Spectral-Adaptive Pyramid Network for Deformable T1-MRI Image Registration

Yuxuan Wang[0009-0002-3577-8261], Chenyang Jiang[0009-0000-1958-7006], and Dongxun Jiang(✉)[0009-0003-2978-0047]

Tongji University, Shanghai 201804, China
{wangyuxuan_5,lighting374,jdx}@tongji.edu.cn

Abstract. Deformable image registration aims to establish dense spatial correspondences between anatomical structures, yet existing pyramid-based methods overlook the inherent correspondence between pyramid levels and frequency bands, leading to spectral interference that degrades anatomical boundary discrimination across scales. In this paper, we propose a spectral-adaptive pyramid registration network to align spectral information across scales. Spectral Adaptive Encoder (SAE) is designed to explicitly separate and adaptively fuse multi-scale spectral features directly at the initial encoding stage. Frequency-aware Spatial Adaptive Clustering Block (FreqSACB) is designed to perform spectral-adaptive purification for spatial pattern aggregation across the multi-scale decoding phase. We further theoretically prove the existence of scale-adaptive frequency bands for each level of a Gaussian pyramid. Experiments on the LPBA40 dataset demonstrate state-of-the-art performance with a Dice of 72.73% and IoU of 57.79%, with ablation studies confirming the complementary contribution of each component.

Keywords: Deformable Image Registration · Spectral Decomposition · Frequency-Aware Feature Extraction · Multi-Scale Evaluation · Spectral-Adaptive Processing.

1 Introduction

Deformable image registration, a cornerstone of medical image analysis, estimates dense spatial correspondences between moving and fixed images, supporting applications such as atlas construction, longitudinal analysis, and surgical planning [10]. Classical energy-based methods such as Demons [14], ANTs [1], and LDDMM [3] achieve decent accuracy through iterative optimization. Although these methods are reliable, their inference times are too long to meet real-time clinical requirements.

Deep learning fundamentally transforms registration by enabling real-time inference. VoxelMorph [2] establishes an end-to-end unsupervised U-Net framework that remains widely adopted. To capture long-range dependencies beyond the

Y. Wang and C. Jiang—Equal contribution.

receptive field of convolutions, TransMorph [4] integrates Swin Transformers for global context, and MambaMorph [6] adopts state space models to reduce the quadratic cost of attention. Nevertheless, these single-scale or globally-attentive architectures struggle to simultaneously resolve large deformations and fine boundary details, motivating the adoption of multi-scale pyramid frameworks.

To this end, pyramid-based coarse-to-fine strategies emerge as the dominant paradigm in brain MRI registration. LapIRN [9] decomposes displacement fields across Laplacian pyramid levels, while cascaded networks [16] and the Recursive Decomposition Network [7] refine estimates progressively from coarse to fine. ModeT [15] decomposes motion fields via a Transformer operating over pyramid features and SACB-Net [5] introduces spatial-awareness convolutions that adaptively aggregate features via clustering. These pyramid-based methods have effectively improved the accuracy of multiscale registration and have become the leading paradigm today.

Nevertheless, the aforementioned pyramid methods typically employ uniform convolutional operations across all pyramid levels, overlooking the fact that features at different resolutions inherently contain varying spectral characteristics. Consequently, existing methods fail to make effective use of the spectral complementary information across pyramid levels, thereby limiting the ability to distinguish anatomical boundaries and reducing registration accuracy.

Motivated by these observations, we propose a spectral-adaptive pyramid registration network that explicitly incorporates frequency-aware feature processing at both the initial encoding stage and throughout the coarse-to-fine pyramid hierarchy. Our main contributions are as follows:

1. We propose the **Spectral Adaptive Encoder (SAE)** to explicitly separate and adaptively fuse multi-scale spectral features directly at the initial encoding stage.
2. We introduce the **Frequency-aware Spatial Adaptive Clustering Block (FreqSACB)** to perform spectral-adaptive purification for spatial pattern aggregation.
3. We theoretically prove the existence of scale-adaptive frequency bands for each level of a Gaussian pyramid to provide a mathematical foundation for spectral-adaptive pyramid registration.

