SACB-Net: Spatial-awareness Convolutions for Medical Image Registration

Xinxing Cheng[1], Tianyang Zhang[1], Wenqi Lu[2], Qingjie Meng[1,5],
Alejandro F. Frangi[3,4], Jinming Duan[1,3,4*]

[1] School of Computer Science, University of Birmingham, UK

[2] Department of Computing and Mathematics, Manchester Metropolitan University, UK

[3] Division of Informatics, Imaging and Data Sciences, University of Manchester, UK

[4] Centre for Computational Imaging and Modelling in Medicine, University of Manchester, UK

[5] Department of Computing, Impeiral College London, UK

{xxc142, txz009}@student.bham.ac.uk, W.Lu@mmu.ac.uk, m.qingjie@bham.ac.uk

alejandro.frangi@manchester.ac.uk, jinming.duan@manchester.ac.uk

Abstract

Deep learning-based image registration methods have shown state-of-the-art performance and rapid inference speeds. Despite these advances, many existing approaches fall short in capturing spatially varying information in non-local regions of feature maps due to the reliance on spatially-shared convolution kernels. This limitation leads to suboptimal estimation of deformation fields. In this paper, we propose a 3D Spatial-Awareness Convolution Block (SACB) to enhance the spatial information within feature representations. Our SACB estimates the spatial clusters within feature maps by leveraging feature similarity and subsequently parameterizes the adaptive convolution kernels across diverse regions. This adaptive mechanism generates the convolution kernels (weights and biases) tailored to spatial variations, thereby enabling the network to effectively capture spatially varying information. Building on SACB, we introduce a pyramid flow estimator (named SACB-Net) that integrates SACBs to facilitate multi-scale flow composition, particularly addressing large deformations. Experimental results on the brain IXI and LPBA datasets as well as Abdomen CT datasets demonstrate the effectiveness of SACB and the superiority of SACB-Net over the state-of-the-art learning-based registration methods. The code is available at `https://github.com/x-xc/SACB_Net`.

1. Introduction

Deformable image registration aims to align the corresponding anatomical structures between a moving and fixed
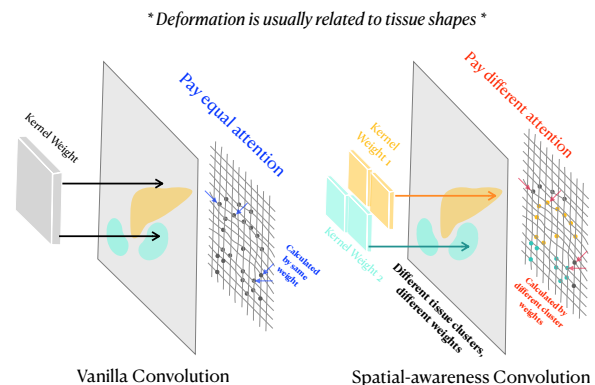
---

*Corresponding author.

Figure 1. **Spatial-awareness Convolution.** Since deformation is usually related to tissue shape in medical image registration, voxels/features from different regions should be given varying levels of attention. However, vanilla convolution methods apply shared kernel weights across all regions, leading to a suboptimal estimation of deformation fields. SAC mechanisms, on the other hand, apply distinct kernel attention weights for different spatial clusters.

image pair with the estimated dense non-linear deformation field, which plays a crucial role in various medical imaging tasks [38, 48]. In the past, conventional methods like LDDMM [6], DARTEL [1], SyN [2], Demons [45], and ADMM [42] are time-consuming and computationally intensive, primarily due to the instance-level (pair-wise) iterative optimization required. Moreover, such methods may involve sophisticated hyper-parameter tuning, limiting their applications in large-scale volumetric registration. To overcome these challenges, learning-based image registration approaches have demonstrated significant success in achieving competitive accuracy along with fast inference speed. Learning-based approaches can be divided into supervised and unsupervised categories. Supervised methods

[7, 34, 49] often rely on the high quality of the synthetic ground truth deformation field or the annotated map [35] to predict a deformation field. Therefore, unsupervised image registration is more popular for estimating an optimal deformation field of pairwise images by minimizing the distance between the warped moving image and the fixed image with a regularization term.

Following the classical unsupervised learning-based framework VoxelMorph [5], some U-Net-based methods focus on the performance improvements by adding stronger constraints over the deformation field such as inverse/cycle consistency [24, 50], gradient inverse consistency [43] and diffeomorphisms [13, 14, 29]. Some existing methods have introduced more advanced neural architectures, such as large-kernel network [20], vision transformers [8, 10, 17, 30, 37, 51]. To enhance the efficiency of the U-Net architecture and accelerate inference speed, several model-driven methods, including B-Spline [32], Fourier-Net [21] and WiNet [12], have been proposed to learn a low-dimensional representation of displacement. The lightweight approach, NCA-Morph [33], utilizes Neural Cellular Automata to improve the efficiency of image registration. Another track of recent works tends to progressively estimate the large and complex deformations using either multiple cascades [15, 31, 52, 53] or pyramid methods [18, 23, 25, 26, 28, 46] compositions of multi-scale flows derived from multi-scale features. The cascaded methods and pyramid coarse-to-fine methods are capable of effectively improving registration accuracy for estimating large deformation.

However, existing learning-based methods typically rely on networks that learn local feature correlations by applying shared convolution across all spatial dimensions within each layer. As illustrated in Fig. 1, deformations are often linked to the morphological characteristics of tissue in medical image registration, where voxels/features from various anatomical regions need to be paid different attention. However, plain convolutional methods utilize shared kernel weights uniformly across all regions, resulting in suboptimal estimation of deformation fields. Specifically, 3D medical images encapsulate complex intensity information relevant to anatomy, which makes it hard to capture spatially varying information in non-local regions of feature maps via spatially-shared convolution kernels. This key challenge is effectively incorporating spatially adaptive information, which can be addressed using region-based adaptive convolution techniques. Inspired by [16], which explores content-adaptive convolution in 2D image pansharpening, we extend this approach to 3D for medical image registration. This adaptation improves spatial awareness by modifying convolutional kernels based on regional content, thereby enhancing registration accuracy.

In this paper, we propose a novel 3D pyramid medical image registration network, named SACB-Net, which in-

tegrates spatial-awareness convolution block (SACB) with a similarity matching module to form a multi-scale flow estimator. The proposed SACB employs the unsupervised KMeans to cluster the learned feature maps. Each feature clustering is believed to capture a specific intensity or structural pattern of the feature maps and then subsequentially used to learn the spatial-aware adaptive convolution kernels. The adaptive kernel generation in SACB allows the network to effectively learn refined feature representations for more accurate flow estimation. We summarize our main contributions as follows:

- We consider the spatial variance between two input images in the context of image registration, which is overlooked by vanilla convolutional-based approaches. The adaptive convolutions in our SACB-Net consider the changing of spatial regions to enhance registration accuracy. To the best of our knowledge, our method is the first work that employs 3D spatial-awareness convolution in 3D medical registration.

