project under Grant No. 202301BF070001-025.

ABSTRACT

One of the major challenges in medical image registration is balancing computational efficiency with the ability to capture large deformations in complex anatomical structures. Existing methods often struggle with high computational costs due to the need for extensive feature extraction and attention computations at various levels of the network. Moreover, some methods do not take into account the spatial relationships of the feature images during registration, and the loss of these spatial relationships leads to suboptimal results for these methods. To this end, we introduce a novel medical image registration network, PSMamba-Net, which leverages optimized iteration and the Mamba framework within a dual-stream pyramid architecture. The network reduces the computational burden by narrowing attention computations at each decoding level, while an optimized iterative registration module at the bottom of the pyramid captures large deformations. This approach eliminates the need for repeated feature extraction, significantly accelerating the registration process. Additionally, the SMB module is incorporated as a decoder to enhance spatial relationship modelling and leverage Mamba's strengths in long-sequence processing. PSMamba-Net balances efficiency and accuracy, surpassing state-of-the-art methods across LPBA40, Mindboggle, and Abdomen CT datasets. Our source code is available at: https://github.com/VCMHE/PSMamba.

1 | Introduction

Correspondences between images through reliable image registration is essential for various clinical tasks, including image fusion, organ mapping, and monitoring tumour growth, and presents a significant challenge [1]. Deformable image registration can be understood as finding a dense nonlinear spatial correspondence (transformation) between a fixed image and a moving image [2]. Through spatial transformations, the moving image can be warped to align with the fixed image. Conventional deformable registration techniques, including [3–6], often treat deformable registration as a complicated optimization challenge. Although these methods are widely used and have a strong mathematical foundation, they require iterative optimization for each pair of new images [7]. This implies that these traditional methods often require significant computational resources to densely assess voxel-level similarities [8]. Moreover, the limited amount of medical training data and the lack of ground truth information both restrict the development of traditional registration methods [9].

In recent years, deep learning methods have been greatly developed in the field of medical image registration. For example, Balakrishnan et al. proposed VoxelMorph [10], which utilizes a five-layer U-Net [11] with three additional convolutional layers at the end. During training, the network's performance is enhanced

by evaluating the similarity between the reference image and the deformed image. VoxelMorph achieves comparable accuracy to state-of-the-art traditional methods, such as [4], while being several orders of magnitude faster in operational speed. Inspired by its success, many subsequent [12–14] registration processes have adopted U-Net-style architectures as the backbone for their registration networks. Stronger constraints have been introduced on the deformation field to ensure better registration results. For instance, CycleMorph [15] enhances image registration performance by providing implicit regularization to maintain topology during deformation, while SynthMorph [16] eliminates reliance on training data by leveraging various synthetic label mappings and image generation strategies. Furthermore, some methods optimize the loss function to preserve specific characteristics of the predicted deformation field. For example, FAIM [17] introduces a penalty for negative Jacobian determinant areas in its loss function to reduce their occurrence. The diffeomorphic variant of VoxelMorph, Voxelmorph-diff [18], employs a probabilistic generative model and a velocity-based transformation representation to maintain topological consistency. TM-TVF [19] draws on the concept of time-invariant velocity fields, utilizing a cyclic structure at the network's end to generate a series of time-invariant velocity fields, with additional constraints imposed. However, these modifications significantly increase the number of network parameters and computational complexity, potentially hindering the overall efficiency of both training and inference processes.

Drawing inspiration from the strengths of transformers [20] in natural language processing (NLP), recent studies have adapted transformers for computer vision tasks, achieving results that often exceed those of convolutional neural networks (CNNs) across various applications. Many transformer-based registration methods have also been proposed, such as TransMorph [21], which utilizes a swin transformer in its encoder to capture long-range dependencies, thereby improving registration accuracy. Due to the transformer's ability to densely route information within contextual windows, allowing it to model complex data effectively. However, this feature also presents a fundamental drawback: it cannot model any content outside of a limited window, which restricts its capability when dealing with long sequences. To address this issue, AttentionReg [22] and XMorpher [23] have improved their strategy by adopting a cross-attention mechanism to enhance volumetric representations. Although [22, 23] perform exceptionally well in registration tasks, their computational cost is relatively high.

Although transformers enhance the ability to capture long-range dependencies, their computational cost is often substantial due to the quadratic relationship of the self-attention mechanism with input size, especially for high-resolution medical images. Recently, a model named Mamba [24], based on the state space model (SSM) [25, 26] framework, has been proposed to address this issue. Mamba introduces a selection mechanism within the SSM to compress the context into a smaller state and combines it with hardware-aware algorithms to model long-range dependencies, significantly improving training and inference efficiency. Some existing studies, such as [27, 28], have already applied Mamba in the field of medical image registration, achieving promising results. However, current Mamba-based registration methods typically involve directly transferring the Mamba architecture to 3D image registration tasks. Mamba requires modelling features by flattening 3D characteristics into 1D sequences [29], which may lead to the loss of some spatial information in 3D medical images and result in the final warped image losing some semantic information.

To tackle the issues mentioned above, we propose a novel architecture called PSMamba-Net, which integrates a dual-stream pyramid structure with Mamba for 3D medical image registration. First, we utilize a dual-stream pyramid registration structure, which effectively reduces the scope of attention calculations required at each decoding level, thereby mitigating computational costs. Next, to minimize computational consumption, we introduce the substrate iterative refinement framework only at the bottom layer of the feature pyramid with the smallest feature image size to capture large deformations. Unlike other iterative registration methods, the advantage of iterating solely at the bottom layer of the pyramid is that it eliminates the need to re-extract image features, significantly accelerating both model training and deformation field prediction, while better addressing the challenges associated with large deformations. Furthermore, to effectively model spatial features, we replace the standard 3D convolutions in the original decoder with an enhanced SMB that fully extracts spatial features, thereby improving the representation of features in the spatial dimension. Our main contributions can be summarized as follows:

- A spatial feature extraction module (SMB) is proposed to effectively model spatial features. Using a gate-like mechanism, SMB extracts spatial information from feature maps for subsequent modelling, retaining Mamba's ability to handle long sequences while better leveraging spatial relationships in the feature maps.
- An optimization iteration module is proposed, performing iterations only at the lowest level of the feature pyramid to reduce memory consumption from redundant feature extraction. Ablation experiments show that combining this module with the SMB decoder significantly improves performance compared to other decoders.
- We propose PSMamba-Net, an end-to-end unsupervised deformable image registration model that uses an SMB module for spatial feature extraction and an optimization iteration module to reduce computational overhead. Experiments on two 3D brain datasets demonstrate that PSMamba-Net surpasses state-of-the-art methods in both accuracy and efficiency.

The structure of this paper is arranged as follows: Section 2 will discuss the related research progress on pyramid registration and iterative registration methods; Section 3 will provide a detailed description of our proposed method; Section 4 will present and analyse the experimental results; finally, Section 5 will summarize the conclusions of the study and outline future directions.

2 | Related Work

2.1 | Pyramid Decoder Architecture

In recent years, pyramid registration methods have rapidly developed alongside advancements in deep learning technology, particularly in the fields of medical image processing and remote sensing image analysis. The unsupervised learning strategy based on pyramid structures has gained attention due to its effective reduction in reliance on real deformation annotations. Notably, pyramid registration networks have excelled in handling deformation tasks between images, especially in capturing features at different scales. Compared to previous single-level networks, the pyramid registration network can progressively refine the deformation field through multi-level feature extraction, allowing it to better handle complex deformations [30]. Building on this foundation, subsequent researchers have proposed various extension techniques to enhance registration accuracy and tackle the challenges of complex medical image registration. For example, the multi-level variational image registration network (mlVIRNET) [12] constructs an image pyramid to gradually obtain the next-level deformation field from a coarse levels, demonstrating efficacy in handling large deformations.

The Laplacian pyramid registration network (LapIRN) [31] further improves this structure by merging feature maps from the coarse decoder into finer levels, enhancing the receptive field. However, such methods often require repeated feature extraction, complicating the training process. The dual-stream pyramid registration network (Dual-PRNet) [32] employs a dual-stream U-Net architecture with shared parameters, utilizing stacked fixed and distorted feature mappings to estimate registration fields from coarse to fine, thus improving feature extraction efficiency. Additionally, PRnet++ [9] introduces 3D correlation layers and residual connections based on the original PR module, significantly enhancing the model's relevance estimation. Nonetheless, directly using the deformation field from the current layer may lead to semantic ambiguity. Im2Grid [33] addresses the pixel correspondence in feature maps through a cross-attention structure, but its computational intensity may affect efficiency in practical applications.

