MGNet: Multi-Axis Similarity Matching and Grouping Contextual Attention Pyramid Network for Deformable Medical Image Registration

Junyuan Huang[1], Mengxiao Yin[1,2*], Jiachao Li[1] and Tao Luo[1]
[1] School of Computer and Electronic Information, Guangxi University, Nanning, China
[2] Guangxi Key Laboratory of Multimedia Communications and Network Technology, Guangxi University, Nanning, China
Email: 2413393017@st.gxu.edu.cn, ymx@gxu.edu.cn, 2413393025@st.gxu.edu.cn, 2413301024@st.gxu.edu.cn

***Abstract*—Accurately modeling complex deformations remains challenging in medical image registration tasks, as actual deformations often involve a mixture of large global displacements and subtle local distortions. Although various advanced registration models have been proposed, they still fall short in achieving a balance between long-range similarity matching and multi-scale representation. To this end, this paper proposes a novel Multi-Axis Similarity Matching and Grouping Contextual Attention Pyramid Network for Deformable Medical Image Registration (MGNet). To overcome the limitations of existing explicit matching methods that are constrained by local windows, we design a Multi-Axis Similarity Matching Module (MASM), which explicitly establishes feature correspondences by jointly processing dense local windows and sparse global grids. In addition, to enhance feature representation at each network layer, we introduce a Grouping Contextual Attention Module (GCA) that aggregates grouped features with different receptive fields, enabling refined modeling of complex deformation patterns. Extensive experiments are conducted on three publicly available 3D brain MRI datasets, including LPBA, Mindboggle, and OASIS. The results demonstrate that MGNet achieves state-of-the-art performance in medical image registration.**

***Index Terms*—Medical Image Registration, Pyramid Network, Similarity Matching, Transformer**

I. Introduction

Deformable image registration aims to establish an accurate nonlinear spatial correspondence between fixed and moving images. This technique is one of the core tasks in medical image analysis and is essential for preoperative planning, intraoperative guidance, disease diagnosis, and longitudinal follow-up [1]–[3]. Traditional registration methods [4]–[7] typically solve the deformation field through an iterative optimization strategy, which results in a high computational cost. Recently, unsupervised deep learning registration methods, represented by VoxelMorph [8], have improved computational efficiency by directly estimating the deformation field. Subsequently, pyramid networks [9]–[12] have been widely adopted to handle large deformations by decomposing the complex registration task into a sequence of subproblems.

Despite the success of pyramid architectures, two critical limitations persist. First, the establishment of feature correspondences lacks a global perspective. Most methods [13]–[15] rely on convolutions to implicitly learn correlations, which lack explicit matching capability. Even though many explicit matching methods have been proposed [16], [17], their search range is often restricted to local windows, making it difficult to establish an accurate correspondence when the true corresponding positions cannot be covered. Second, the ability to represent multi-scale deformation within a single layer is limited. In medical images, large organ-level displacements and subtle tissue distortions are often coupled. However, most pyramidal networks [9]–[11] estimate a single deformation field at each resolution level, forcing it to fit multi-scale information simultaneously and thus making it difficult to handle complex deformations.

To address the above limitations, we propose a novel Multi-Axis Similarity Matching and Grouping Contextual Attention Pyramid Network for Deformable Medical Image Registration (MGNet). To solve the locality bottleneck, we propose a Multi-Axis Similarity Matching Module (MASM) that decomposes the matching process into parallel local and global axes. Specifically, the local axis captures fine-grained structural correspondences through dense local windows, while the global axis establishes long-range dependencies using sparse grid sampling, ensuring that correspondences are accurately established at all scales. Furthermore, to enhance the modeling of complex deformation, we introduce a Grouping Contextual Attention Module (GCA). It groups features and utilizes progressive convolution to assign different receptive fields, while incorporating contextual attention mechanisms to capture deeper dependencies. By generating multiple deformation subfields in parallel, the GCA enables accurate modeling of multi-scale deformations within a single layer. In summary, the main contributions of this paper are as follows:

- We propose a Multi-Axis Similarity Matching Module (MASM) that performs explicit similarity matching along local and global axes in parallel, and subsequently fuses the resulting correspondences to enable effective long-range correspondence modeling.
- We introduce a Grouping Contextual Attention Module (GCA) that enhances intra-layer multi-scale deformation modeling by generating and fusing multi-scale deformation representations from grouped features using multi-scale contextual attention.