2 Related Works

2.1 U-Net and Its Variants

U-Net [12], with its encoder-decoder structure and skip connections, serves as the backbone for many learning-based image registration methods. VoxelMorph [2] first demonstrates the effectiveness of unsupervised learning in estimating dense deformation fields, achieving fast inference with competitive accuracy. TransMorph [4] later incorporates Swin Transformer encoders to enhance long-range spatial dependencies, establishing new performance benchmarks. Recently, MambaMorph [6] employes selective state space models for scalable context modeling

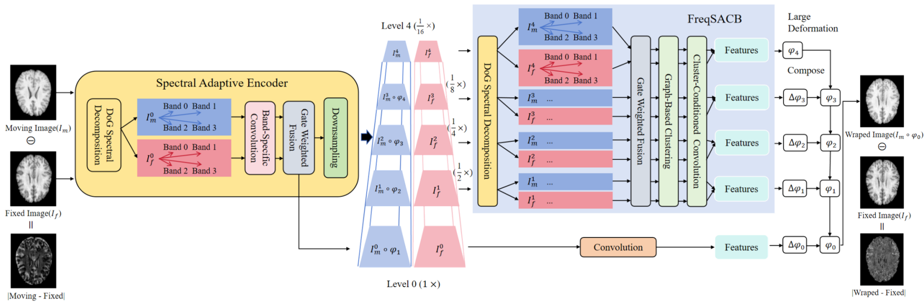

Fig. 1. Overview of the proposed spectral-adaptive pyramid registration network.

in high-resolution 3D volumes. However, these single-scale methods struggle with large deformations, emphasizing the need for pyramid-based cascaded approaches.

2.2 Pyramid-based and Cascaded Methods

Pyramid-based and cascaded approaches follow a coarse-to-fine paradigm to refine deformation estimates across multiple resolutions. Zhao et al. [16] proposes a recursive cascaded network, decomposing large displacements into smaller transformations. ModeT [15] introduces a motion decomposition transformer to disentangle deformation into interpretable components across scales, while SACB-Net [5] incorporates spatial adaptive convolutions with graph-based clustering to capture structurally coherent features at each pyramid level. While effective for handling large deformations, these methods operate purely in the spatial domain and do not account for the spectral variation across pyramid levels, limiting the discriminability of multi-scale feature representations.

3 Method

3.1 Motivation and Overview

Pyramid-based registration methods often apply uniform operations across all levels, overlooking the inherent correspondence between pyramid levels and frequency bands. At the initial encoding stage, standard convolutions indiscriminately mix all frequency components, entangling high-frequency details with low-frequency structures. During the multi-scale decoding phase, existing methods perform clustering on features that are not purified according to the spectral content of each level, causing further spectral interference.

To address this challenge, we propose a spectral-adaptive pyramid registration network. At the initial encoding stage, the Spectral Adaptive Encoder (SAE) explicitly separates and adaptively fuses multi-scale spectral features. Across the multi-scale decoding phase, the Frequency-aware Spatial Adaptive Clustering Block (FreqSACB) performs spectral-adaptive purification for spatial pattern aggregation.

3.2 Preliminaries

Difference-of-Gaussians (DoG). Given input $\mathbf{x} \in \mathbb{R}^{B\times C\times D\times H\times W}$, Gaussian blurring operations with increasing standard deviations $\{\sigma_k\}$ decompose $\mathbf{x}$ into complementary frequency bands:

$$\begin{aligned} B_0(\mathbf{x}) &= \mathbf{x} - G_{\sigma_1}(\mathbf{x}), \quad \text{(high-freq.)} \\ B_k(\mathbf{x}) &= G_{\sigma_k}(\mathbf{x}) - G_{\sigma_{k+1}}(\mathbf{x}), \;\; k = 1, \ldots, K{-}1, \quad \text{(mid-freq.)} \\ B_K(\mathbf{x}) &= G_{\sigma_K}(\mathbf{x}). \quad \text{(low-freq.)} \end{aligned} \tag{1}$$

where $B_k(\mathbf{x})$ presents the k-th frequency band after DoG decomposition. The decomposition ensures $\sum_{k=0}^{K} B_k = \mathbf{x}$, producing distinct features for anatomical structures from fine sulcal edges to global intensity gradients.

Gate Weighted Fusion. Given input $\mathbf{x} \in \mathbb{R}^{B\times C\times D\times H\times W}$, the per-band mean absolute strength s_k is computed as:

$$s_k = \frac{1}{BCDHW} \sum_{b,c,d,h,w} \left| B_k(\mathbf{x}) \right|_{bcdhw}, \tag{2}$$

The strength vector $\mathbf{s} = (s_0, \ldots, s_K) \in \mathbb{R}^{K+1}$ is mapped to gate weights $\mathbf{w}$ via a two-layer ReLU MLP followed by softmax normalisation:

$$\mathbf{w} = \mathrm{softmax}(\mathrm{MLP}(\mathbf{s}))\,, \tag{3}$$

and the bands are fused via an adaptive spectral weighting residual gate:

$$\mathbf{x}' = \sum_{k=0}^{K} w_k \cdot B_k(\mathbf{x}) + \mathrm{Conv}_{1\times 1}(\mathbf{x}), \tag{4}$$

where the residual $\mathrm{Conv}_{1\times 1}(\mathbf{x})$ preserves the original feature representation.