2.2 | Traditional Iterative Optimization Based Registration

Iterative optimization registration is a widely used method in medical image processing, particularly suitable for 3D image registration. In this approach, the similarity between the fixed image and the deformed image is maximized through an iterative process to optimize the deformation model. These methods typically involve nonlinear deformation models, such as optical flow, B-spline constraints based on free-form deformation (FFD), and large deformation diffeomorphic metric mapping (LDDMM) based on time-independent velocity fields [34]. Over the past few decades, researchers have continuously proposed methods to enhance the efficiency of iterative optimization registration, especially when dealing with 3D medical images, where each pair of images requires extensive computation, posing challenges for real-time applications. This challenge is particularly pronounced in the field of diffeomorphic image registration.

To accelerate the registration process, some studies have proposed algorithms that scale and smooth the static velocity field (SVF) [35, 36], such as the DARTEL and Demons methods. Additionally, Zhang and Fletcher introduced a Fourier approximation method based on Lie algebra (Flash) [37], which speeds up the computation of differential equations by calculating finite-band frequencies in the Fourier domain for velocity field deformation estimation. Hernández further improved the Stokes-LDDMM [38] variational problem by introducing GPU parallelization, representing it as a bounded non-stationary vector field, thus further accelerating processing. Despite the widespread application of iterative optimization registration in medical image processing, its limitations in processing speed and computational complexity remain pressing issues to be addressed. To tackle these challenges, researchers have begun to explore more efficient registration methods. For instance, deep learning technologies are gradually being introduced into the field of image registration, allowing for significant improvements in registration speed and reductions in computational burden through learned feature extraction and deformation mapping. Additionally, advanced network architectures such as convolutional neural networks (CNNs) and generative adversarial networks (GANs) have been applied to image registration tasks, leveraging their strong feature representation capabilities to enhance registration accuracy. These new methods employ end-to-end training approaches to automatically learn the similarities between images and generate precise deformation fields, greatly reducing the reliance on manual feature selection and complex model design. Overall, as deep learning and other advanced technologies continue to evolve, finding a way to better integrate traditional iterative concepts with deep learning and achieving a balance between accuracy and speed remains an urgent problem to solve.

3 | Method

3.1 | Network Overview

In this section, we will introduce our proposed PSMamba-Net. As shown in the Figure 1, we adopt a dual-stream pyramid registration structure, which alleviates computational consumption by reducing the scope of attention calculations required at each decoding level. At the bottom of the pyramid network, we introduce a substrate iterative refinement module to capture large deformations. Additionally, we incorporate the SMB to reduce computational demands and enhance Mamba's spatial feature extraction capabilities.

Specifically, the dual-stream weight-sharing feature encoder takes the moving image I_m and the fixed image I_f as inputs, using five layers of convolutional blocks to extract hierarchical features, resulting in two sets of feature maps $\{F_1, F_2, F_3, F_4, F_5\}$and $\{M_1, M_2, M_3, M_4, M_5\}$. Initially, F_5 and M_5 are fed into the Substrate Iterative Refinement module to obtain the initial deformation field ϕ_1. Then, using ϕ_1 to deform M_4, we combine it with F_4 and send it to the SMB module to produce φ_2. After merging φ_2 with the previously obtained initial deformation field ϕ_1, we acquire the deformation field ϕ_2. By following similar operations, we can obtain ϕ_3, ϕ_4, and the final deformation field ϕ. Finally, the moving image I_m is deformed using the final deformation field ϕ to obtain the registered image. To guide the network

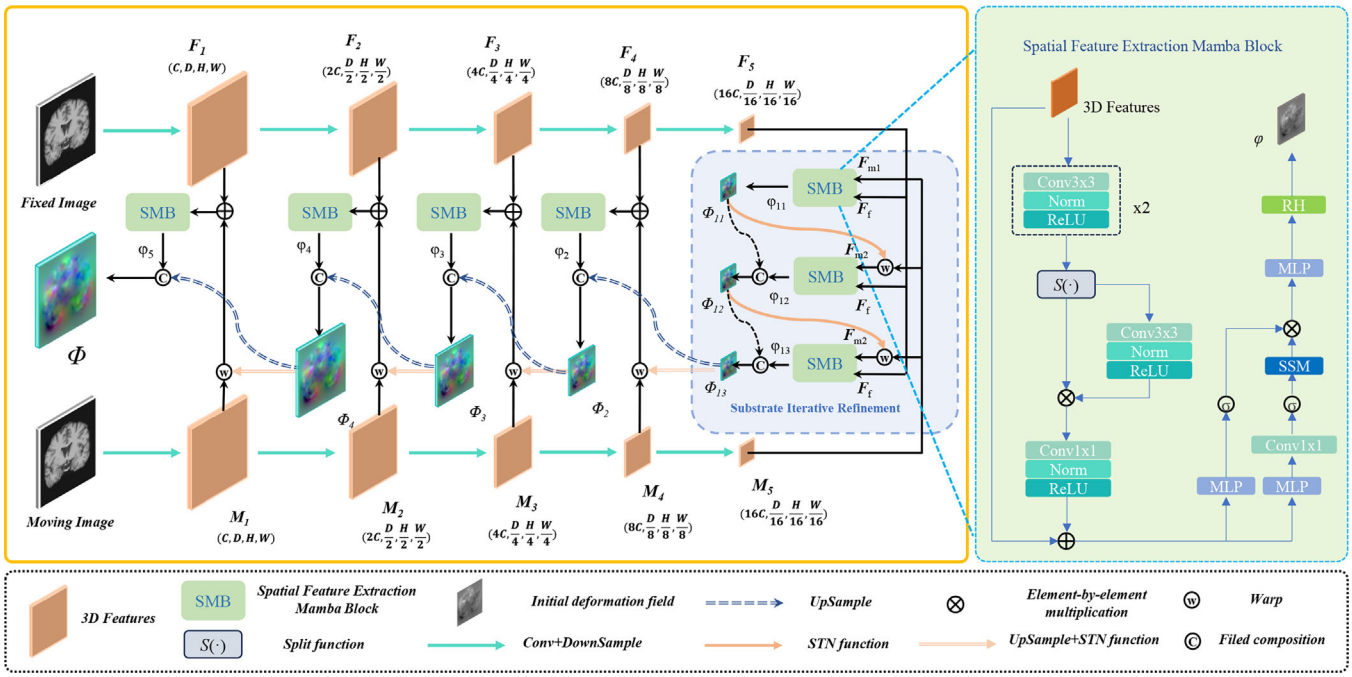

FIGURE 1 | The proposed PSMamba-Net architecture is illustrated above. It features a dual-stream pyramid framework designed to extract features from coarse to fine levels. At the base of the pyramid, we employ a bottom-up iterative approach for repeated feature extraction. The decoder utilizes our proposed SMB module to effectively capture the spatial relationships among the features.

training, we mimic VoxelMorph by using normalized cross-correlation $\mathcal{L}_{\text{ncc}}$ [39] to evaluate image similarity and employ deformation regularization $\mathcal{L}_{\text{reg}}$ to regularize the smoothness of the deformation field. Thus, our loss function is defined as:

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{ncc}}(I_{\text{f}}, I_{\text{m}} \circ \phi) + \lambda \mathcal{L}_{\text{reg}}(\phi) \tag{1}$$

where $\circ$ represents the warping operation, and λ is the weight of the regularization term.