* Corresponding author

- Extensive experiments on three publicly available 3D brain MRI datasets (LPBA, Mindboggle, and OASIS) demonstrate that MGNet achieves state-of-the-art registration performance.

II. Related Work

A. Traditional Methods

Traditional image registration is typically formulated as an iterative optimization problem based on energy minimization. The core idea is to iteratively update a parameterized deformation field to maximize image similarity. Classical approaches include optical flow-based Demons [4], B-spline constrained free-form deformation [5], large deformation diffeomorphic metric mapping [6] based on time-independent velocity fields, and symmetric normalization [7]. Although these methods offer strong theoretical guarantees and interpretability, their iterative optimization nature results in high computational overhead, which limits their applicability in real-time clinical scenarios.

B. Learning-Based Methods

VoxelMorph [8] proposed a UNet-like [18] unsupervised registration method, which enables training without relying on the ground-truth deformation fields and directly estimates the deformation field during inference, thereby solving the efficiency bottleneck of the traditional methods. However, such single-stage networks [8], [19], [20] are unable to progressively refine the deformation field within the network, since the final deformation field is estimated directly, which limits their ability to handle complex deformations. Pyramid networks [9]–[15] have been demonstrated to be effective architectures for large deformation registration by adopting a coarse-to-fine strategy, in which deformation fields are estimated at each resolution level and progressively synthesized. This approach of capturing global displacements at the low-resolution level and refining local details at the high-resolution level effectively improves the registration accuracy.

C. Correspondence Establishment

Accurate spatial correspondence estimation lies at the core of deformable image registration. Early pyramid networks, such as NICE-Net [13], typically adopted implicit correspondence learning strategies. These methods concatenate fixed features with warped moving features at the decoder stage and directly predict deformation fields using convolutional layers. However, due to the localized and implicit nature of convolutional operations, the reliability of learned correspondences is difficult to guarantee. To improve interpretability and matching accuracy, recent studies have shifted toward explicit correspondence modeling. PRNet++ [21] introduces a correlation volume to guide deformation estimation by explicitly calculating matching costs between feature points. ModeT [17] interprets attention mechanisms as motion pattern decomposition, using attention weights to assign coefficients to predefined displacement basis vectors. Similarly, SLNet [16] calculates local correspondence by computing a local similarity matrix that weights a sampled grid containing relative coordinates.

Although these explicit methods improve correspondence accuracy, they limit the search window to a small region to control computational complexity. This assumption presumes that corresponding points lie within the predefined window. However, in the presence of large displacements caused by inter-subject anatomical variability or changes in scanning pose, such methods often fail to establish effective long-range dependencies and may converge to suboptimal solutions. To address this limitation, we propose a Multi-Axis Similarity Matching Module (MASM), which jointly models complementary local and global similarities to establish reliable long-range feature correspondences.

D. Feature Representation

The performance of registration networks strongly depends on the quality of feature representation. Convolutional neural networks have become the standard backbone for most single-stage [8], [19] and pyramid-based networks [9], [13] due to their effectiveness in capturing local texture information. However, the inherently limited receptive field of convolutional kernels restricts their ability to model long-range dependencies. To alleviate this limitation, Transformers and their variants have been introduced into medical image registration to enhance global modeling capability. For example, TransMorph [20] adopts a Swin Transformer [22] as the encoder to explicitly establish long-range dependencies.

Despite these advances, fine-grained modeling of multi-scale contextual information within a single layer remains a challenge. Studies on CoTNet [23] demonstrate that exploiting dynamic contextual relationships among features is critical for improving feature discriminability. To address multi-scale deformation modeling, GroupMorph [24] employs feature grouping and multi-path convolution to simulate receptive fields of different sizes. However, this approach is still fundamentally based on static convolutional operations and lacks dynamic interaction and aggregation of features within groups. To this end, we introduce a Grouping Contextual Attention Module (GCA), which incorporates feature grouping with multi-scale contextual attention to enable multi-scale deformation representation within a single layer.