- The proposed SACB aims to boost the moving and fixed feature maps using adaptive convolution kernels for different spatial regions, where the spatial regions are divided based on the clustering of features, and the convolution kernels are learned from the centroids of each feature clustering.

- We propose an innovative pyramid flow estimator that integrates SACB and similarity matching module to perform multi-scale flow estimation, which can not only estimate local small deformation but also handle large deformation.

- Experimental results on atlas-based and inter-subject registration using two brain datasets (IXI and LPBA) demonstrate the effectiveness of SACB-Net. Further evaluation on inter-subject abdominal CT registration shows improved performance in handling large deformations.

2. Related Works

2.1. U-Net-based Registration Method

VoxelMorph [4, 5], a general framework for learning-based image registration, demonstrates the effectiveness of unsupervised learning in pairwise image registration using a simplified U-Net architecture. Building on this foundation, recent methods have enhanced registration performance by incorporating stronger constraints on the deformation field, such as inverse/cycle consistency [24, 50], gradient inverse consistency [43], and diffeomorphism constraints [13, 14, 29]. Another track of methods tried to capture long-range dependencies of feature correlations using either large kernel convolution [20] or vision transformer [8, 10, 17, 30, 37, 51]. TransMorph [10] incorporates Swin transformer blocks in the encoder, improving the network's capacity to capture long-range dependencies, though this

comes with increased computational cost. LKU [20], on the other hand, employs a large kernel convolutional encoder to model long-range relationships with fewer parameters and reduced computational load. NCA-Morph [33] employs Neural Cellular Automata mechanisms, drawing inspiration from biological processes, to improve registration efficiency. While U-Net-based methods have shown comparable performance for small deformations, they struggle to accurately estimate large deformations or register images that are not well pre-aligned.

2.2. Model-driven Registration Method

To reduce redundant convolution operations in the U-Net decoder and enhance training and inference efficiency, model-driven approaches such as B-Spline [32] and Fourier-Net [21] focus on learning a low-dimensional representation of the displacement field. However, displacement fields are not inherently smooth, and as a result, methods that rely solely on low-frequency components [21] or interpolate displacement from regular control points [32] may fail to fully and accurately capture the entire displacement field. To address this limitation, WiNet [12] introduces a wavelet-based decoder to preserve high-frequency information in the deformation field while maintaining network efficiency. Nonetheless, similar to U-Net-based methods, model-driven approaches also face challenges in effectively estimating large and complex deformations.

2.3. Cascaded and Pyramid-based Method

Recent works that estimate large deformation can be either network cascade [15, 31, 52, 53] or coarse-to-fine pyramid [18, 23, 25, 26, 28, 46]. LapIRN [31] utilizes a Laplacian pyramid network to capture large deformations by cascading registered flows at three different scales. PRNet++ [23] utilizes a dual-stream pyramid architecture for coarse-to-fine flow estimation, performing sequential warping on multi-scale feature maps and incorporating local 3D correlation layers. Im2grid [25] embedded the coordinate information on extracted multi-scale feature maps and estimated the scale flow by calculating the matching score on the neighborhood feature. ModeT [46] leverages neighbourhood attention mechanisms to estimate multi-head, multi-scale sub-deformation flows from pyramid feature maps. It then applies weighting modules to fuse these multi-head flows at each scale, generating the final deformation by sequential composition. To solve the insufficient decomposition problems, RDN [18] compose the multi-scale flows by performing level-wise recursion at each resolution level and stage-wise recursion at feature pyramid level.

2.4. Adaptive Convolution

Unlike standard convolution operators that use shared convolution kernels, adaptive convolution operators apply different kernels based on variations in the input, providing enhanced feature extraction capabilities and greater flexibility. Following [19], recent methods have applied adaptive convolution, demonstrating its effectiveness in various 2D computer vision tasks, such as image upsampling [39], classification [11, 54], segmentation [9] and remote sensing pansharpening [16, 22]. However, spatially adaptive convolution has not yet been explored in 3D medical image registration tasks.

3. Method

3.1. Network Overview

Given a pair of moving image I_m and fixed image I_f over the spatial domain $\Omega \subset \mathcal{R}^3$, the SACB-Net$(I_m, I_f; \boldsymbol{\theta})$ network aims to learn the optimal parameters $\boldsymbol{\theta}_*$ that estimate a deformation field $\phi_{m \to f} : \mathcal{R}^3 \to \mathcal{R}^3$, mapping the corresponding coordinates between the paired images. In unsupervised setting, the loss function contains $\mathcal{L}_{sim}(\cdot)$ measuring the similarity between the warped moving image $I_m \circ \phi$ and I_f, and the regularization term. In our experiments, the smoothness term $\mathcal{L}_{reg}(\phi) = \|\nabla \phi(\boldsymbol{\theta})\|_2^2$ is computed using the L_2-norm of the gradient of the deformation field. The objective function of our network is

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_{sim}(I_m \circ (\phi(\boldsymbol{\theta}) + \mathrm{Id}), I_f) + \lambda \mathcal{L}_{reg} \quad (1)$$

where $\boldsymbol{\theta}$ denotes the learnable network parameters, $\circ$ represents the warping operator, Id is the identity grid, ∇ refers to the first-order gradient, and λ is the weight of the regularization term. We employ Normalized Cross-Correlation (NCC) to measure the similarity.

As illustrated in Fig. 2, SACB-Net consists of a shared encoder and a pyramid flow estimator. The shared encoder is designed to extract multi-scale features $\{F_m^i\}$ and $\{F_f^i\}$ for the moving image I_m and the fixed image I_f, respectively, in a pyramid structure. Here, $i \in \{1, 2, 3, 4, 5\}$ represents the scale of the features, with resolutions of $\left[\frac{D}{2^{i-1}}, \frac{H}{2^{i-1}}, \frac{W}{2^{i-1}}\right]$ for the depth, height and width. The shared encoder consists of five 3D convolutional layers and four down-sampling layers. The convoluational layer is composed of 3D convolution operation, instance normalization [44] and LeakyReLU activations [27] with the 0.1 slope. Each convolutional layer is followed by an average pooling operation with a kernel size of 2. The extracted features $\{F_m^i\}$ and $\{F_f^i\}$ are then input into our pyramid flow estimators, where they are used to perform flow estimation and composition as outlined in Sections 3.2 and 3.3.