Spatial Adaptive Clustering Block Network (SACB-Net). SACB-Net is a clustering-based module that adaptively aggregates spatially coherent features at each pyramid level. Each SACB includes a GPU-accelerated k-means clustering module that partitions spatial feature vectors into k semantically coherent clusters. Given the feature map $\mathbf{x} \in \mathbb{R}^{B\times C\times D\times H\times W}$, the input is first projected via a $3 \times 3 \times 3$ convolution with instance normalisation, then unfolded into local patches. Each spatial location (d, h, w) is represented by a C-dimensional vector $\mathbf{x}_{dhw}$ obtained by averaging over the patch dimension. Clustering minimises within-cluster Euclidean distances:

$$\min_{\{\mathcal{C}_i\}_{i=1}^{k}} \sum_{i=1}^{k} \sum_{\mathbf{x}_{dhw} \in \mathcal{C}_i} \|\mathbf{x}_{dhw} - \boldsymbol{\mu}_i\|_2^2, \tag{5}$$

where $\boldsymbol{\mu}_i$ is the centroid of cluster $\mathcal{C}_i$. The resulting cluster assignments $\{c_{dhw}\}$ and centroids $\{\boldsymbol{\mu}_i\}$ define structurally meaningful regions for adaptive convolution.

Using the cluster centroids, SACB-Net generates cluster-specific convolutional kernels. For cluster i with centroid $\boldsymbol{\mu}_i$, a two-branch MLP predicts a kernel modulation weight $\mathbf{a}_i \in \mathbb{R}^{K^3}$ and bias $\mathbf{b}_i \in \mathbb{R}^{C_{\text{out}}}$:

$$\mathbf{a}_i = \sigma(\mathrm{MLP}_w(\boldsymbol{\mu}_i)), \qquad \mathbf{b}_i = \mathrm{MLP}_b(\boldsymbol{\mu}_i), \tag{6}$$

where σ is sigmoid. The cluster-specific kernel is obtained by element-wise modulation of a shared base kernel $\mathbf{W}$ along the spatial dimension: $\mathbf{W}_i = \mathbf{a}_i \odot \mathbf{W}$. The local output vector $\mathbf{y}_{dhw}$ is assembled as:

$$\mathbf{y}_{dhw} = \sum_{i=1}^{k} \big(\mathbf{W}_i * \mathbf{x}_{dhw} + \mathbf{b}_i\big) \cdot \mathbf{1}[c_{dhw} = i], \tag{7}$$

These local responses are subsequently assembled to form the complete feature map $\mathbf{y} \in \mathbb{R}^{B \times C \times D \times H \times W}$. The final output is then obtained by applying an activation function and a residual connection $\mathbf{y} \leftarrow \mathbf{y} + \mathbf{x}_{\text{in}}$, where $\mathbf{x}_{\text{in}}$ denotes the input to the SACB-Net prior to projection. This enables spatially adaptive convolutions conditioned on local structural context rather than a fixed volume-wide kernel.

3.3 Spectral Adaptive Encoder (SAE)

Due to the fact that standard convolutions indiscriminately mix all frequency components at the very first layer, causing high-frequency boundary details to be irreversibly entangled with low-frequency structural variations from the start, we propose Spectral Adaptive Encoder (SAE) as a replacement for the standard convolutional layer at the initial stage. SAE explicitly decomposes the raw input into separate frequency bands and processes each band independently, and adaptively fuses them to preserve multi-scale spectral features.

Specifically, the raw input image I is decomposed into four frequency bands $\{B_0, \ldots, B_3\}$ using a Difference-of-Gaussians (DoG) decomposition (eq. (1)). To extract the highest-frequency details that require the most precise registration in the pyramid from the raw image, we set the $(\sigma_1^0, \sigma_2^0, \sigma_3^0)$ to $(0.5,\ 1.0,\ 2.0)$. Each band is then processed by a dedicated $3 \times 3 \times 3$ convolution Conv_k, allowing the network to learn frequency-specific filters that are not compromised by cross-band interference.

The outputs of these band-specific convolutions are fused via an adaptive spectral weighting residual gate:

$$\mathbf{y} = \sum_{k=0}^{3} w_k \cdot \mathrm{Conv}_k(B_k) + \mathrm{Conv}_{1 \times 1}(I), \tag{8}$$

where the gate weights $\mathbf{w}$ are computed from the per-band mean absolute strength (eqs. (2) and (3)). This gating mechanism adaptively emphasises frequency bands that are most informative for the current input, suppressing irrelevant or noisy bands. The 1×1×1 skip connection with $Conv_{1\times1}(I)$ preserves the full-spectrum information as a safety residual.