III. Method

A. Overview

Fig. 1 illustrates the overall architecture of MGNet, which consists of a dual-stream encoder with shared weights and a pyramid decoder. The overall structure and detailed configuration of the dual-stream encoder are shown in Fig. 1(a) and (c). The encoder is composed of five successive convolutional (Conv) modules, and a $2 \times 2 \times 2$ average pooling layer (AvgPool) is inserted between adjacent Conv modules. The fixed image I_f and the moving image I_m are fed into their respective encoders, generating two sets of features $F_f \in \{F_{f1}, F_{f2}, F_{f3}, F_{f4}, F_{f5}\}$ and $F_m \in$

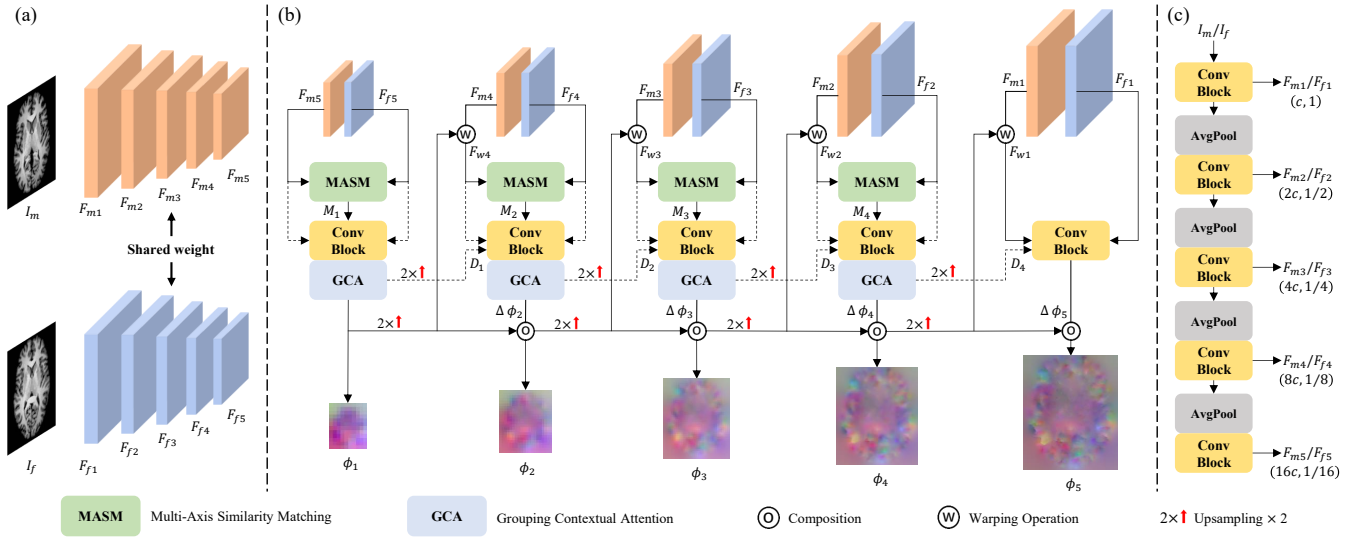

Fig. 1. (a) Dual-stream encoder with shared weights. (b) Detailed structure of the decoder. (c) Detailed structure of the encoder.

$\{F_{m1}, F_{m2}, F_{m3}, F_{m4}, F_{m5}\}$, respectively. Here, F_{fi} and F_{mi} denote the outputs of the $i\text{-}th$ Conv module, and the final feature resolution is $\left(\frac{1}{16}\right)^3$ of the original input. Each Conv module consists of two $3 \times 3 \times 3$ convolutional layers followed by instance normalization and LeakyReLU activation. The c represents the number of channels.

The pyramid decoder is illustrated in Fig. 1(b). The Multi-Axis Similarity Matching Module (MASM) and the Grouping Contextual Attention Module (GCA) are employed at each decoding stage except for the last layer. Specifically, in the first decoding layer, the MASM computes the spatial correspondence M_1 between F_{f5} and F_{m5}. Subsequently, F_{f5}, F_{m5}, and M_1 are concatenated and fed into the Conv module followed by the GCA to generate the initial deformation field ϕ_1 and the semantic feature D_1. In the following decoding layer, F_{m4} is warped using the upsampled ϕ_1 to obtain the warped moving feature F_{w4}. The warping operation is implemented using a spatial transformer network [25]. The MASM then computes M_2 between F_{f4} and F_{w4}. These features are concatenated with the upsampled D_1 and fed into the Conv module and the GCA to produce the deformation field $\Delta\phi_2$. The $\Delta\phi_2$ is subsequently composed with the upsampled ϕ_1 to obtain the deformation field ϕ_2. Similar operations are performed in the remaining decoding layers to obtain the final deformation field ϕ_5.

B. Multi-Axis Similarity Matching

The Multi-Axis Similarity Matching Module (MASM) extends feature matching to two complementary axes and derives spatial correspondences between moving and fixed features by jointly modeling local and global information.