3.2. 3D Spatial-awareness Convolution Block

To effectively extract region-based context information for flow estimation, the proposed 3D SAC incorporates prior knowledge of spatial context through a spatial context estimation module and adaptive kernel generator. First, the
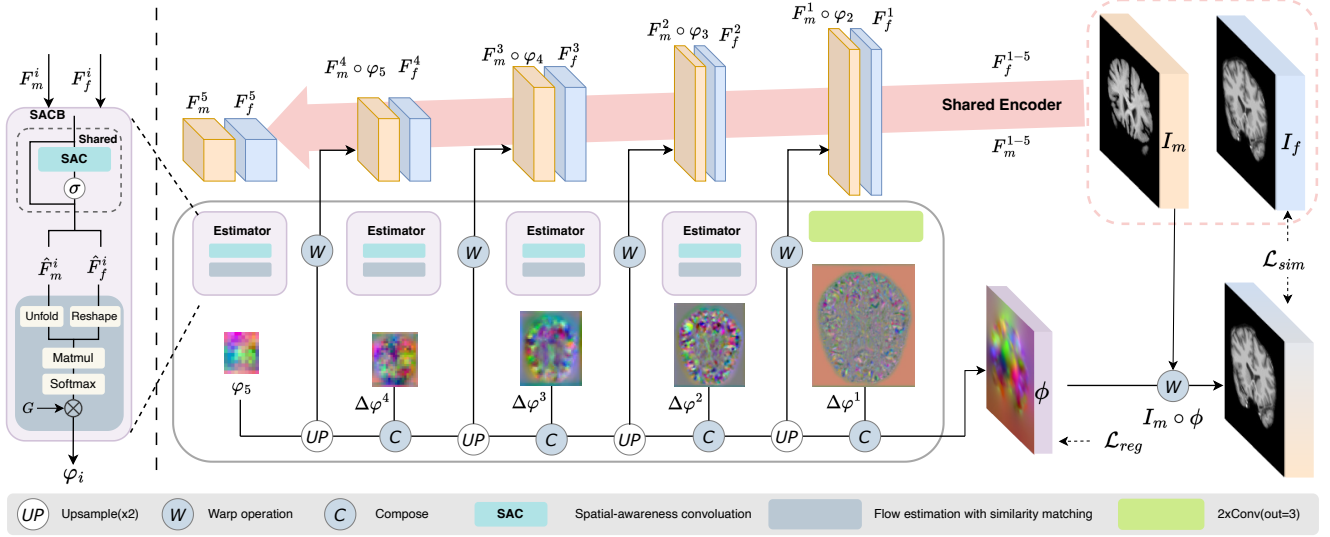
Figure 2. Illustration of a 5-level pyramid SACB-Net. SACB-Net includes a shared encoder that extracts multi-scale feature maps $\{F_m^i\}$ and $\{F_f^i\}$ for the moving image I_m and the fixed image I_f, as well as pyramid flow estimators at each scale. At the lowest level, the flow estimator learns deformation (φ_5) from the extracted moving and fixed image features (F_m^5 and F_f^5). The following flow estimators take the level-wise features and the output deformation from its preceding level to compose the deformation. Each flow estimator includes a Spatial-Awareness Convolution Block (SACB) to enhance spatially adaptive feature representation, along with a similarity matching module for flow estimation.

spatial context estimation module identifies spatial regions of interest, producing spatial contextual information. This contextual data is subsequently utilized by the adaptive kernel generator to dynamically produce convolutional kernels tailored to capture non-local, region-specific information.

Spatial Context Estimation. After obtaining the multi-scale features $\{F_m^i\}$ and $\{F_f^i\}$, prior spatial region information is incorporated to refine these features. To achieve this, we employ unsupervised clustering method K-Means to estimate spatial region for feature maps. The reason not to use label information is due to potential inconsistencies in the labeling of I_m and I_f within the feature space, as well as to address the challenge of limited label availability. Let $\mathbf{F} \in \mathbb{R}^{C_{in} \times D \times H \times W}$ denote the input feature map, where C_{in} represents the number of input channels, and D, H, and W correspond to the depth, height, and width of the spatial dimensions, respectively. As shown in Fig. 3, first, the feature map $\mathbf{F}$ is unfolded into local patches to capture neighborhood information using sliding windows of size k. A spatial mean is then applied to each patch to reduce its spatial dimensions. For each voxel $V_{(d,h,w)}$ located at coordinates (d,h,w) within $\mathbf{F}$, suppose $p(\cdot)$ is the unfolding operation, the corresponding unfolded patch is defined as:

$$p_{(V_{(d,h,w)})} = \left\{ V_{(d+d',h+h',w+w')} \mid d',h',w' \in \left[-\frac{k}{2}, \frac{k}{2}\right] \right\}. \quad (2)$$

Here, $p_{(V_{(d,h,w)})}$ represents a patch of voxels centered at $V_{(d,h,w)}$, with offsets d',h',w' that range from $-\frac{k}{2}$ to $\frac{k}{2}$ (assuming k is odd). The unfolded feature denotes as $p(F) \in \mathbb{R}^{C_{in} \times (D \times H \times W) \times k^3}$. The spatial mean of each local patch $\overline{p}_{(V_{(d,h,w)})}$ [1] based on the neighborhood voxel is calculated as:

$$\overline{p}_{(V_{(d,h,w)})} = \frac{1}{k^3} \sum_{d'=-\lfloor\frac{k}{2}\rfloor}^{\lfloor\frac{k}{2}\rfloor} \sum_{h'=-\lfloor\frac{k}{2}\rfloor}^{\lfloor\frac{k}{2}\rfloor} \sum_{w'=-\lfloor\frac{k}{2}\rfloor}^{\lfloor\frac{k}{2}\rfloor} p_{(V_{(d,h,w)})}. \quad (3)$$

Afterward, the precessed feature $\mathbf{F}$ is reshaped to dimensions $[D \times H \times W, C_{in}]$ to estimate a cluster index matrix $S \in \mathbb{Z}^{D \times H \times W}$, which assigns each voxel to a cluster $S_n = \{(d,w,h) \mid S_{(d,w,h)} = n\}; n \in \{1, \cdots, N\}$. Here, N represents the maximum number of clusters used for GPU-based KMeans clustering. Additionally, the centroid voxel of each cluster $S_n^c \in \mathbb{R}^{C_{in}}$ is computed as

$$S_n^c = \frac{1}{|S_n|} \sum_{(d,w,h) \in S_n} V_{(d,w,h)}, \quad (4)$$

where $|S_n|$ denotes the total number of voxels belonging to the nth spatial region. The cluster index matrix S and the centroid voxels S^c for each cluster are then utilized to generate adaptive kernel weights for convolution.