The fused output y is normalised with instance normalisation and passed through a LeakyReLU activation before being fed into the rest of the encoder. By routing each frequency band through a dedicated convolution from the very first layer and adaptively fusing them via an adaptive spectral weighting residual gate, SAE ensures that the foundation of the feature pyramid is cleanly separated, providing frequency-purified representations that empower the downstream multi-scale FreqSACB modules to perform accurate and scale-adaptive filtering.

3.4 Frequency-aware Spatial Adaptive Clustering Block (FreqSACB)

To ensure that spatial clustering is driven by the optimal frequency band for the current scale, we propose the Frequency-aware Spatial Adaptive Clustering Block (FreqSACB).

Given a feature map $\mathbf{x}_\ell \in \mathbb{R}^{B\times C\times D\times H\times W}$ at pyramid level ℓ, FreqSACB first applies a DoG decomposition (eq. (1)). To avoid overlap between the physical frequency bands of different pyramid levels and to prevent the equivalent physical receptive field at the deepest level from exceeding the global image dimensions, we let pixel-domain $(\sigma_1^\ell, \sigma_2^\ell, \sigma_3^\ell)$ grow exponentially with the pyramid level. For $\ell = 1$ to $\ell = 4$, we adopt (0.25, 0.5, 1.0), (0.5, 1.0, 2.0), (1.0, 2.0, 4.0), and (2.0, 4.0, 8.0) respectively. After mapping to the physical domain by the down-sampling factor 2^ℓ, the equivalent physical $(\sigma_1^{p_\ell}, \sigma_2^{p_\ell}, \sigma_3^{p_\ell})$ become (0.5, 1.0, 2.0), (2.0, 4.0, 8.0), (8.0, 16.0, 32.0), and (32.0, 64.0, 128.0), forming a contiguous spectrum partitioning. This produces four complementary frequency bands $B_0, \ldots, B_3$ on each layer, ranging from high-frequency edges to low-frequency global structures. For each band, the mean absolute strength s_k is computed via eq. (2). A two-layer MLP with softmax maps the strength vector e to gate weights w eq. (3). The bands are then fused with a residual connection to obtain a frequency-purified feature map x_ℓ^{pur}:

$$\mathbf{x}_\ell^{\mathbf{pur}} = \sum_{k=0}^{3} w_k \cdot B_k(\mathbf{x}_\ell) + \mathrm{Conv}_{1\times1}(\mathbf{x}_\ell), \tag{9}$$

where the 1×1 convolution preserves the original information as a safeguard. This gating mechanism adaptively emphasises frequency bands that are most informative for the current scale while suppressing bands that would cause cross-band interference.

The purified feature map is then fed into the SACB-Net. Inside SACB-Net, the features are projected by a $3 \times 3 \times 3$ convolution, unfolded into local patches, and clustered via GPU-accelerated k-means using the Euclidean distance (eq. (5)). Each cluster centroid generates modulation weights and biases for a shared base kernel via two-branch MLPs (eq. (6)). The output for each spatial location is the sum of cluster-specific modulated convolutions over its assigned cluster, followed by a residual connection (eq. (7)). By integrating DoG decomposition and adaptive spectral weighting residual gating directly into the clustering block, FreqSACB ensures that SACB-Net always receives spectral-adaptive purified inputs, thereby avoiding the confusion of anatomical boundaries caused by band mixing.

3.5 Coarse-to-Fine Displacement Field Estimation

After FreqSACB processes the moving and fixed feature maps at level ℓ, a cross-similarity estimator computes a local displacement field. For each spatial location in the fixed feature map I_F^ℓ, the estimator attends over a $3\times3\times3$ neighbourhood in I_M^ℓ and produces an attention-weighted displacement vector:

$$\delta\varphi_\ell = \sum_{\mathbf{g}\in\mathcal{N}} \mathrm{softmax}\Big(\mathbf{f}_{I_F^\ell} \cdot \mathbf{f}_{I_M^\ell}^{(\mathbf{g})}\Big) \cdot \mathbf{g}, \tag{10}$$

where $\mathbf{g}$ enumerates the 27 offset vectors in the local neighbourhood. The displacement field is refined in a coarse-to-fine pyramid. At the coarsest level ($\ell = 4$), the initial field is estimated directly and upsampled. At subsequent levels $\ell \in \{3, 2, 1\}$, the moving features are first warped by the propagated field, and the residual displacement is accumulated:

$$\varphi_\ell = \uparrow\!\big(2 \cdot \left(\varphi_{\ell+1} \circ \delta\varphi_\ell + \delta\varphi_\ell\right)\big)\,, \tag{11}$$

capturing large deformations at coarse-scale levels and refining fine structural details at finer levels.