3.2 | Spatial Feature Extraction Mamba Block

To better utilize the spatial features inherent in the feature maps, we propose the SMB (spatial feature extraction Mamba block) module to extract and preserve the spatial relationships within the feature maps. For preliminary spatial feature extraction, the input features first pass through two consecutive spatial convolution blocks. The specific workflow is illustrated in the figure. First, a 3D convolution is applied to the input data to obtain the convolved feature map. Next, instance normalization is performed on the convolution output. Finally, the ReLU activation function is applied to maintain non-linearity in the output. We can represent this process as:

$$C_{\text{sfe}}(F) = \max\left(0, \frac{C^{\kappa}(F) - \mu(C^{\kappa}(F))}{\sqrt{\sigma(C^{\kappa}(F)) + \epsilon}} \cdot \gamma + \beta\right) \tag{2}$$

where F is the input 3D feature map; C represents the convolution block; $\mu(\cdot)$ and $\sigma(\cdot)$ are the mean and variance for each channel; γ and β are learnable scaling and shifting parameters; ϵ is a small constant that prevents division by zero; κ is the size of the convolution kernel, and the kernel size used here is $3 \times 3 \times 3$.

After the preliminary extraction of spatial features, the segmentation function $S(\cdot)$ is used to divide the initially extracted features into two independent parts along the channel dimension. One part is then fed back into the spatial convolution to enhance the effectiveness of spatial feature extraction. This part is multiplied element-wise with the other part. Finally, we use convolution to further fuse the features and apply a residual connection to reuse the features, as represented by the following formula:

$$(F_u, F_v) = S(C_{\text{sfe}}(F)) \tag{3}$$

$$F_{\text{sfe}} = F + C^{\kappa}(F_u \cdot C_{\text{sfe}}(F_v)) \tag{4}$$

where F_u and F_v are two independent parts divided along the channel dimension of the input features; $\cdot$ denotes element-wise multiplication. Notably, the size of the convolution kernel κ is $1 \times 1 \times 1$.

Compared to transformer, Mamba is better at handling relatively long sequences with linear complexity. Therefore, after extracting spatial features, we model them using Mamba blocks and then obtain the initial deformation field ϕ through the RH module. To facilitate the computation of subsequent Mamba modules, we first flatten the feature map that has been integrated with spatial features into a 1D sequence. Then, we sequentially pass it through MLP linear projection layers, 1D convolutional layers, and SiLU/Swish layers. Finally, the tensor is processed in the SSM layer. The state space model (SSM) is commonly used to describe linear time-invariant systems. It generates an output $h_t \in R^N$ from a one-dimensional input sequence through an intermediate hidden state $y_t \in R$. This can be represented using linear ordinary differential equations as follows:

$$h'(t) = Ah(t) + Bx(t) \tag{5}$$

$$y(t) = Ch(t) + Dx(t) \tag{6}$$

where $x(t) \in R$ is a one-dimensional sequence obtained by flattening the feature map; $A \in R^{N\times N}$ is the transition matrix; $B, C \in R^{N\times 1}$ are projection parameters; $D \in R$ is for skip connections. Given the mismatch between the continuous-time characteristics of SSM and the form of discrete data in deep learning, it is necessary to discretize the ordinary differential equation through a discretization process. Referring to VMamba, the discrete version of the linear model can be obtained using zero-order hold rules with the given time scale parameter $\Delta \in R^D$, and its discretized form is as follows:

$$h_k = \bar{A}h_{k-1} + \bar{B}x_k \tag{7}$$

$$y(t) = \bar{C}h_k + Dx_k \tag{8}$$

$$\bar{A} = e^{\Delta A} \tag{9}$$

$$\bar{B} = (\Delta A)^{-1}(e^{\Delta A} - I)\Delta B \tag{10}$$

$$\bar{C} = C \tag{11}$$

where$B, C \in R^D$, and I is the identity matrix. Currently, Mamba-based registration methods typically directly transfer the Mamba structure to 3D image registration tasks. However, due to the initial design of Mamba, it requires flattening 3D features into a 1D sequence to model the features. In contrast, the SMB we propose utilizes a spatial extraction module with a gated mechanism to obtain spatial features and applies these to subsequent modelling. Additionally, we not only use the SMB as a decoder in the substrate iterative refinement but also employ SMB on higher-scale larger feature maps to acquire feature information, thereby compensating for the lack of detailed information when using only low-scale features.

3.3 | Substrate Iterative Refinement

For other iterative registration methods, iterations typically occur at different scales, and each iteration may require re-extracting image features, which is undoubtedly very computationally intensive in terms of both processing time and memory. Therefore, we propose an iterative registration module based on SMB blocks at the lowest level. In this module, multiple iterations are performed only on the last layer of the feature pyramid, effectively avoiding the computational costs associated with repeatedly extracting image features. The SMB blocks are used to optimize the precision and speed of iterative registration. The specific process is illustrated in the substrate iterative Refinement module in Figure 1. In the kth iteration, the feature images F_{mk} and F_{f} are fed into the SMB block to generate the temporary deformation field φ_{1k}. Expressed in formula form as:

$$\varphi_{1k} = Mamba(F_{sfe}(F_f + F_{mk})) \tag{12}$$

where $Mamba$ denotes the spatial modelling operation performed by the Mamba module, k represents the current iteration step.

Subsequently, the newly generated temporary deformation field φ_{1k} is used to refine the residual deformation field from the previous iteration Φ_{1k-1}, resulting in an updated residual deformation field Φ_{1k}. This process can be represented by the following formula:

$$\Phi_{1k} = \begin{cases} \varphi_{11}, & k = 1 \\ \Phi_{1k-1} + \varphi_{1k-1}, & k = 2, \ldots, K \end{cases} \tag{13}$$

where + denotes element-wise summation of deformation fields, k represents the current iteration step, and K is the maximum number of iterations. In our experiments, we set K= 3.

Unlike other iterative registration methods, in the proposed substrate iterative refinement method, multiple iterations are performed only at the lowest level of the feature pyramid. This means there is no need to re-extract image features. Instead, our specially designed SMB module is used for spatial feature extraction, combining the advantages of Mamba for computing long sequences while simultaneously reducing the algorithm's computational complexity and time consumption, thus accelerating the speed of model training and deformation field prediction.

4 | Experiments

4.1 | Datasets

Our experiments were conducted on three publicly available medical imaging datasets: LPBA40, Mindboggle, and abdomen CT–CT. Among them, LPBA40 and Mindboggle are brain MRI datasets, while abdomen CT–CT is an abdominal CT dataset.

For the LPBA40 dataset [40], each brain MRI volume contains 54 manually annotated regions of interest (ROIs). All volumes were centrally cropped and resampled to a uniform voxel resolution of 160 × 192 × 160 (1 mm × 1 mm × 1 mm). A total of 30 volumes (30 × 29 pairs) were used for training, and 10 volumes (10 × 9 pairs) were used for testing.

For the Mindboggle dataset [41], each volume contains 62 manually labelled ROIs. All volumes are pre-aligned to the MNI152 template. Preprocessing steps included min-max normalization and skull stripping performed using FreeSurfer [42]. After central cropping, all volumes were resized to 160 × 192 × 160 (1 mm × 1 mm × 1 mm). We used 42 volumes (42 × 41 pairs) from the NKI-RS-22 and NKI-TRT-20 subsets for training, and 20 volumes (20 × 19 pairs) from the OASIS-TRT-20 subset for testing.

Additionally, we introduced an abdominal CT (abdomen CT–CT) dataset, which contains 13 abdominal organs labelled as ROIs, including the spleen, left and right kidneys, gallbladder, oesophagus, liver, stomach, aorta, inferior vena cava, portal vein, splenic vein, pancreas, and left and right adrenal glands. All volumes were centrally cropped and resized to 160 × 160 × 192 (2 mm × 2 mm × 2 mm). The dataset consists of 30 volumes, with 24 used for training and 6 for testing.

4.2 | Comparison Methods

We conducted a comprehensive comparison between our proposed method and several state-of-the-art learning-based medical image registration networks, encompassing a variety of architectural paradigms, including different transformer and Mamba structures as well as pyramid-based designs.

TransMorph: A single-stage registration network based on swin transformer, which employs a hierarchical encoder to extract multi-scale features and incorporates self-attention mechanisms to improve registration accuracy.

Im2Grid (I2G): A pyramid registration network that integrates coordinate Transformers to explicitly model the structure of displacement fields, enhancing smoothness and structural consistency of registration results.

PR++: A pyramid registration network that integrates 3D correlation convolutions to more effectively capture multi-scale spatial deformations and improve adaptation to complex non-rigid transformations.

ModeT: A motion decomposition transformer framework that innovatively transforms multi-head neighbourhood attention into multi-coordinate modelling, and uses a competitive weighting mechanism to fuse multiple deformation sub-fields, resulting in more accurate estimation of complex non-rigid deformations.