Adaptive Kernel Generator. To adaptively generate a spatial convolution kernel that incorporates region-based information, we employ a multi-layer perception to model the

---

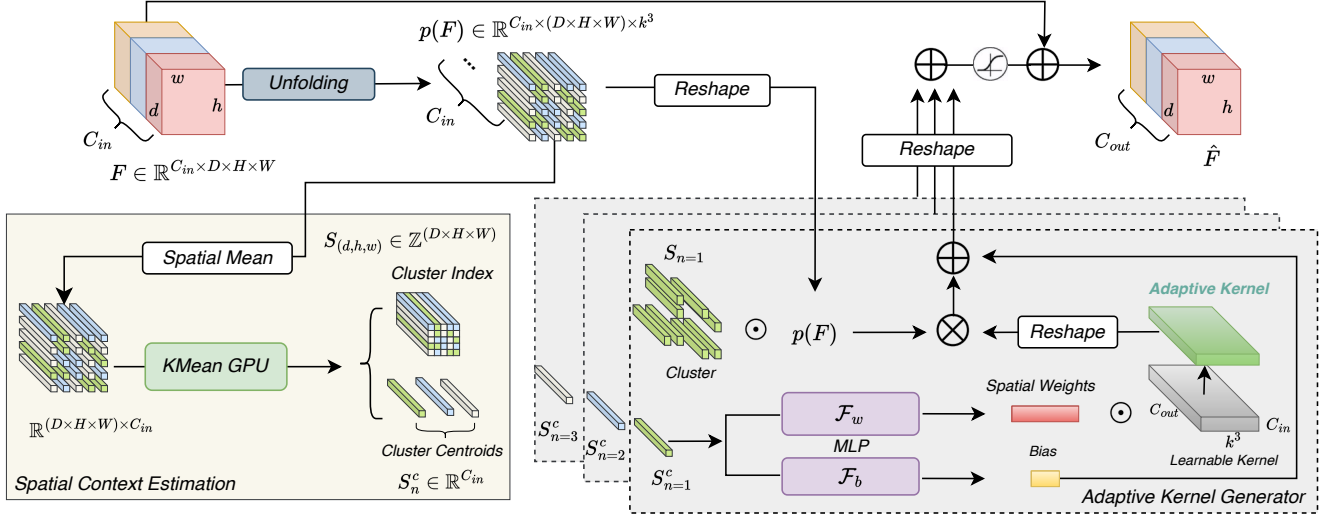[1] The choices of $\overline{p}$ are investigated in Sec. 6.

Figure 3. Architecture of the 3D spatial-awareness convolution block (example in three clusters). This block aims to refine the input feature $\mathbf{F}$ to $\hat{\mathbf{F}}$ by adaptive convolution learned from spatial feature clustering. SACB consists of three parts: 1) the spatial context estimation module employs KMeans to cluster similar spatial features on the patched unfolding features, in which the features belonging to the same cluster centroid S_n^c will be indicated with a cluster index map S_n. 2) the adaptive kernel generator leverages each cluster centroid to generate cluster-specific spatial weights and bias via two MLPs, the resulting weights and bias will be used to form the final spatial adaptive convolution kernel, which will be imposed on the cluster-indexed unfolding features, again highlighting the spatial-awareness. 3) the original features and spatial-aware features are residually connected to form the final refined $\hat{\mathbf{F}}$.

mapping $\mathcal{F}_w$, which produces an adaptive convolution kernel $\mathbf{W}_n \in \mathbb{R}^{C_{out} \times C_{in} \times k^3}$ for the nth spatial region based on S_n^c. To simplify the process, we first initialize a global learnable convolution kernel $\mathbf{W} \in \mathbb{R}^{C_{out} \times C_{in} \times k^3}$ and then use the mapping function $\mathcal{F}_w$ to generate a spatial weight that can be applied to the global kernel $\mathbf{W}$. Thus, $\mathbf{W}_n$ is obtained as

$$\mathbf{W}_n = \mathcal{F}_w(S_n^c) \odot \mathbf{W}, \qquad (5)$$

where $\odot$ denotes element-wise dot product. Then, a mapping function $\mathcal{F}_b$ using a multi-layer perception is applied to compute the convolution bias via obtained centroid S_n^c. Consequently, the region-based spatially adaptive convolution is represented as:

$$SAC(\mathbf{F}_{(d,w,h)}) = p_{(V_{(d,h,w)})} \otimes \mathbf{W}_n + \mathcal{F}_b(S_n^c), \qquad (6)$$

where $p(\cdot)$ denotes the unfolding operation outlined in the spatial context estimation section. Subsequently, the process of applying the SAC block to produce refined features is as follows

$$\hat{\mathbf{F}} = SACB(\mathbf{F}) = \mathbf{F} + \sigma(SAC(\mathbf{F})), \qquad (7)$$

where $\sigma(\cdot)$ denotes activation function.

3.3. Pyramid Flow Estimator

Let $\hat{F}_m^{\,i}$ and $\hat{F}_f^{\,i}$ represent the multi-scale features refined by the proposed SACB at the ith level, the sub-deformation

flows at each scale are then estimated based on the matching score using the refined features. Specifically, we calculate the similarity matching score between each voxel in the fixed image features and corresponding neighbouring voxels in the moving image features. Similar to [25, 40], this operation can be efficiently implemented as a matrix operation using the dot product. For instance, at the coarsest level $i = 5$, the matching score can be computed as

$$M_{sim}(\hat{F}_f^{\,5}, \hat{F}_m^{\,5}) = \text{Softmax}(\hat{F}_f^{\,5} p(\hat{F}_m^{\,5})^T), \qquad (8)$$

where $p(\cdot)$ represents the unfolding operation as described in Section 3.2. For each voxel V at $\hat{F}_m^{\,5}$, the similarity matching score can be calculated as

$$M_{sim}(\hat{F}_f^{\,5}(V), \hat{F}_m^{\,5}(V)) = \qquad (9)$$
$$\text{Softmax}\left(\left\{\left\langle \hat{F}_f^{\,5}(V), \hat{F}_m^{\,5}(V_c)\right\rangle\right\}_{V_c \in \mathcal{N}(V)}\right),$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, and $\mathcal{N}(V)$ represents the set of voxels within an $k \times k \times k$ neighborhood centered around voxel V, k represents the window size of the search region of $\hat{F}_m^5$. We set $k = 3$ in our experiments. Then, the sub-deformation flow at the 5th scale can be obtained as

$$\varphi^5 = M_{sim}(\hat{F}_f^{\,5}, \hat{F}_m^{\,5})G \qquad (10)$$

$G \in \mathbb{R}^{(k \times k \times k) \times 3}$ is a 3D grid that encodes the relative positions with respect to the centroid voxel location.

At the $i = 4$ level, we first up scaling the flow φ^5 to form $\hat{\varphi}^5$, then $\hat{F_m}^4$ is warpped with $\hat{\varphi}^5$. Sequentially, the M_{sim} need to calculated between $\hat{F_f}^4$ and $\hat{F_m}^4 \circ \hat{\varphi}^5$ for obtaining $\Delta\varphi^4$. Afterward, the overall process can be formulated as:

$$\begin{cases} \hat{\varphi}^i = \text{up}_{2\times}(\varphi^i), \\ \hat{F_f}^{i-1} = SACB(F_f^{i-1}), \\ \hat{F_m}^{i-1} = SACB(F_m^{i-1} \circ \hat{\varphi}^i), \qquad i \in \{5, 4, 3, 2\}. \\ \Delta\varphi^{i-1} = M_{sim}(\hat{F_f}^{i-1}, \hat{F_m}^{i-1})G, \\ \varphi^{i-1} = \hat{\varphi}^i \circ \Delta\varphi^{i-1}, \end{cases}$$
(11)

The $\text{up}_{2\times}(\cdot)$ is flow rescaling operation with $2 \times$ resolution.