Finally, at the finest level ($\ell = 0$), the network directly leverages the high-fidelity features preserved by the Spectral Adaptive Encoder. The moving feature I_m^0 is warped by φ_1 and then directly concatenated with the fixed feature I_F^0. A dedicated dense convolutional block regresses the final microscopic residual displacement $\delta\varphi_0$. The ultimate full-resolution deformation field Φ is composited as $\Phi = \varphi_1 \circ \delta\varphi_0 + \delta\varphi_0$, which is then used to warp the original moving image.

3.6 Proof of Scale-Adaptive Frequency Band Existence

We hereby prove that, under a Gaussian pyramid with anti-aliasing filtering, there exists a scale-matched frequency band for each pyramid level.

Band-limitedness and Nyquist Frequency. Let $I : \mathbb{R}^d \to \mathbb{R}$ (typically $d = 2$ or 3) be a square-integrable image with bounded spatial support. Define a Gaussian pyramid with $L + 1$ levels $\ell = 0, 1, \ldots, L$, where $\ell = 0$ is the finest (original) level and $\ell = L$ the coarsest. The pyramid is constructed recursively:

$$I_{\ell+1} = \mathcal{D}\left(G_{\sigma_{\text{base}}} * I_\ell\right), \tag{12}$$

where $I_{\ell+1}$ is the image at pyramid level, $G_{\sigma_{\text{base}}}$ is a Gaussian kernel with standard deviation σ_{base}, $*$ denotes convolution, and $\mathcal{D}$ is downsampling by factor 2. The feature map at level ℓ is denoted F_ℓ.

Because the downsampling step is preceded by low-pass filtering (Gaussian blur) to avoid aliasing, the effective frequency content $\Omega^{\ell}_{\mathrm{Nyq}}$ of I_ℓ (and hence F_ℓ) is confined to the Nyquist frequency of that level:

$$\Omega^{\ell}_{\mathrm{Nyq}} = \frac{\Omega_{\max}}{2^{\ell}}, \tag{13}$$

where $\Omega_{\max}$ is the maximum frequency present in the original image I. In practice, $\Omega_{\max}$ is determined by the pixel spacing.

We now prove that for each pyramid level ℓ, there exists a choice of DoG parameters such that one of the bands B_k isolates a frequency range that is *unique* to that scale and can be used for enhancement.

Lemma 1 (Frequency support of pyramid levels). *For any level $\ell \geq 1$, the detail information that distinguishes level ℓ from the next coarser level $\ell+1$ lies exclusively in the frequency interval*

$$\mathcal{I}_\ell := \left(\Omega^{\ell+1}_{Nyq},\ \Omega^{\ell}_{Nyq}\right]. \tag{14}$$

Frequencies below $\Omega^{\ell+1}_{Nyq}$ are preserved at both levels; frequencies above Ω^{ℓ}_{Nyq} are already absent at level ℓ.

Proof. By construction, $I_{\ell+1}$ is obtained by low-pass filtering I_ℓ with cutoff $\approx \Omega^{\ell}_{\mathrm{Nyq}}$ (the Gaussian blur) followed by downsampling. The downsampling reduces the Nyquist frequency to $\Omega^{\ell+1}_{\mathrm{Nyq}} = \frac{\Omega^{\ell}_{\mathrm{Nyq}}}{2}$. Therefore, any frequency component ω with $\omega > \Omega^{\ell+1}_{\mathrm{Nyq}}$ cannot be represented in $I_{\ell+1}$ without aliasing. Conversely, such components are present in I_ℓ (since $\omega \leq \Omega^{\ell}_{\mathrm{Nyq}}$ for ℓ). Hence $\mathcal{I}_\ell$ is precisely the set of frequencies that are present at level ℓ but irrecoverably lost at level $\ell+1$.

Lemma 2 (DoG Peak Alignment and Strength Concentration). *For any interval $[\omega_{\min}, \omega_{\max}] \subset (0, \Omega^{\ell}_{Nyq}]$, there exists a choice of (σ_k, σ_{k+1}) such that the corresponding DoG band B_k aligns its peak frequency ω_{peak} strictly within this interval, thereby capturing a high concentration of the target spectral strength associated with the image of the $\mathcal{I}_\ell$.*

Proof. The frequency response of $B_k = G_{\sigma_k} - G_{\sigma_{k+1}}$ (in 1D radial sense) is

$$\widehat{B}_k(\omega) = e^{-2\pi^2\sigma_k^2\omega^2} - e^{-2\pi^2\sigma_{k+1}^2\omega^2}. \tag{15}$$

This function is zero at $\omega = 0$, increases to a maximum at $\omega_{\mathrm{peak}} = \frac{1}{\pi}\sqrt{\frac{\ln(\frac{\sigma_{k+1}}{\sigma_k})}{\sigma_{k+1}^2 - \sigma_k^2}}$, and then decays. Therefore, by properly determining the standard deviations σ_k and σ_{k+1}, we can deliberately constrain the peak frequency ω_{peak} within the target interval $[\omega_{\min}, \omega_{\max}]$. This precise positioning ensures that the interval inherently captures a high concentration of the desired spectral strength.