PIViT: A pyramid-based registration method leveraging optical flow modelling, which incorporates iterative swin transformer blocks to enable progressive alignment at low resolution, balancing efficiency and accuracy.

MambaMorph: The single-stage registration network constructed based on Mamba encoder integrates the efficient sequence modelling capability and lightweight feature extractor to achieve efficient long-distance correspondence modelling and high-dimensional feature learning.

TransMatch: An end-to-end transformer-based registration framework that introduces cross-attention modules at each layer of the encoder and decoder to enhance both local and global context modelling.

Table 1 summarizes the key characteristics and hyperparameter settings of all comparison methods. Methods marked with an asterisk (*) are unofficial implementations provided by the ModeT authors. For the remaining methods, hyperparameters were configured based on the recommendations in their original publications to ensure optimal performance.

4.3 | Implementation Details

We tested our method using PyTorch on an NVIDIA GeForce RTX 4090 GPU equipped with 24 GB of memory. To ensure the smoothness of the deformation field, the regularization term λ was set to 1.We apply a learning rate decay strategy where the learning rate lr_m decreases following a power-law function with an exponent of 0.8, ensuring a smooth decay over M epochs:

$$lr_m = lr_{\text{init}} \cdot \left(1 - \frac{m-1}{M}\right)^{0.8}, m = 1, 2, \ldots, M \tag{14}$$

where lr_m represents the learning rate of mth epoch and $lr_{\text{init}} = 0.0001$ represents the learning rate of initial epoch. We set the batch size as 1, M as 30 for training.

TABLE 1 | Hyperparameter settings for each method.

Journal/ Conference	Methods	Loss	Hyperparameters
2022 MIA	TransMorph	NCC	$\lambda = 1, l_r = 0.0001$
2022 MMMI	I2G*	NCC	$\lambda = 1, l_r = 0.0001$
2022 MIA	PR++*	NCC	$\lambda = 1, l_r = 0.0001$
2023 MICCAI	ModeT	NCC	$\lambda = 1, l_r = 0.0001$
2023 MICCAI	PIVIT	NCC	$\lambda = 1, l_r = 0.0001$
2024 arXiv	MambaMorph	Dice	$\lambda = 0.1, l_r = 0.0001$
2024 IEEE-TMI	TransMatch	NCC	$\lambda = 4, l_r = 0.0004$

4.4 | Experimental Result and Analysis

We use the dice score (DSC) as the primary similarity metric to evaluate the overlap between corresponding regions. Additionally, the quality of the predicted non-rigid deformations is assessed by the percentage of voxels with a negative Jacobian determinant($\mathcal{J}_\phi$).

Tables 2 and 3 present the quantitative performance of various state-of-the-art medical image registration methods on the LPBA40, Mindboggle, and abdominal CT datasets. The values highlighted in red and bold indicate the best performance, while those in blue and bold denote the second-best. Our proposed PSMamba-Net consistently achieves the highest DSC scores across all three datasets, demonstrating its strong adaptability and generalization capability in diverse anatomical contexts, including both brain and abdominal structures. Specifically, PSMamba-Net outperforms the second-best method by 2.3% on the Mindboggle dataset, 1.0% on LPBA40, and achieves a notable 4.3% improvement on the abdominal CT abdominal dataset. These results indicate that our model exhibits strong robustness in handling complex and highly non-rigid deformations across a wide range of anatomical structures.

The performance gains can be attributed to the synergistic design of several key architectural components. First, the optimized iterative registration module integrated at the bottom of the network refines the deformation field by leveraging fixed high-level features, enabling it to effectively model large-scale anatomical deformations while significantly accelerating the registration process by avoiding repeated feature extraction. Second, the spatial-Mamba block (SMB), serving as a core component of the decoder, enhances the modelling of spatial relationships between features. By harnessing Mamba's strengths in efficient long-sequence modelling, the SMB module addresses the limitations of traditional transformer-based models in preserving spatial

TABLE 2 | Quantitative comparison of the research methods was conducted on the datasets of Mindboggle and LPBA40.

Method	Mindboggle (62 ROIs)		LPBA40 (54 ROIs)	
	DSC ↑	$\lvert \mathcal{J}_\phi \rvert \leq 0\%$ ↓	DSC ↑	$\lvert \mathcal{J}_\phi \rvert \leq 0\%$ ↓
Affine	0.340 ± 0.019	—	0.537 ± 0.048	—
TransMorph [21] (2022)	0.555 ± 0.065	<1.6%	0.694 ± 0.027	<0.2%
I2G [33] (2022)*	0.594 ± 0.015	**<0.03%**	0.704 ± 0.016	<0.02%
PR++ [32] (2022)*	0.593 ± 0.019	<0.5%	0.694 ± 0.023	<0.1%
ModeT [40] (2023)	0.622 ± 0.011	**<0.03%**	0.721 ± 0.014	<0.008%
PIVIT [30] (2023)	0.602 ± 0.013	<0.04%	0.7163 ± 0.014	<0.03%
MambaMorph [27] (2024)	0.598 ± 0.021	<0.05%	0.685 ± 0.029	**<0.001%**
TransMatch [41] (2024)	0.527 ± 0.016	<0.09%	0.6921 ± 0.020	<0.007%
PSMamba-Net (present)	**0.644 ± 0.013**	<0.4%	**0.731 ± 0.015**	<0.09%
PSMamba-Net_diff (present)	**0.644 ± 0.013**	**<0.003%**	**0.732 ± 0.014**	**<0.00003%**

TABLE 3 | Quantitative comparison of the research methods was conducted on the datasets of Abdominal CT.

Method	Abdominal CT (13 ROIs)	
	DSC ↑	$\lvert \mathcal{J}_\phi \rvert \leq 0\%$ ↓
TransMorph [21] (2022)	0.318 ± 0.046	<0.00001%
I2G [33] (2022)*	0.456 ± 0.039	<0.002%
PR++ [32] (2022)*	0.422 ± 0.059	<0.02%
ModeT [40] (2023)	0.493 ± 0.036	<0.005%
PIVIT [30] (2023)	0.316 ± 0.029	<0.001%
MambaMorph [27] (2024)	0.446 ± 0.061	<0.001%
TransMatch [41] (2024)	0.428 ± 0.058	<0.06%
PSMamba-Net (present)	**0.536 ± 0.004**	<0.04%
PSMamba-Net_diff (present)	**0.526 ± 0.052**	**<0.0001%**

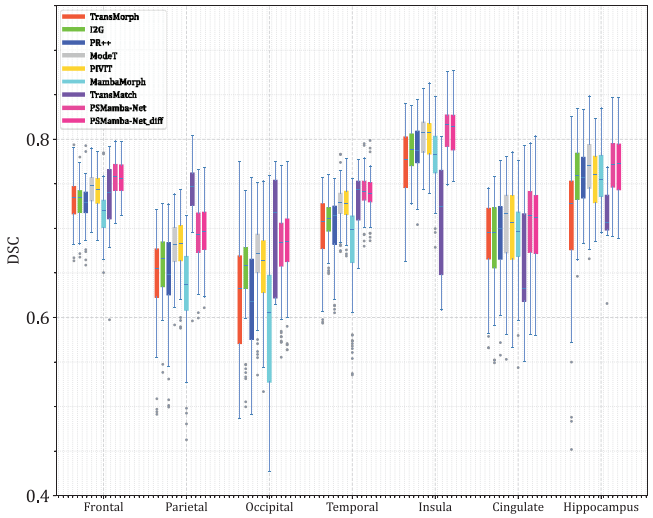

FIGURE 2 | Dice performance across anatomical structures (LPBA40).