We note that the matching score calculation was not applied for flow estimation at the 1st scale, as the search region with $k = 3$ can only determine a maximum displacement of 1 voxel using full-resolution features. However, the calculations at the 1st scale significantly increase memory and computational costs, yielding limited benefit. Therefore, we used a simple approach with two convolutional layers for flow estimation at the final scale.

4. Experiments

Datasets. Our experiments use two publicly available 3D brain MRI datasets IXI[2] and LPBA [36], along with one public abdominal CT dataset[3] [47]. The preprocessed IXI dataset [10] includes 576 MRI scans (160×192×224) from healthy subjects. We followed the same protocol as [10], partitioning the data into 403, 58, and 115 scans for training, validation, and testing sets, respectively. Atlas-based brain registration was conducted using a generated atlas from [24]. The LPBA dataset consists of 40 T1-weighted MRI volumes. Each MRI scan has been rigidly pre-aligned to the MNI305 template and manually annotated with 54 regions of interest (ROIs). We center-crop all volumes to dimensions of $160 \times 192 \times 160$ with voxel size $1 \times 1 \times 1mm^3$. LPBA dataset was used to perform inter-subject registration with divided three subsets: 25 volumes (25×24 pairs) for training, 5 (5×4 pairs) for validation, and 10 (10×9 pairs) for testing. The Abdomen CT dataset consists of 30 abdominal scans with 13 manually labelled anatomical structures. We clamp the Hounsfield Units (HU) to $[-1000, 1000]$ and normalize each scan to $[0, 1]$. Each scan is center-cropped from $192 \times 160 \times 256$ to $192 \times 160 \times 224$ with voxel size $2 \times 2 \times 2$ mm^3. The dataset is divided into 19 training volumes (19×18 pairs), 3 validation volumes (3×2 pairs) and 8 testing volumes (8×7 pairs) for inter-subject registration.

Implementation Details. Our method is implemented using PyTorch. Both the training and testing phases are deployed on an A100 GPU with 40GB VRAM. All models

are optimized using Adam followed with the same learning rate 10^{-4} and batch size 1. A fast GPU K-Means implementation[4] is adopted.

Evaluation Metrics. To qualify the registration performance, we use the Dice similarity coefficient (Dice) to evaluate the region of interest (ROI) segmentation overlap, the average symmetric surface distance (ASSD) [41] and the 95 percentile of the Hausdorff distance (HD95) to measure the registration accuracy of ROI structures. For evaluating the diffeomorphism of deformation fields, we use the percentage of negative values of the Jacobian determinant ($|J|_{<0}\%$) to calculate the percentage of folding voxels.

Comparative Methods. We compare our method against state-of-the-art approaches, including the conventional SyN method [2], U-Net-based models such as VoxelMorph (VM) [5], LKU [20], and TransMorph [10], model-driven methods like B-Spline-Diff [32] and Fourier-Net [21], the lightweight NCA-Morph [33], the cascaded method LapIRN [31], and four pyramid-based methods: PR++ [23], Im2Grid [25], RDN [18] and ModeT [46]. All models were trained using their officially released code, with optimal hyperparameters tuned on the validation sets.

5. Results

In this section, we evaluate our SACB-Net against comparative methods on three tasks: atlas-based brain registration using the IXI dataset, inter-subject brain registration using the LPBA dataset, and large deformation estimation for inter-subject abdominal CT registration.

5.1. Atlas-based Brain Registration

Table 1 presents the quantitative results of atlas-based brain registration for 30 anatomical structures evaluated on the IXI dataset. It can be seen that SACB-Net outperforms all comparable methods in terms of Dice score while achieving plausible deformations and third-best results on HD95 and ASSD. Compared to the best-performing U-Net-based method (LKU), our SACB-Net achieves a 0.4% improvement in Dice score while using 0.98M fewer parameters. In terms of model-driven methods, SACB-Net surpasses Fourier-Net by a margin of 0.6% in the Dice score and uses 3.09M fewer parameters. For cascade and pyramid-based methods, those methods have a similar number of parameters apart from RDN. Our SACB-Net outperforms the second-best cascade method LapIRN with 0.6% Dice, 0.038mm HD95, and 0.019mm ASSD, respectively.

5.2. Inter-subject Brain Registration

The inter-subject registration of the LPBA dataset tends to have a slightly larger deformation than the atlas-based registration on the IXI dataset. As shown in Table 1, most UNet-

Table 1. Registration performance comparison on the IXI and LPBA datasets.

Method	IXI (30 ROIs)				LPBA (54 ROIs)				# Param
	Dice↑	HD95↓	ASSD↓	$|J|_{<0}\%\downarrow$	Dice↑	HD95↓	ASSD↓	$|J|_{<0}\%\downarrow$	
Affine	0.386±0.195	6.479±0.666	2.445±0.280	-	0.525±0.047	8.039±0.861	2.586±0.350	-	-
SyN [3]	0.645±0.152	6.394±1.048	1.551±0.286	<0.0001	0.707±0.016	6.254±0.444	1.479±0.131	<0.0001	-
VM-1 [5]	0.729±0.129	3.798±0.757	0.937±0.162	1.590±0.339	0.664±0.025	6.873±0.654	1.717±0.200	0.649±0.261	0.27M
VM-2 [5]	0.732±0.123	3.723±0.680	0.926±0.158	1.522±0.336	0.669±0.025	6.847±0.659	1.698±0.200	0.591±0.242	0.30M
NCA-Morph [33]	0.753±0.136	3.109±0.525	0.796±0.121	0.506±0.190	0.679±0.023	6.666±0.634	1.631±0.188	0.264±0.121	0.37M
TransMorph [10]	0.754±0.124	3.543±0.721	0.862±0.168	1.579±0.328	0.695±0.022	6.564±0.619	1.559±0.182	0.474±0.176	46.77M
LKU [20]	0.765±0.129	2.967±0.494	0.757±0.114	0.109±0.054	0.706±0.032	6.452±0.951	1.603±0.250	0.594±0.203	2.09M
B-Spline-Diff [32]	0.742±0.128	3.256±0.538	0.832±0.117	<0.0001	0.665±0.023	6.792±0.622	1.713±0.186	**0.0±0.0**	0.27M
Fourier-Net [21]	0.763±0.129	**2.857±0.456**	**0.748±0.114**	0.024±0.019	0.672±0.022	6.716±0.601	1.666±0.180	0.216±0.104	4.20M
LapIRN [31]	0.763±0.133	3.166±0.608	0.779±0.126	0.312±0.106	0.716±0.016	6.116±0.454	1.426±0.133	0.024±0.009	1.20M
PRNet++ [23]	0.755±0.130	3.593±0.748	0.857±0.162	1.052±0.302	0.701±0.021	6.492±0.597	1.520±0.177	0.072±0.027	1.24M
ModeT [46]	0.758±0.125	3.496±0.732	0.828±0.151	0.114±0.057	0.721±0.013	5.969±0.416	1.375±0.110	0.010±0.004	1.03M
Im2Grid [25]	0.761±0.127	3.316±0.668	0.799±0.128	<0.0002	0.713±0.014	6.062±0.428	1.419±0.118	0.007±0.003	0.89M
RDN [18]	0.759±0.123	3.476±0.802	0.823±0.161	<0.0001	0.713±0.017	6.208±0.497	1.436±0.142	<0.0002	28.65M
Ours	**0.769±0.127**	3.128±0.631	0.760±0.125	0.083±0.045	**0.731±0.012**	**5.862±0.436**	**1.326±0.114**	0.018±0.006	1.11M