Theorem 1 (Existence of scale-matched frequency band). *For each pyramid level $\ell = 0, 1, \ldots, L-1$, there exists a frequency band $\mathcal{B}_\ell$ and a DoG filter H_ℓ such that:*

1. $\mathcal{B}_\ell \subseteq (0, \Omega^\ell_{Nyq}]$,
2. H_ℓ *acts as a soft band-pass filter for* $\mathcal{B}_\ell$ *(i.e., its frequency response is concentrated on* $\mathcal{B}_\ell$*),*
3. *The band* $\mathcal{B}_\ell$ *captures information that is present at level* ℓ *but absent at level* $\ell+1$ *(or for* $\ell = 0$*, captures the highest frequencies of the original image).*

Consequently, a network can be designed to extract $\mathcal{B}_\ell$ from F_ℓ and enhance it (e.g., via a learnable gate or a dedicated convolution) without affecting other frequency bands.

Proof. We treat two cases.
Case 1: $\ell \geq 1$. Set $\mathcal{B}_\ell$ as the upper half of the detail band:

$$\mathcal{B}_\ell = \left(\frac{\Omega^\ell_{\mathrm{Nyq}} + \Omega^{\ell+1}_{\mathrm{Nyq}}}{2},\ \Omega^\ell_{\mathrm{Nyq}} \right]. \tag{16}$$

By Lemma 1, this interval is a subset of $\mathcal{I}_\ell$ and thus contains scale-distinct information. By Lemma 2, we can choose σ^ℓ_k and σ^ℓ_{k+1} to construct a DoG band that possesses substantial spectral strength within it. Hence, applying H_ℓ to F_ℓ acts as a soft band-pass filter that effectively emphasizes the components within the desired band.
Case 2: $\ell = 0$ (finest scale). The original image I may contain frequencies up to $\Omega_{\max}$. The finest pyramid level has Nyquist frequency $\Omega^0_{\mathrm{Nyq}} = \Omega_{\max}$. The band that is lost after one downsampling step (i.e., going from $\ell = 0$ to $\ell = 1$) is $(\Omega^1_{\mathrm{Nyq}}, \Omega_{\max}]$. Setting $\mathcal{B}_0 = (\frac{\Omega_{\max}}{2}, \Omega_{\max}]$, we can again invoke the principles of Lemma 2 to construct a DoG band that possesses substantial spectral strength within it.

4 Experimental Results

4.1 Dataset and Experimental Settings

LPBA40 Dataset. We evaluate on the LPBA40 dataset [13], consisting of 40 T1-weighted brain MRI volumes with 54 manually delineated anatomical structures. Following standard practice [4, 9], subjects S_1–S_{25} are used for training, S_{26}–S_{30} for validation, and S_{31}–S_{40} for testing, with all possible pairs within each split forming the registration set.
Implementation Details. All experiments are implemented in PyTorch and trained on a single NVIDIA L40 GPU using the Adam optimizer [8] with an initial learning rate of 1×10^{-4} and polynomial decay (power 0.9). The loss function combines NCC similarity and ℓ_2 gradient regularization:

$$\mathcal{L} = \mathcal{L}_{\mathrm{NCC}} + 0.3\,\mathcal{L}_{\mathrm{reg}}. \tag{17}$$

Training runs for 10 epochs with batch size 1, cluster number $k = 7$, and input volume size $160 \times 192 \times 160$.
Evaluation Metrics We report two complementary metrics: (1) Dice Similarity Coefficient (Dice) averaged over all 54 anatomical structures; (2) Intersection over Union (IoU) as a label-based overlap measure.

Table 1. Performance comparison on the LPBA40 dataset. Results are reported as mean values.

methods	Dice (%)↑	IoU (%)↑
Random Forest(ML'01)	56.27	39.90
Support Vector Machine(ML'95)	45.78	30.83
VoxelMorph(TMI'19)	62.31	46.23
TransMorph(MIA'22)	61.99	46.04
B-Spline-Diff(MIDL'21)	54.44	38.60
ModeT(MICCAI'23)	71.88	56.73
SACB-Net(CVPR'25)	71.64	56.49
Ours	**72.73**	**57.79**

Table 2. Ablation study on the LPBA40 dataset. FreqSACB denotes Frequency-aware Spatial Adaptive Clustering Block.