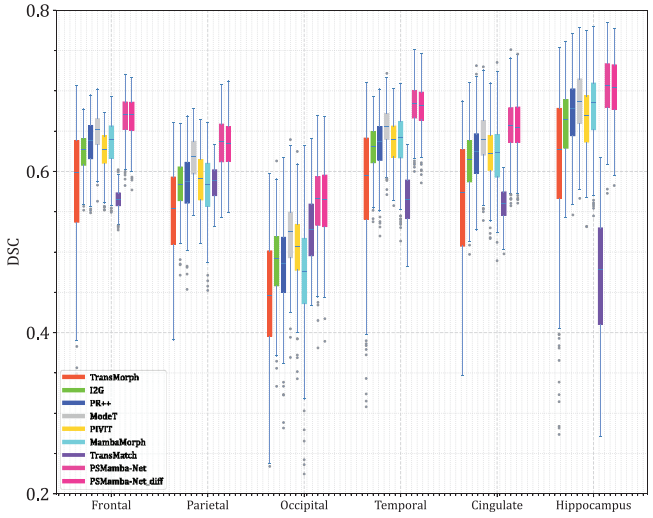

FIGURE 3 | Dice performance across anatomical structures (Mindboggle).

structure, thereby improving the anatomical continuity and consistency of the deformation fields. Besides this, PSMamba-Net not only achieves superior registration accuracy but also maintains favourable smoothness and physical plausibility in terms of the Jacobian determinant metric ($\lvert \mathcal{J}_\phi \rvert \leq 0$), demonstrating an excellent balance between accuracy and deformation quality.

Figures 2 and 3 provide the DSC values for different methods in specific regions across the two datasets. Figure 2 presents the DSC values for the seven regions: frontal, parietal, occipital, temporal, insula, cingulate, and hippocampus, while Figure 3 shows the DSC values for the six regions: frontal, parietal, occipital, temporal, cingulate, and hippocampus. From the figures, it can be observed that our method achieves more significant results in regions with relatively complex structures or greater individual variability, such as the hippocampus and insula. This phenomenon suggests that the preserved spatial features in our method can better guide the registration of brain regions with complex shapes and individual differences, especially deep brain structures.

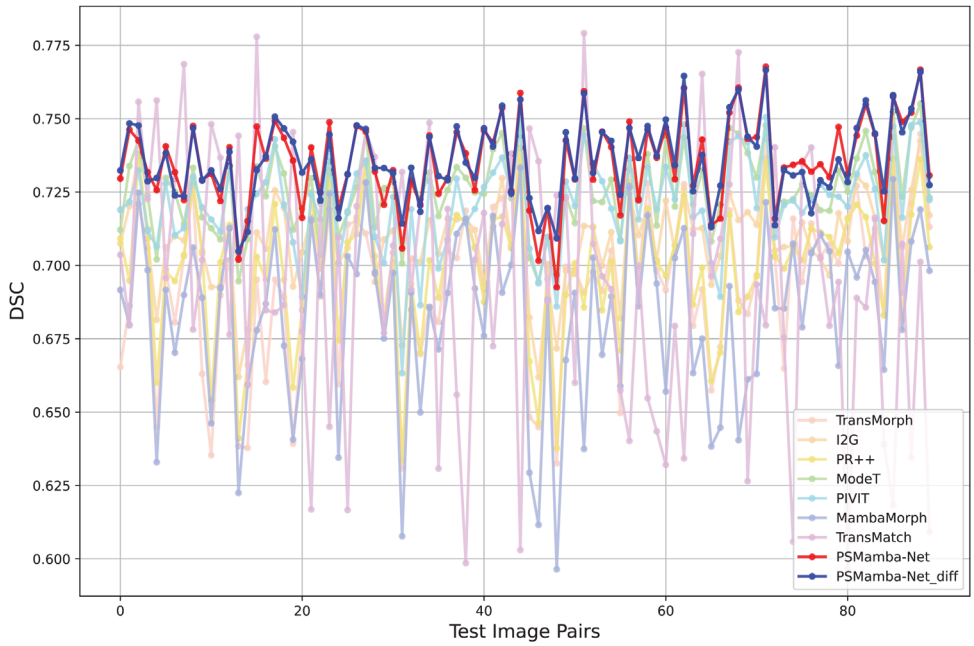

FIGURE 4 | DSC evaluation of registration methods on LPBA40 dataset.

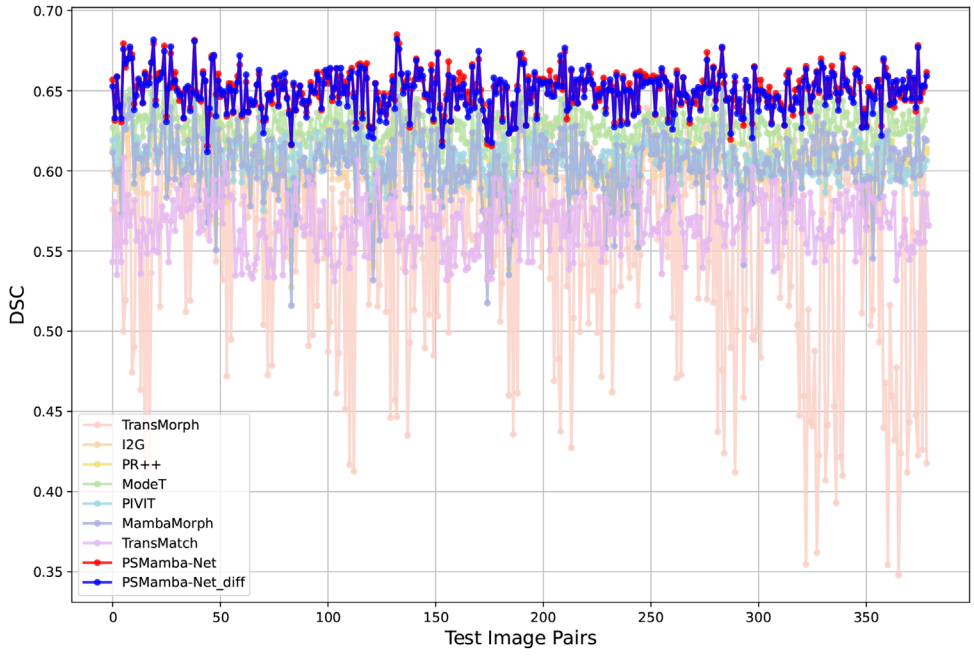

FIGURE 5 | DSC evaluation of registration methods on Mindboggle dataset.

Furthermore, Figures 4 and 5 illustrate the DSC values for each test image pair using different registration methods across the two datasets. For the LPBA40 dataset, the test set consists of ten images (10×9 pairs) that are not included in the training set. For the Mind dataset, the test set comprises 20 images from the OASIS-TRT-20 subset. From the figures, it is evident that our method not only achieves the best registration results but also exhibits smaller fluctuations in DSC values for each epoch compared to other methods, which is particularly noticeable in Figure 5. This demonstrates the robustness of our approach.

Figures 6 and 7 elucidate the registration results of various methodologies across distinct slices of the two datasets. The first five rows showcase coronal views, while the subsequent five rows present axial views. Each row, arranged from top to bottom, encapsulates the warped image, the segmentation labels corresponding to the warped image, the colour jet map of the warped image, the deformation field (rendered in RGB), and the deformation grid. To enhance the differentiation of various regions within the label images, we utilize a heatmap instead of a conventional greyscale representation. The findings in the initial three rows for each method reveal that our approach surpasses