based and model-driven methods do not perform as well as cascade and pyramid-based methods. Only LKU surpasses the pyramid-based method PR++, achieving improvements of 0.5% in Dice, 0.04 mm in HD95 and 0.083 mm in ASSD, respectively. Our SACB outperforms all comparable methods in Dice, HD95 and ASSD with plausible deformations. Compared to the second-best method, ModeT, SACB achieves improvements of 1% in Dice, 0.107mm in HD95, and 0.049mm in ASSD, respectively. The visualization in Fig. 4 shows that our SACB-Net captures more detailed correspondences in the registration process, as highlighted in the red box.

5.3. Inter-subject Abdomen Registration

The primary challenge in abdominal CT registration tasks is caused by the substantial differences in the spatial distribution of organs across scans, making it difficult to accurately estimate large deformations during the registration process. Table 2 demonstrates the numerical results of the inter-subject abdomen registration task. Our results show that our method outperforms all compared approaches in terms of Dice, HD95, and ASSD. Notably, U-Net-based and model-driven methods perform inferior to pyramid and cascaded approaches for estimating large deformations. Even the best-performing method in such approaches, TransMorph, still exhibits a significant gap compared to our SACB-Net with 14.4% difference in Dice score, 5.934mm in HD95 and 3.153mm in ASSD, respectively. Compared to the second-best pyramid method, ModeT, SACB-Net achieves improvements of 3.8% in Dice, 2.098 mm in HD95, and 0.781 mm in ASSD, respectively.

The visualization of the wrapped moving images and labels, as shown in Fig. 4, illustrates that most parts of the kidneys are registered with our method. However, Trans-

Table 2. Registration performance comparison on Abdomen CT.

Method	Abdomen CT (13 ROIs)			
	Dice↑	HD95↓	ASSD↓	$|J|_{<0}\%\downarrow$
Affine	0.305±0.047	27.790±4.868	10.869±2.075	-
VM-2 [5]	0.406±0.058	25.226±4.942	8.890±2.066	3.882±1.512
NCA-Morph [33]	0.406±0.061	25.959±4.982	9.083±2.223	5.197±1.166
TransMorph [10]	0.444±0.065	24.187±4.955	8.156±2.065	6.045±1.750
LKU [20]	0.423±0.058	24.252±4.604	8.395±1.946	4.316±1.778
B-Spline-Diff [32]	0.375±0.053	25.338±4.866	9.217±1.960	**0.0±0.0**
Fourier-Net [21]	0.417±0.048	23.084±4.201	8.000±1.532	2.167±2.391
LapIRN [31]	0.522±0.057	20.728±4.318	6.274±1.379	3.203±0.848
PRNet++ [23]	0.478±0.060	23.791±4.691	7.524±1.771	2.252±1.192
ModeT [46]	0.550±0.054	20.351±4.231	5.784±1.236	2.239±0.906
Im2Grid [25]	0.530±0.047	20.763±3.786	6.028±1.146	2.421±0.653
RDN [18]	0.527±0.055	21.782±4.722	6.461±1.613	1.529±0.989
Ours	**0.588±0.049**	**18.253±3.610**	**5.003±1.011**	3.930±1.416

Morph, PR++, and RDN failed to register the kidneys, leaving a large portion missing on the warped mask.

6. Ablation Study

First, we note that $\overline{p}(\cdot)$ in Eq. (3) can also be defined on the channel dimension or both the spatial+channel (Mix) dimension. As shown in Table 4, using the spatial mean of unfolded patches achieves slightly better performance compared to the channel mean and the Mix setting.

We then determine the optimal number of scales at which SACB should be applied and the ideal number of clusters to achieve accurate registration performance. As shown in Table 3, the optimal configuration is applying SACB from $scale_5$ to $scale_2$. Additionally, increasing the N from 5 to 7 boosts the performance while a too-large value slightly decreases performance (e.g. 11).
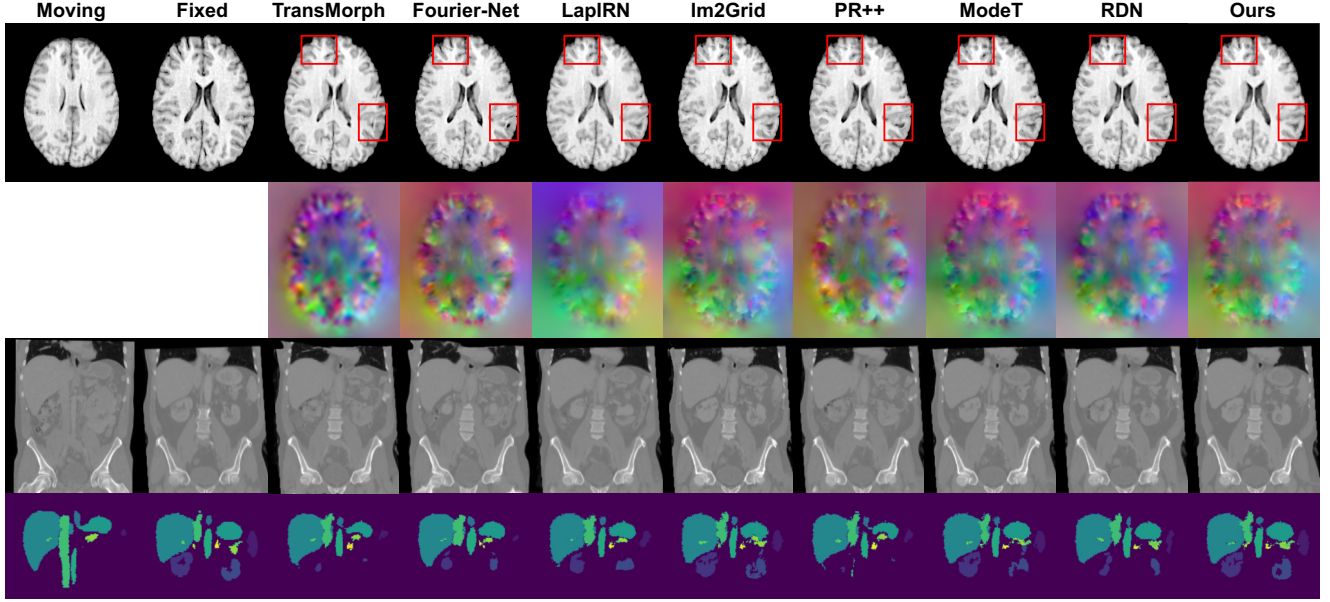
Figure 4. Visual comparisons on the Brain LPBA dataset (top two rows) and the Abdomen CT dataset (bottom two rows). Columns 3–10 display warped moving images (row 1 and 3), displacement fields as RGB images (row 2) and warped segmentation masks (row 4).