Spectral Adaptive Encoder	FreqSACB	Dice (%)↑	IoU (%)↑
✗	✗	71.64	56.49
✓	✗	71.74	56.61
✓	✓	**72.73**	**57.79**

4.2 Overall Performance

We compare against two categories of baselines. For traditional machine learning, we include Random Forest and SVM. For deep learning, we compare against VoxelMorph [2], TransMorph [4], B-Spline-Diff [11], ModeT [15], and SACB-Net [5]. All models are trained using their officially released code, with optimal hyperparameters tuned on the validation sets.

Table 1 reports the Dice and IoU of our method against all baselines on the LPBA40 dataset. Traditional machine learning methods (Random Forest and SVM) perform significantly below deep learning approaches, confirming the necessity of end-to-end feature learning. Among deep learning methods, our approach achieves the best performance on Dice (72.73%), and IoU (57.79%), outperforming the strongest comparable baseline ModeT by 0.85% in Dice and 1.06% in IoU, which confirms the effectiveness of our spectral-adaptive design.

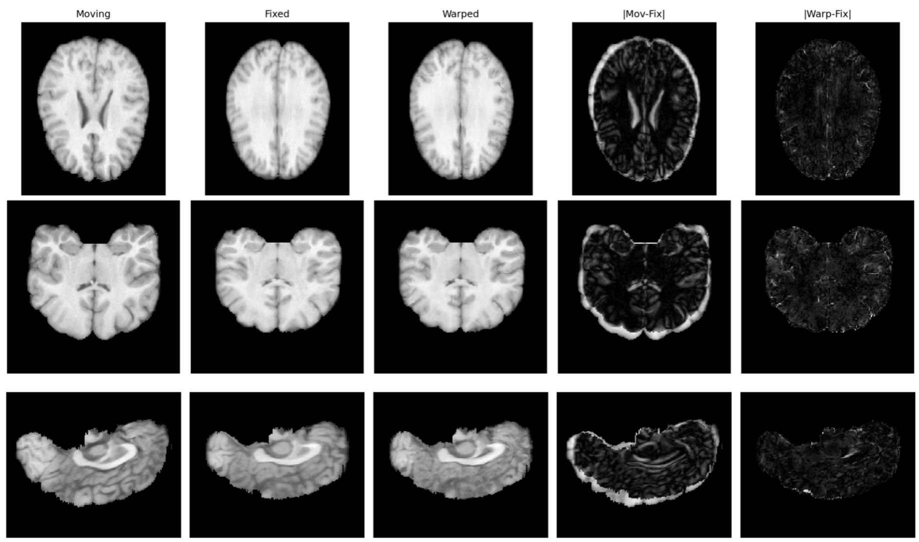

Fig. 2. Visualization of registration results for Pair 0006 on the LPBA40 dataset. Rows show axial, coronal, and sagittal slices; columns show the moving image, fixed image, warped image, and pre-/post-registration difference maps $|I_m - I_f|$ and $|I_m \circ \varphi - I_f|$.

4.3 Ablation Study

Table 2 evaluates the individual contribution of each proposed component. Removing both SAE and FreqSACB (Row 1) corresponds to the SACB-Net baseline. Adding SAE alone (Row 2) yields consistent improvements across all metrics, confirming that frequency-specific feature extraction at the finest encoder layer provides complementary structural cues. Enabling both components (Row 3) achieves the best performance across all metrics, demonstrating that SAE and FreqSACB contribute complementarily.

4.4 Registration Visualization

To visually assess the registration results, we generate qualitative visual comparison images for each sample pair before and after registration during the testing phase (as Figure 2). Given a moving image I_m and fixed image I_f, the model predicts a dense deformation field φ and resamples I_m via bilinear interpolation to obtain the warped image $I_m \circ \varphi$. For each volume, three intermediate slices are extracted along orthogonal directions (axial, coronal, and sagittal), forming a three-row layout. Each row contains five columns: the moving image, fixed image, warped image, pre-registration difference map $|I_m - I_f|$, and post-registration difference map $|I_m \circ \varphi - I_f|$. Brighter regions in the difference maps indicate greater misalignment. Therefore, the globally darker $|I_m \circ \varphi - I_f|$ produced by effective registration confirms improved structural alignment.

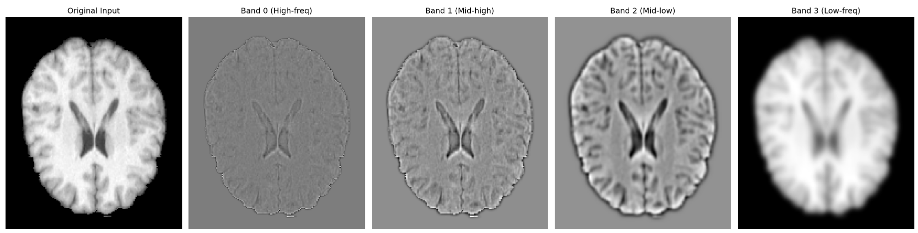

Fig. 3. Spectral decomposition of subject S_{36} using DoGSpectralDecomp3D with $\boldsymbol{\sigma} = (0.5, 1.0, 2.0)$. From left to right: original input, Band 0 (high-frequency edges and sharp boundaries), Band 1 and Band 2 (intermediate-scale sulcal transitions and subcortical structures), and Band 3 (low-frequency global brain morphology and smooth intensity gradients).