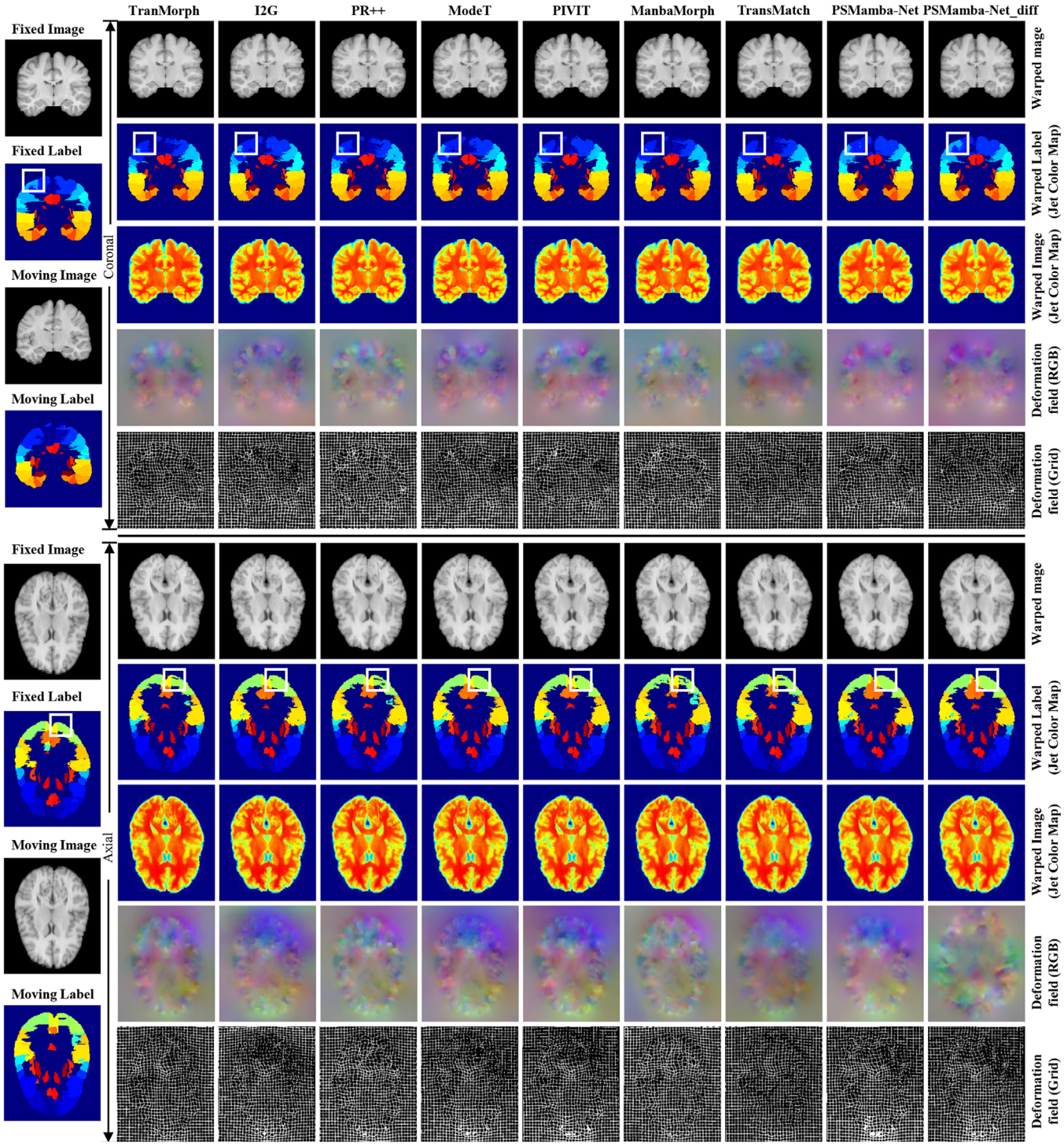

FIGURE 6 | Registration output comparison on LPBA40 dataset (Coronal & Axial views). Each column corresponds to a registration method (with the last two representing ours), and each row presents the deformed image, segmentation, Jet colourmap, deformation field in RGB, and grid representation.

others in its ability to capture intricate texture details. Moreover, our methodology demonstrates superior performance in addressing significant deformations compared to techniques that do not incorporate iterative registration. The results depicted in the last two rows further underscore that the deformation fields produced by our method exhibit notably smoother characteristics, thereby augmenting the overall efficacy of the registration process. This synthesis of precision in detail and adeptness in managing substantial deformations highlights the merits of our proposed approach in the realm of medical image registration tasks.

5 | Ablation Study

To verify the effectiveness of the individual modules in PSMamba-Net, we conducted ablation experiments. Table 4 presents the

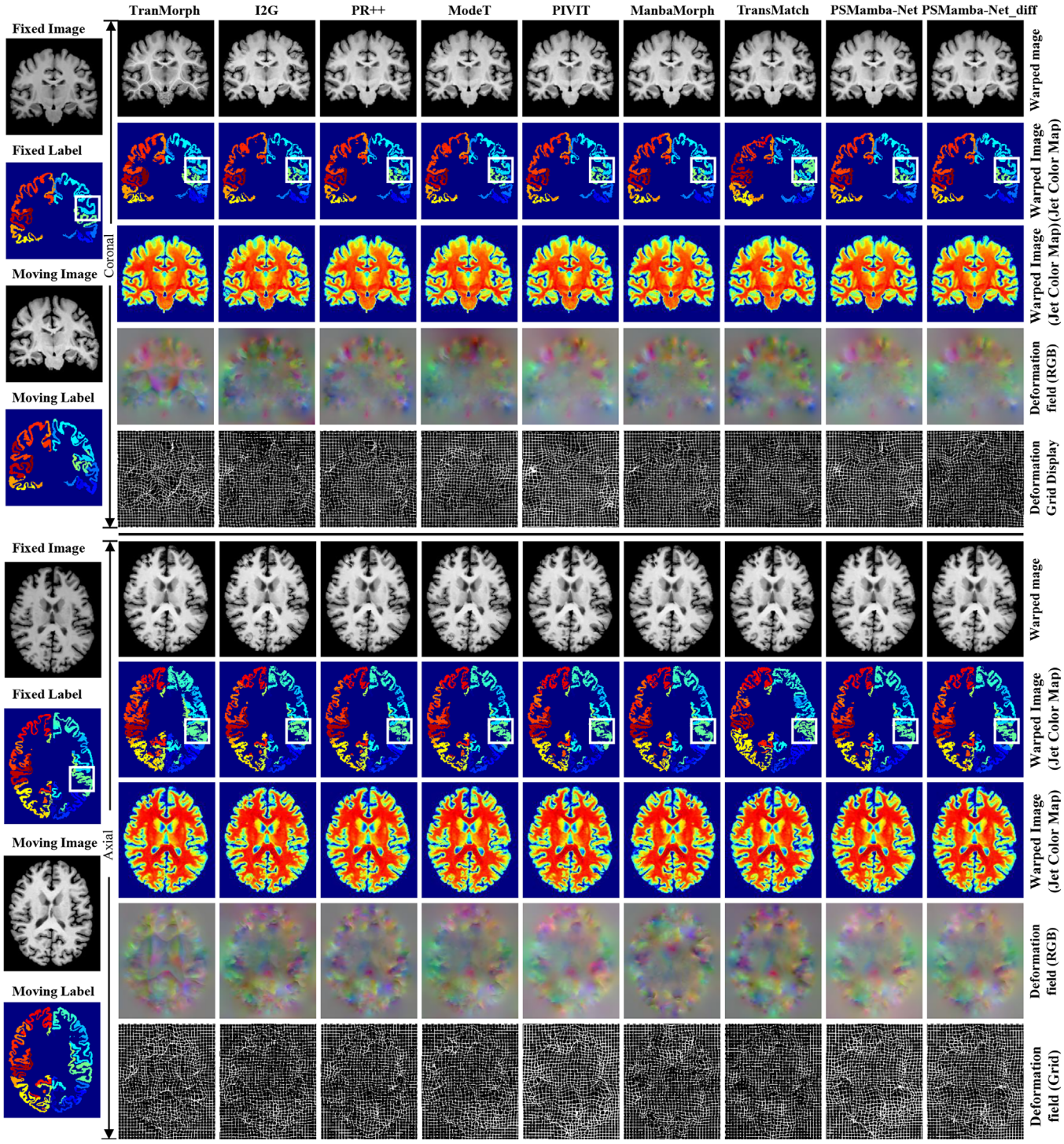

FIGURE 7 | Registration output comparison on Mindboggle dataset (coronal & axial views).

results of these ablation studies conducted on the LPBA40 dataset. We replaced our module with standard 3D convolutions and used this as the baseline method. Based on this, we sequentially replaced components with our proposed SMB blocks and the substrate iterative refinement module to assess the effectiveness of each module. Figure 8 shows the visualization results from our ablation experiments.