Assuming that H, W, D, and C denote the height, width, depth, and number of channels, respectively, the inputs are the moving feature $F_m \in \mathbb{R}^{H \times W \times D \times C}$ and the fixed feature $F_f \in \mathbb{R}^{H \times W \times D \times C}$. First, F_m and F_f are normalized, after which window partitioning is performed according to a

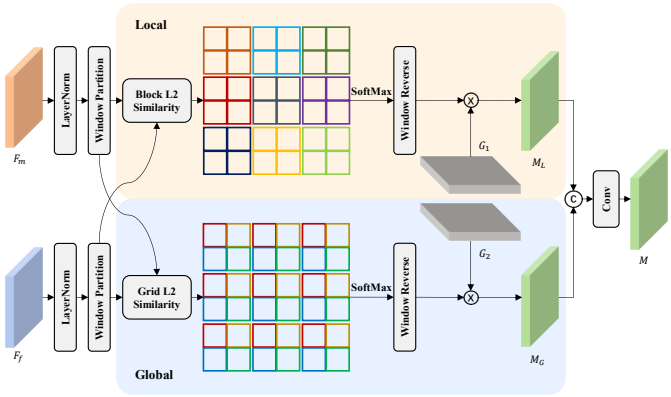

Fig. 2. Detailed structure of the Multi-Axis Similarity Matching Module.

predefined region size R. As illustrated in Fig. 2, MASM mainly consists of a local axis and a global axis. For the local axis, features are partitioned into a tensor shape of $\left(\frac{H \times W \times D}{R^3}, R^3, C\right)$, whereas for the global axis, features are partitioned into a tensor shape of $\left(R^3, \frac{H \times W \times D}{R^3}, C\right)$. The local axis partitions features into a set of non-overlapping local windows [22], while the global axis employs sparse sampling to construct global windows across different regions. The local similarity (LS) and global similarity (GS) between features are computed using the negative squared Euclidean distance (i.e., $L2$ similarity) [26] and are normalized by Softmax to obtain similarity matrices, as commonly done in Transformer. The above process can be defined as follows:

$$LS = Softmax\left(Block\ \mathcal{L}2\left(LN\left(F_m\right), LN\left(F_f\right)\right)\right), \quad (1)$$

$$GS = Softmax\left(Grid\ \mathcal{L}2\left(LN\left(F_m\right), LN\left(F_f\right)\right)\right), \quad (2)$$

where LN denotes the layer normalization operation, $Block\ \mathcal{L}2$ and $Grid\ \mathcal{L}2$ represent the $L2$ similarity computation along the local and global axes, respectively. Sub-

sequently, two sampling grids G_1 and G_2 are introduced to transform the similarity matrices into coordinate correspondences between features. The sampling grids $G_1, G_2 \in \mathbb{R}^{(H*W*D)\times R^3\times 3}$ are multiplied with the similarity matrices $LS, GS \in \mathbb{R}^{(H*W*D)\times R^3}$ to produce the coordinate correspondences of the local and global axes, denoted as M_L and M_G, respectively. The resulting M_L and M_G are then concatenated and fused using a $3 \times 3 \times 3$ convolution to obtain the final multi-axis coordinate correspondence M. This process can be defined as follows:

$$M_L = matmul\left(LS, G_1\right), \tag{3}$$

$$M_G = matmul\left(GS, G_2\right), \tag{4}$$

$$M = Conv\left(\left[M_L, M_G\right]\right). \tag{5}$$

C. Grouping Contextual Attention

The Grouping Contextual Attention Module (GCA) extracts multi-scale information from features within a single decoder layer in parallel and generates corresponding deformation subfields for each scale, which are finally fused into a deformation field with a more comprehensive perception of structures across different scales. Specifically, for each feature group in GCA, a Variable Scale Contextual Attention Module (VSCA) is employed to perform information extraction at a specific scale.