Table 3. Ablation studies on the use of SACB at different scale levels in SACB-Net and varying numbers of spatial clusters (N), along with the corresponding average training time (in seconds), including data loading time, and GPU training memory usage (in MiB). The SACB configurations are based on the best results shown in Table 4.

	With SACB				IXI				LPBA				Abdomen CT			
N	$Scale_5$	$Scale_4$	$Scale_3$	$Scale_2$	Dice ↑	$|J|_{<0}\%$ ↓	Memory	Training	Dice ↑	$|J|_{<0}\%$ ↓	Memory	Training	Dice ↑	$|J|_{<0}\%$ ↓	Memory	Training
–	✗	✗	✗	✗	0.7643	0.059	15838	0.762	0.7141	0.020	11574	0.545	0.5374	2.907	15838	0.764
5	✓	✗	✗	✗	0.7668	0.072	15880	0.799	0.7217	0.087	11664	0.579	0.5444	3.243	15880	0.797
5	✓	✓	✗	✗	0.7679	0.076	16104	0.879	0.7241	0.104	11770	0.631	0.5498	2.619	16104	0.864
5	✓	✓	✓	✗	0.7671	0.060	16960	0.945	0.7266	0.076	12418	0.714	0.5685	2.781	16960	0.950
5	✓	✓	✓	✓	0.7683	0.086	21638	1.255	0.7294	0.017	15748	0.918	0.5849	3.548	21638	1.234
7	✓	✓	✓	✓	0.7691	0.083	21638	1.376	0.7309	0.018	15778	1.084	0.5881	3.930	21662	1.360
9	✓	✓	✓	✓	0.7685	0.069	21680	1.550	0.7304	0.016	15792	1.110	0.5875	4.405	21680	1.562
11	✓	✓	✓	✓	0.7684	0.084	21732	1.679	0.7300	0.015	15806	1.193	0.5875	4.419	21718	1.728

Table 4. Ablations of the SACB configurations. After padding and unfolding on the feature map $\mathbf{F}$, the cluster maps can be calculated based on spatial mean on the local unfolded patch, channel mean and Mix (spatial+channel). In this study, we set $N = 5$ and apply SACBs from $scale_5$ to $scale_2$.

SACB			LPBA	
Spatial	Channel	Mix	Dice ↑	$|J|_{<0}\%$ ↓
✓	✗	✗	0.7294	0.017
✗	✓	✗	0.7281	0.020
✗	✗	✓	0.7285	0.021

7. Conclusion

We introduced SACB-Net, which incorporates a 3D Spatial Awareness Convolution Block (SACB) that adaptively generates spatial convolution kernels based on the feature clusters, thereby improving the network's ability to capture spatially varying information. Additionally, our proposed flow estimators combine SACB with a similarity-matching mechanism to estimate displacements in a pyramid coarse-to-fine manner. Results on two brain datasets and one abdomen dataset demonstrate that SACB-Net effectively handles both small local and large deformations. Furthermore, the proposed flow estimator is a plug-in module that can be applied in other networks for better feature representation learning.

References

[1] John Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007. 1

[2] 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. *Med Image Anal*, 12(1):26–41, 2008. 1, 6

[3] Brian B Avants, Nicholas J Tustison, Gang Song, Philip A Cook, Arno Klein, and James C Gee. A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, 2011. 7

[4] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9252–9260, 2018. 2

[5] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019. 2, 6, 7

[6] M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61:139–157, 2005. 1

[7] Xiaohuan Cao, Jianhua Yang, Jun Zhang, Qian Wang, Pew-Thian Yap, and Dinggang Shen. Deformable image registration using a cue-aware deep regression network. *IEEE Transactions on Biomedical Engineering*, 65(9):1900–1911, 2018. 2

[8] Junyu Chen, Yufan He, Eric C Frey, Ye Li, and Yong Du. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. *arXiv preprint arXiv:2104.06468*, 2021. 2

[9] Jin Chen, Xijun Wang, Zichao Guo, Xiangyu Zhang, and Jian Sun. Dynamic region-aware convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8064–8073, 2021. 3

[10] Junyu Chen, Eric C Frey, Yufan He, William P Segars, Ye Li, and Yong Du. Transmorph: Transformer for unsupervised medical image registration. *Medical image analysis*, 82:102615, 2022. 2, 6, 7

[11] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11030–11039, 2020. 3

[12] Xinxing Cheng, Xi Jia, Wenqi Lu, Qiufu Li, Linlin Shen, Alexander Krull, and Jinming Duan. Winet: Wavelet-based incremental learning for efficient medical image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 761–771. Springer, 2024. 2, 3

[13] Adrian Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical image analysis*, 57:226–236, 2019. 2

[14] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 729–738. Springer, 2018. 2

[15] Bob D De Vos, Floris F Berendsen, Max A Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52:128–143, 2019. 2, 3

[16] Yule Duan, Xiao Wu, Haoyu Deng, and Liang-Jian Deng. Content-adaptive non-local convolution for remote sensing pansharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27738–27747, 2024. 2, 3

[17] Morteza Ghahremani, Mohammad Khateri, Bailiang Jian, Benedikt Wiestler, Ehsan Adeli, and Christian Wachinger. H-vit: A hierarchical vision transformer for deformable image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11513–11523, 2024. 2

[18] Bo Hu, Shenglong Zhou, Zhiwei Xiong, and Feng Wu. Recursive decomposition network for deformable image registration. *IEEE Journal of Biomedical and Health Informatics*, 26(10):5130–5141, 2022. 2, 3, 6, 7

[19] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Advances in neural information processing systems*, 29, 2016. 3

[20] Xi Jia, Joseph Bartlett, Tianyang Zhang, Wenqi Lu, Zhaowen Qiu, and Jinming Duan. U-net vs transformer: Is u-net outdated in medical image registration? In *International Workshop on Machine Learning in Medical Imaging*, pages 151–160. Springer, 2022. 2, 3, 6, 7