As shown in the table, when only the SMB module was utilized, our DSC metric reached 72.7% on the LPBA40 dataset, surpassing the baseline method using only standard convolutions. Additionally, the percentage of voxels with non-positive Jacobian determinants was lower than that of the baseline method, indicating that the spatial features extracted by our SMB module allowed for the generation of more accurate and smoother deformation fields. When only the substrate iterative refinement module was used, the DSC metric increased to 72.8% on the LPBA40 dataset, again outperforming the baseline method. However, the percentage of voxels with non-positive Jacobian determinants was higher than that of the baseline, suggesting that while

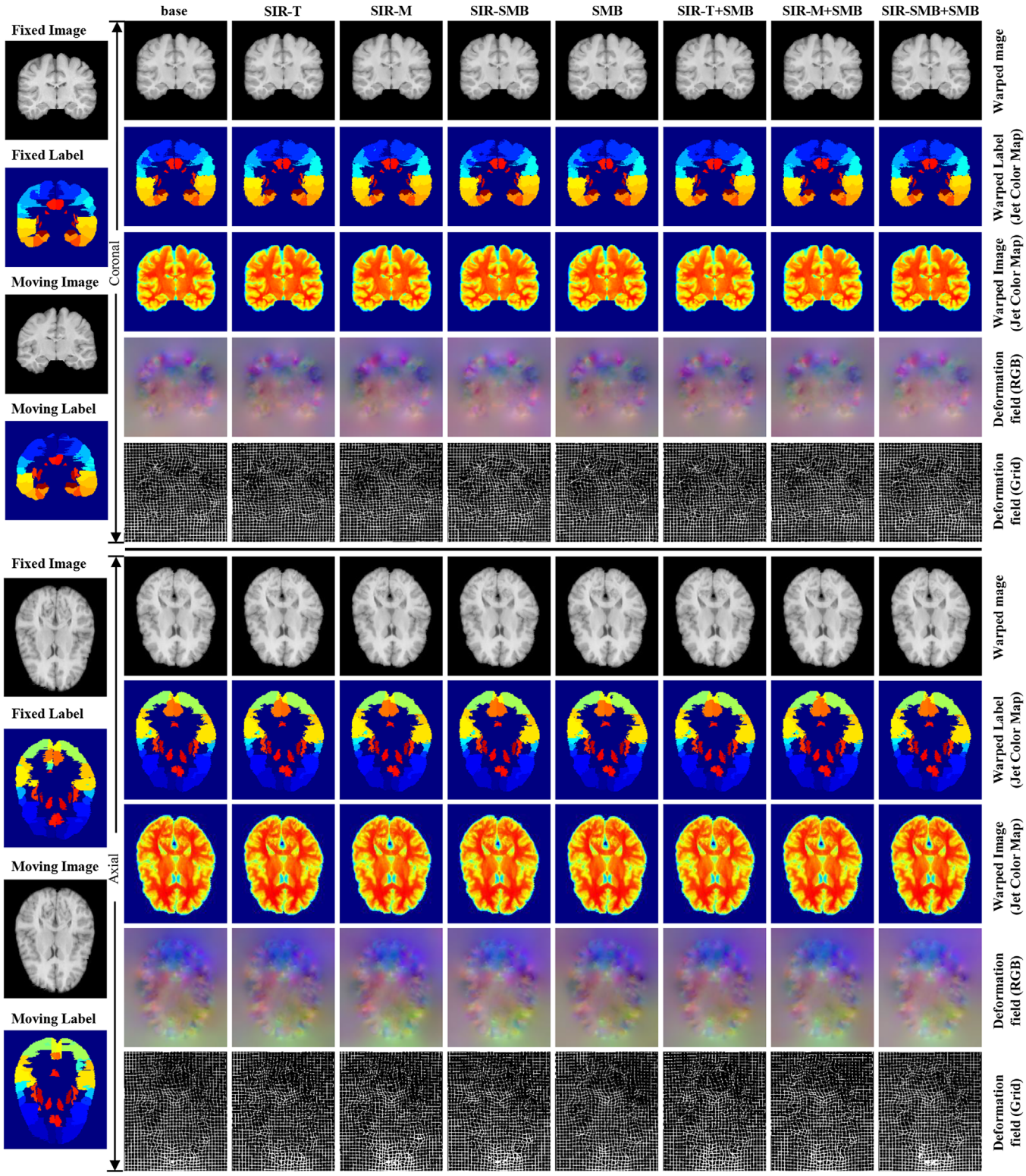

FIGURE 8 | Comparison of ablation study results on LPBA40 dataset (coronal & axial views).

multiple iterations in the substrate iterative refinement module helped create a more accurate deformation field transitioning from coarse to fine, they also slightly increased the distortion of the deformation field.

Furthermore, we assessed the effects of using transformer, standard Mamba, and SMB blocks within the substrate iterative refinement module to examine the compatibility of the SMB blocks. The results showed that the use of transformer and standard Mamba yielded DSC values of 72.5% and 72.7%, respectively, while the application of the SMB block achieved a DSC value of 73.1%. Notably, when we implemented the PSMamba-Net_diff variant, not only did the DSC metric peak at 73.2%, but it also achieved the best performance in terms of the percentage of

TABLE 4 | Quantitative evaluation of models using different modules in PSMamba-Net.

SIR-T	SIR-M	SIR-SMB	SMB	DSC ↑	$\lvert \mathcal{J}_\phi \rvert \leq 0\%$ ↓
—	—	—	—	0.724	**<0.02%**
√	—	—	—	0.721	<0.03%
—	√	—	—	0.726	<0.03%
—	—	√	—	0.728	<0.05%
—	—	—	√	0.727	**<0.001%**
√	—	—	√	0.725	<0.06%
—	√	—	√	**0.728**	<0.06%
—	—	√	√	**0.731**	<0.09%

SIR-T, SIR-M, and SIR-SMB represent the use of Transformer, standard Mamba, and SMB blocks, respectively, within the Substrate Iterative Refinement module.

voxels with non-positive Jacobian determinants. This indicates that our SMB block is more compatible with the substrate iterative refinement module than either the transformer or the standard Mamba.

6 | Conclusion

This paper presents a medical image registration network based on optimized iteration and Mamba, featuring a dual-stream pyramid structure. We adopt a dual-stream pyramid registration architecture that reduces the scope of attention computations required at each decoding level, thereby alleviating computational costs. At the bottom of the pyramid network, we introduce a substrate iterative refinement module to capture large deformations. Since iterations are performed only at the lowest level of the feature pyramid, there is no need to repeatedly extract image features, which significantly speeds up the registration process. Additionally, we incorporate the SMB module as a decoder, which not only leverages the spatial relationships that Mamba previously overlooked but also takes advantage of Mamba's strengths in modelling long sequences. This further reduces computational costs and model storage, achieving a better balance between the performance and efficiency of the registration model. The experimental results on LPBA 40, Mindboggle and Abdominal CT indicate that our PSMamba-Net outperforms the state-of-the-art methods in terms of speed while maintaining good registration accuracy.

When we conducted hypothesis testing using paired t-tests combined with Bonferroni correction, we found that the *p*-values were less than 0.0005 when comparing PSMamba-Net with all other methods. In addition, we performed one-way ANOVA and Tukey's HSD post-hoc test, which further confirmed that PSMamba-Net significantly outperforms the other methods.

Although our method has achieved promising results in experiments, there is still room for improvement. In the future, we will explore more advanced spatial-aware architectures to further enhance feature representation. Additionally, our current model has been trained and evaluated primarily on brain MRI datasets. As the next step, we plan to extend this approach to other medical imaging modalities, such as CT, PET, and ultrasound, to comprehensively validate its robustness and adaptability across different imaging types.

Author Contributions

The source manuscript was written by Zilong Xue and Kangjian He. Funding for this work was provided by Kangjian He and Dan Xu, while Jian Gong verified the experimental results. All authors contributed to the review of the manuscript.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62202416, 62462066, 62162068, in part by the Yunnan Fundamental Research Projects under Grant 202401AU070204, 202501AT070228, in part by the Yunnan Provincial Science and Technology Department-Yunnan University Double First Class Construction Joint Fund Project under Grant No. 202301BF070001-025.

Ethics Statement

This study used publicly available medical imaging data. In accordance with the license for open-access data, ethical approval was considered not necessary.

Conflicts of Interest

The authors declare no conflicts of interest.

Code Availability

Our source code is available at: https://github.com/VCMHE/PSMamba.

Data Availability Statement

The publicly available datasets utilized in this study include the LPBA dataset, accessible at https://resource.loni.usc.edu/resources/atlases-downloads/, and the Mindboggle dataset, accessible at https://osf.io/yhkde/, and the Abdomen CT dataset, accessible at https://learn2reg.grand-challenge.org/Datasets/. Access to these datasets is governed by their respective licenses.