As shown in Fig. 3(a), given an input feature $X \in \mathbb{R}^{H\times W\times D\times C}$, the GCA first divides X into k independent groups equally along the channel dimension to obtain a set of features $X_i \in \mathbb{R}^{H\times W\times D\times (C/k)}$, $i = 1, 2 \ldots, k$ ($k = 4$ in Fig. 3). Each feature group X_i is fed into a corresponding VSCA branch, denoted as $VSCA_i$. Each $VSCA_i$ has a distinct internal structure that enables it to capture deformation patterns at a specific scale and outputs a deformation subfield $\Delta\phi_{Gi}$ containing the corresponding scale information. Finally, all parallel deformation subfields are averaged to generate the final deformation field $\Delta\phi$ at the current decoding level. The above process can be defined as follows:

$$\Delta\phi = \frac{1}{k}\sum_{i=1}^{k}\Delta\phi_{Gi}. \tag{6}$$

The VSCA is the core computational unit of the GCA. Inspired by [23], we design a variable-scale static contextual extraction mechanism to adaptively support multi-scale deformation modeling. The detailed structure of VSCA is illustrated in Fig. 3(b). For the input features of each VSCA, a static contextual representation containing neighborhood information is first extracted using convolution operations to serve as the Key for subsequent attention computation. To enable scale variability, the receptive field is adjusted by stacking different numbers of convolutional layers. Specifically, assuming that the input feature of the i-th group is X_i, a static contextual key SK_i is generated using a progressive stacked convolution (PSConv) composed of N_i $3 \times 3 \times 3$ convolutional layers:

$$SK_i = PSConv_{N_i}\left(X_i\right). \tag{7}$$

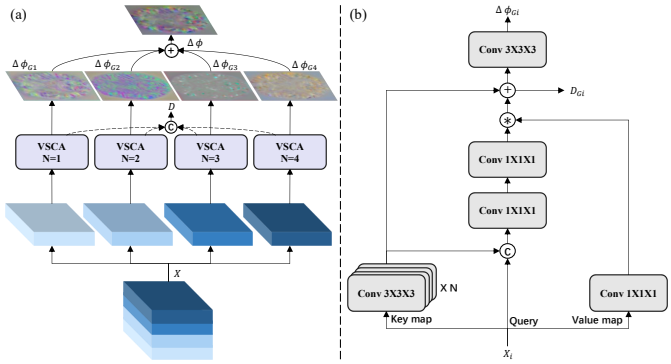

Fig. 3. (a) Detailed structure of the Grouping Contextual Attention Module. (b) Detailed structure of the Variable Scale Contextual Attention Module.

By setting different numbers of stacked layers N_i to each parallel branch in the GCA, each branch obtains a different receptive field size. Notably, the progressive stacked convolutions across different branches share weights. The Query and Value are then defined as $Q_i = X_i$ and $V_i = X_i W_{vi}$, where W_{vi} represents a $1 \times 1 \times 1$ convolution operation. The Query Q_i and static contextual key SK_i are concatenated and passed through two consecutive $1 \times 1 \times 1$ convolutions to generate the dynamic attention matrix A_i:

$$A_i = \left[SK_i, Q_i\right] W_\theta W_\delta, \tag{8}$$

where W_θ and W_δ denote $1 \times 1 \times 1$ convolution operations, respectively. The dynamic contextual key DK_i is obtained by element-wise multiplication of A_i and V_i, and is then additively fused with the static contextual key SK_i to generate the feature D_{Gi}:

$$D_{Gi} = DK_i + SK_i. \tag{9}$$

Finally, a $3 \times 3 \times 3$ convolution is employed as a registration head to generate the deformation subfield $\Delta\phi_{Gi}$ for each branch, with the registration head weights shared across all branches. Meanwhile, the features D_{Gi} from all branches are concatenated to generate the semantic feature D, which is transferred to the next decoder layer.

D. Loss Functions

We utilize two types of loss functions to guide the learning process of the registration network, which are defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{sim}\left(I_f, I_m \circ \phi\right) + \lambda\mathcal{L}_{reg}, \tag{10}$$

where λ is the weighting of $\mathcal{L}_{reg}$, and $\circ$ denotes the warping operation. The normalized cross-correlation (NCC) [27] is used as $\mathcal{L}_{sim}$ to evaluate the similarity between the fixed image I_f and the warped moving image $I_w = I_m \circ \phi$, which is defined as:

$$\mathcal{L}_{sim}\left(I_f, I_w\right) = -\sum_{p\in\Omega}\frac{\left(\sum_{p_i}\left(I_f\left(p_i\right) - \bar{I_f}\left(p\right)\right)\left(I_w\left(p_i\right) - \bar{I_w}\left(p\right)\right)\right)^2}{\left(\sum_{p_i}\left(I_f\left(p_i\right) - \bar{I_f}\left(p\right)\right)^2\right)\left(\sum_{p_i}\left(I_w\left(p_i\right) - \bar{I_w}\left(p\right)\right)^2\right)} \tag{11}$$

where Ω denotes the entire volumetric domain. Voxel p_i represents the local neighborhood within an n^3 volumetric patch (with $n = 9$ in our setting) centered at voxel p. $\bar{I}_f(p)$ and $\bar{I}_w(p)$ denote the average voxel values within this local neighborhood.