4.5 Analysis of the Effectiveness of DoG Spectral Decomposition

Fig. 3 visualizes the four frequency bands extracted from a representative brain MRI volume. Band 0 isolates sharp cortical fold boundaries and grey-white matter interfaces at the finest spatial scale. Bands 1 and 2 reveal intermediate structural transitions, from major sulcal boundaries to subcortical contours. Band 3 captures smooth, large-scale grey/white matter contrast and global brain morphology. Together, the four bands form a complete and non-redundant decomposition whose superposition exactly reconstructs the input volume.

This visualization directly supports the motivation for our two spectral modules. Without explicit frequency separation and adaptive fusion, the dominant strength in Band 3 suppresses the fine structural signals in Band 0, leading clustering modules to conflate boundary voxels with anatomically distinct interior regions of similar low-frequency profiles. By explicitly isolating these spectral components and adaptively fusing them, SAE protects the vulnerable high-frequency boundaries right at the source, while FreqSACB performs spectral-adaptive purification to drive spatial clustering with the most scale-appropriate bands, thereby effectively eliminating this cross-frequency suppression.

5 Conclusion

We present a frequency-aware pyramid registration network that addresses the spectral interference overlooked by existing pyramid-based methods. The Spectral Adaptive Encoder and Frequency-aware Spatial Adaptive Clustering Block collectively enable spectral-adaptive processing by preventing early frequency entanglement at the initial encoding stage and enabling scale-specific spectral-adaptive purification across the multi-scale decoding phase. We further provide a theoretical proof of the existence of scale-adaptive frequency bands for each level of a Gaussian pyramid. Experiments on LPBA40 demonstrate state-of-the-art performance with a Dice of 72.73% and IoU of 57.79%, with ablation studies confirming the complementary contribution of each component.

References

1. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical image analysis **12**(1), 26–41 (2008)
2. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging **38**(8), 1788–1800 (2019)
3. Beg, M.F., Miller, M.I., Trouvé, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. International journal of computer vision **61**(2), 139–157 (2005)
4. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y.: Transmorph: Transformer for unsupervised medical image registration. Medical image analysis **82**, 102615 (2022)
5. Cheng, X., Zhang, T., Lu, W., Meng, Q., Frangi, A.F., Duan, J.: Sacb-net: spatial-awareness convolutions for medical image registration. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5227–5237. IEEE (2025)
6. Guo, T., Wang, Y., Shu, S., Yuan, W., Chen, D., Tang, Z., Meng, C., Bai, X.: Mambamorph: a mamba-based framework for medical mr-ct deformable registration. arXiv preprint arXiv:2401.13934 (2024)
7. Hu, B., Zhou, S., Xiong, Z., Wu, F.: Recursive decomposition network for deformable image registration. IEEE Journal of Biomedical and Health Informatics **26**(10), 5130–5141 (2022)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Mok, T.C., Chung, A.C.: Large deformation diffeomorphic image registration with laplacian pyramid networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. LNCS, vol. 12263, pp. 211–221. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_21
10. Oliveira, F.P., Tavares, J.M.R.: Medical image registration: a review. Computer methods in biomechanics and biomedical engineering **17**(2), 73–93 (2014)
11. Qiu, H., Qin, C., Schuh, A., Hammernik, K., Rueckert, D.: Learning diffeomorphic and modality-invariant registration using b-splines. In: Medical imaging with deep learning. ML Research Press (2021)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. LNCS, vol. 9351, pp. 234–241. Springer, cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
13. Shattuck, D.W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K.L., Poldrack, R.A., Bilder, R.M., Toga, A.W.: Construction of a 3d probabilistic atlas of human cortical structures. Neuroimage **39**(3), 1064–1080 (2008)
14. Thirion, J.P.: Image matching as a diffusion process: an analogy with maxwell's demons. Medical image analysis **2**(3), 243–260 (1998)
15. Wang, H., Ni, D., Wang, Y.: Modet: Learning deformable image registration via motion decomposition transformer. In: International conference on medical image computing and computer-assisted intervention. LNCS, vol. 14229, pp. 740–749. Springer, cham (2023). https://doi.org/10.1007/978-3-031-43999-5_70
16. Zhao, S., Dong, Y., Chang, E.I., Xu, Y., et al.: Recursive cascaded networks for unsupervised medical image registration. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10600–10610. IEEE (2019)