References

1. P. K. R. Yelampalli, J. Nayak, and V. H. Gaidhane, "Daubechies Wavelet-Based Local Feature Descriptor for Multimodal Medical Image Registration," *IET Image Processing* 12, no. 10 (2018): 1692–1702.

2. H. Yu, Q. Zheng, F. Hu, C. Ma, S. Wang, and S. Wang, "MSCAReg-Net: Multi-scale Complexity-Aware Convolutional Neural Network for Deformable Image Registration," *IET Image Processing* 18, no. 4 (2024): 839–855.

3. T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Diffeomorphic Demons: Efficient Non-Parametric Image Registration," *NeuroImage* 45, no. 1 (2009): S61–S72.

4. J. Glaunes, A. Qiu, M. I. Miller, and L. Younes, "Large Deformation Diffeomorphic Metric Curve Mapping," *International Journal of Computer Vision* 80 (2008): 317–336.

5. B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric Diffeomorphic Image Registration With Cross-Correlation: Evaluating Automated Labeling of Elderly and Neurodegenerative Brain," *Medical Image Analysis* 12, no. 1 (2008): 26–41.

6. M. F. Beg, M. I. Miller, A Trouvé, and L. Younes, "Computing Large Deformation Metric Mappings Via Geodesic Flows of Diffeomorphisms," *International Journal of Computer Vision* 61 (2005): 139–157.

7. C. Liu, K. He, D. Xu, H. Shi, H. Zhang, and K. Zhao, "RegFSC-Net: Medical Image Registration via Fourier Transform with Spatial Reorganization and Channel Refinement Network," *IEEE Journal of Biomedical and Health Informatics* 28, no. 6 (2024): 3489–3500.

8. M. Kang, X. Hu, W. Huang, M. R. Scott, and M. Reyes, "Dual-stream Pyramid Registration Network," *Medical Image Analysis* 78 (2022): 102379.

9. L. Li, L. Li, Y. Zhang, et al., "Explicit–Implicit Symmetric Diffeomorphic Deformable Image Registration With Convolutional Neural Network," *IET Image Processing* 18, no. 13 (2024): 3892–3903.

10. G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: A Learning Framework for Deformable Medical Image Registration," *IEEE Transactions on Medical Imaging* 38, no. 8 (2019): 1788–1800.

11. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference* (Springer, 2015), 234–241.

12. A. Hering, B. Van Ginneken, and S. Heldmann, "mlVIRNet: Multilevel Variational Image Registration Network," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2019: 22nd International Conference* (Springer, 2019), 257–265.

13. S. Zhao, Y. Dong, E. I. Chang, et al., "Recursive Cascaded Networks for Unsupervised Medical Image Registration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, 2019), 10600–10610.

14. T. C. Mok and A. Chung, "Fast Symmetric Diffeomorphic Image Registration With Convolutional Neural Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2020), 4644–4653.

15. B. Kim, D. H. Kim, S. H. Park, J. Kim, J. G. Lee, and J. C. Ye, "CycleMorph: Cycle Consistent Unsupervised Deformable Image Registration," *Medical Image Analysis* 71 (2021): 102036.

16. M. Hoffmann, B. Billot, D. N. Greve, J. E. Iglesias, B. Fischl, and A. V. Dalca, "SynthMorph: Learning Contrast-Invariant Registration Without Acquired Images," *IEEE Transactions on Medical Imaging* 41, no. 3 (2021): 543–558.

17. D. Kuang and T. Schmah, "Faim–a Convnet Method for Unsupervised 3d Medical Image Registration," in *Machine Learning in Medical Imaging: 10th International Workshop* (Springer, 2019), 646–654.

18. A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised Learning of Probabilistic Diffeomorphic Registration for Images and Surfaces," *Medical Image Analysis* 57 (2019): 226–236.

19. J. Chen, E. C. Frey, and Y. Du, "Unsupervised Learning of Diffeomorphic Image Registration via Transmorph," in *International Workshop on Biomedical Image Registration* (Springer, 2022), 96–102.

20. A. Vaswani, "Attention is All You Need," in *Advances in Neural Information Processing Systems* (ACM, 2017), 6000–6010.

21. J. Chen, E. C. Frey, Y. He, W. P. Segars, Y. Li, and Y. Du, "Transmorph: Transformer for Unsupervised Medical Image Registration," *Medical Image Analysis* 82 (2022): 102615.

22. X. Song, H. Guo, X. Xu, et al., "Cross-Modal Attention for MRI and Ultrasound Volume Registration," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference* (Springer, 2021), 66–75.

23. J. Shi, Y. He, Y. Kong, et al., "XMorpher: Full Transformer for Deformable Medical Image Registration via Cross Attention," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2022), 217–226.

24. A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling With Selective State Spaces," *arXiv:2312.00752* (2023).

25. A. Gu, K. Goel, and C. Ré, "Efficiently Modeling Long Sequences With Structured State Spaces," *arXiv:2111.00396* (2021).

26. A. Gu, I. Johnson, K. Goel, et al., "Combining Recurrent, Convolutional, and Continuous-Time Models With Linear State Space Layers," *Advances in Neural Information Processing Systems* 34 (2021): 572–585.

27. T. Guo, Y. Wang, and C. Meng, "Mambamorph: A Mamba-Based Backbone With Contrastive Feature Learning for Deformable mr-ct Registration," *arXiv:2401.13934* (2024).

28. Z. Wang, J. Q. Zheng, C. Ma, and T. Guo, "VMambaMorph: A Multi-Modality Deformable Image Registration Framework Based on Visual State Space Model With Cross-scan Module," *arXiv:2404.05105* (2024).

29. Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, "SegMamba: Long-Range Sequential Modeling Mamba for 3d Medical Image Segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2024), 578–588.

30. T. Ma, X. Dai, S. Zhang, and Y. Wen, "PIViT: Large Deformation Image Registration With Pyramid-Iterative Vision Transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2023), 602–612.

31. T. C. Mok and A. C. Chung, "Large Deformation Diffeomorphic Image Registration With Laplacian Pyramid Networks," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2020: 23rd International Conference* (Springer, 2020), 211–221.

32. M. Kang, X. Hu, W. Huang, M. R. Scott, and M. Reyes, "Dual-stream Pyramid Registration Network," *Medical Image Analysis* 78 (2022): 102379.

33. Y. Liu, L. Zuo, S. Han, Y. Xue, J. L. Prince, and A. Carass, "Coordinate Translator for Learning Deformable Medical Image Registration," in *International Workshop on Multiscale Multimodal Medical Imaging* (Springer, 2022), 98–109.

34. P. K. R. Yelampalli, J. Nayak, and V. H. Gaidhane, "Daubechies Wavelet-Based Local Feature Descriptor for Multimodal Medical Image Registration," *IET Image Processing* 12, no. 10 (2018): 1692–1702.

35. A. Legouhy, O. Commowick, F. Rousseau, and C. Barillot, "Unbiased Longitudinal Brain Atlas Creation Using Robust Linear Registration and Log-Euclidean Framework for Diffeomorphisms," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (IEEE, 2019), 1038–1041.

36. V. Arsigny, O. Commowick, X. Pennec, and N. Ayache, "A Log-Euclidean Framework for Statistics on Diffeomorphisms," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2006: 9th International Conference* (Springer, 2006), 924–931.

37. M. Zhang and P. T. Fletcher, "Fast Diffeomorphic Image Registration via Fourier-Approximated Lie Algebras," *International Journal of Computer Vision* 127 (2019): 61–73.

38. M. Hernandez, "Band-Limited Stokes Large Deformation Diffeomorphic Metric Mapping," *IEEE Journal of Biomedical and Health Informatics* 23, no. 1 (2018): 362–373.

39. Y. R. Rao, N. Prathapani, and E. Nagabhooshanam, "Application of Normalized Cross Correlation to Image Registration," *International Journal of Research in Engineering and Technology* 3, no. 5 (2014): 12–16.

40. H. Wang, D. Ni, and Y. Wang, "ModeT: Learning Deformable Image Registration Via Motion Decomposition Transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2023), 740–749.

41. Z. Chen, Y. Zheng, and J. C. Gee, "TransMatch: A Transformer-Based Multilevel Dual-Stream Feature Matching Network for Unsupervised Deformable Image Registration," *IEEE Transactions on Medical Imaging* 43, no. 1 (2024): 15–27.