[21] Xi Jia, Joseph Bartlett, Wei Chen, Siyang Song, Tianyang Zhang, Xinxing Cheng, Wenqi Lu, Zhaowen Qiu, and Jinming Duan. Fourier-net: Fast image registration with band-limited deformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1015–1023, 2023. 2, 3, 6, 7

[22] Zi-Rong Jin, Tian-Jing Zhang, Tai-Xiang Jiang, Gemine Vivone, and Liang-Jian Deng. Lagconv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1113–1121, 2022. 3

[23] Miao Kang, Xiaojun Hu, Weilin Huang, Matthew R Scott, and Mauricio Reyes. Dual-stream pyramid registration network. *Medical image analysis*, 78:102379, 2022. 2, 3, 6, 7

[24] Boah Kim, Dong Hwan Kim, Seong Ho Park, Jieun Kim, June-Goo Lee, and Jong Chul Ye. Cyclemorph: cycle consistent unsupervised deformable image registration. *Medical image analysis*, 71:102036, 2021. 2, 6

[25] Yihao Liu, Lianrui Zuo, Shuo Han, Yuan Xue, Jerry L Prince, and Aaron Carass. Coordinate translator for learning deformable medical image registration. In *International workshop on multiscale multimodal medical imaging*, pages 98–109. Springer, 2022. 2, 3, 5, 6, 7

[26] Tai Ma, Suwei Zhang, Jiafeng Li, and Ying Wen. Iirp-net: Iterative inference residual pyramid network for enhanced

image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11546–11555, 2024. 2, 3

[27] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, page 3. Atlanta, GA, 2013. 3

[28] Mingyuan Meng, Dagan Feng, Lei Bi, and Jinman Kim. Correlation-aware coarse-to-fine mlps for deformable medical image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9645–9654, 2024. 2, 3

[29] Tony CW Mok and Albert Chung. Fast symmetric diffeomorphic image registration with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4644–4653, 2020. 2

[30] Tony CW Mok and Albert Chung. Affine medical image registration with coarse-to-fine vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20835–20844, 2022. 2

[31] Tony CW Mok and Albert CS Chung. Large deformation diffeomorphic image registration with laplacian pyramid networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 211–221. Springer, 2020. 2, 3, 6, 7

[32] Huaqi Qiu, Chen Qin, Andreas Schuh, Kerstin Hammernik, and Daniel Rueckert. Learning diffeomorphic and modality-invariant registration using b-splines. In *Medical Imaging with Deep Learning*, 2021. 2, 3, 6, 7

[33] Amin Ranem, John Kalkhof, and Anirban Mukhopadhyay. Nca-morph: Medical image registration with neural cellular automata. *arXiv preprint arXiv:2410.22265*, 2024. 2, 3, 6, 7

[34] Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. Svf-net: learning deformable image registration using shape matching. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pages 266–274. Springer, 2017. 2

[35] Yudi Sang, Xianglei Xing, Yingnian Wu, and Dan Ruan. Imposing implicit feasibility constraints on deformable image registration using a statistical generative model. *Journal of Medical Imaging*, 7(6):064005–064005, 2020. 2

[36] David W Shattuck, Mubeena Mirza, Vitria Adisetiyo, Cornelius Hojatkashani, Georges Salamon, Katherine L Narr, Russell A Poldrack, Robert M Bilder, and Arthur W Toga. Construction of a 3d probabilistic atlas of human cortical structures. *Neuroimage*, 39(3):1064–1080, 2008. 6

[37] Jiacheng Shi, Yuting He, Youyong Kong, Jean-Louis Coatrieux, Huazhong Shu, Guanyu Yang, and Shuo Li. Xmorpher: Full transformer for deformable medical image registration via cross attention. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 217–226. Springer, 2022. 2

[38] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. Deformable medical image registration: A survey. *IEEE transactions on medical imaging*, 32(7):1153–1190, 2013. 1

[39] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11166–11175, 2019. 3

[40] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 5

[41] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15:1–28, 2015. 6

[42] Alexander Thorley, Xi Jia, Hyung Jin Chang, Boyang Liu, Karina Bunting, Victoria Stoll, Antonio de Marvao, Declan P O'Regan, Georgios Gkoutos, Dipak Kotecha, et al. Nesterov accelerated admm for fast diffeomorphic image registration. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*, pages 150–160. Springer, 2021. 1

[43] Lin Tian, Hastings Greer, François-Xavier Vialard, Roland Kwitt, Raúl San José Estépar, Richard Jarrett Rushmore, Nikolaos Makris, Sylvain Bouix, and Marc Niethammer. Gradicon: Approximate diffeomorphisms via gradient inverse consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18084–18094, 2023. 2

[44] D Ulyanov. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 3

[45] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: Efficient nonparametric image registration. *NeuroImage*, 45(1):S61–S72, 2009. 1

[46] Haiqiao Wang and Yi Ni, Dongand Wang. Modet: Learning deformable image registration via motion decomposition transformer. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 740–749, 2023. 2, 3, 6, 7

[47] Zhoubing Xu, Christopher P Lee, Mattias P Heinrich, Marc Modat, Daniel Rueckert, Sebastien Ourselin, Richard G Abramson, and Bennett A Landman. Evaluation of six registration methods for the human abdomen on clinically acquired ct. *IEEE Transactions on Biomedical Engineering*, 63(8):1563–1572, 2016. 6

[48] Xiaofeng Yang, Ning Wu, Guanghui Cheng, Zhengyang Zhou, S Yu David, Jonathan J Beitler, Walter J Curran, and Tian Liu. Automated segmentation of the parotid gland based on atlas registration and machine learning: a longitudinal mri study in head-and-neck radiation therapy. *International Journal of Radiation Oncology*Biology*Physics*, 90 (5):1225–1233, 2014. 1

[49] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. Quicksilver: Fast predictive image registration–a deep learning approach. *NeuroImage*, 158:378–396, 2017. 2

[50] Jun Zhang. Inverse-consistent deep networks for unsu-

pervised deformable image registration. *arXiv preprint arXiv:1809.03443*, 2018. 2

[51] Yungeng Zhang, Yuru Pei, and Hongbin Zha. Learning dual transformer network for diffeomorphic registration. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*, pages 129–138. Springer, 2021. 2

[52] Shengyu Zhao, Yue Dong, Eric I Chang, Yan Xu, et al. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10600–10610, 2019. 2, 3

[53] Shengyu Zhao, Tingfung Lau, Ji Luo, I Eric, Chao Chang, and Yan Xu. Unsupervised 3d end-to-end medical image registration with volume tweening network. *IEEE journal of biomedical and health informatics*, 24(5):1394–1404, 2019. 2, 3

[54] Jingkai Zhou, Varun Jampani, Zhixiong Pi, Qiong Liu, and Ming-Hsuan Yang. Decoupled dynamic filter networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6647–6656, 2021. 3