We use the spatial gradient of the deformation field as the regularization term $\mathcal{L}_{reg}$ to ensure the reasonable deformation. Notably, the deformation fields produced at all levels undergo regularization. Therefore, the regularization loss is defined as:

$$\mathcal{L}_{reg} = \sum_{l=1}^{L} \|\nabla \phi_l\|^2, \tag{12}$$

where L denotes the number of feature levels ($L = 5$ in this paper), and l indexes the different feature levels.

IV. Experiments

A. Datasets

We evaluate the effectiveness of the proposed method on three publicly available 3D brain MRI datasets, including LPBA [28], Mindboggle [29], and OASIS [30].

LPBA dataset contains 40 T1-weighted MRI volumes, each containing 54 manually labeled region-of-interests (ROIs). The MRI volumes in LPBA have been pre-aligned and centrally cropped to a size of $160 \times 192 \times 160$. We performed inter-subject registration on this dataset. 30 volumes (30×29 pairs) are employed for training and 10 volumes (10×9 pairs) are used for testing.

Mindboggle dataset contains 61 T1-weighted MRI volumes, each containing 62 manually labeled ROIs. This dataset has been pre-aligned and centrally cropped to a size of $160 \times 192 \times 160$. We performed inter-subject registration on this dataset. 42 MRI volumes (42×41 pairs from NKI-RS-22 and NKI-TRT-20 subsets) are involved for training, while MMRR-19 (19×18 pairs) are used for testing.

OASIS dataset contains 414 T1-weighted MRI volumes, each containing 35 manually labeled ROIs. This dataset has been pre-aligned and cropped to a size of $160 \times 192 \times 224$. We performed inter-subject registration on this dataset. 394 volumes are used for training and 8 volumes (8×7 pairs) are used for validation and 12 volumes (12×11 pairs) for testing.

B. Evaluation Metrics

The Dice similarity coefficient (DSC) measures the overlap between corresponding regions and is therefore adopted as the primary accuracy metric. DSC is defined as:

$$DSC(A, B) = 2\frac{|A \cap B|}{|A| + |B|}, \tag{13}$$

where A and B represent the corresponding regions from different volumes. In addition, the average symmetric surface distance (ASSD) and the 95% maximum Hausdorff distance (HD95) are used to evaluate the similarity of region contours. The ASSD is defined as:

$$ASSD(A, B) = \frac{1}{2}\left(\frac{\sum_{a \in A} d(a, B) \cdot s_a}{\sum_{a \in A} s_a} + \frac{\sum_{b \in B} d(b, A) \cdot s_b}{\sum_{b \in B} s_b}\right), \tag{14}$$

where A and B denote the sets of surface points of the two volumes, s_a and s_b represent the areas of the surface elements a and b, and $d(x, Y)$ denotes the Euclidean distance from point x to the nearest point on surface Y. Similarly, HD95 is defined as follows:

$$HD95(A, B) = \max(h_{95}(A, B), h_{95}(B, A)), \tag{15}$$

$$h_{95}(A, B) = \min\left\{\delta \;\middle|\; \frac{\sum_{a \in A, d(a,B) \leq \delta} s_a}{\sum_{a \in A} s_a} \geq 0.95\right\}. \tag{16}$$

Finally, the percentage of voxels with non-positive Jacobian determinant ($|J_\phi| \leq 0$) is used to evaluate the quality of the deformation field, which also reflects the fold rate of the deformation field. A better registration result is expected to achieve larger DSC and smaller ASSD, HD95, and Jacobian.

C. Implementation Details

Our method is implemented with PyTorch on an NVIDIA Tesla V100 GPU with 32 GB of memory. The regularization weight λ is set to 1. The number of channels c in the first encoder layer is set to 16, while the window size R and the number of groups k are set to 2 and 4, respectively. The Adam optimizer with a learning rate decay strategy is employed, which is defined as follows:

$$lr_m = lr_{init} \cdot \left(1 - \frac{m-1}{M}\right)^{0.9}, m = 1, 2, \ldots, M \tag{17}$$

where lr_m denotes the learning rate at the $m\text{-}th$ epoch, and $lr_{init} = 0.0001$ represents the initial learning rate. During training, the batch size is set to 1. The total number of training epochs M is set to 30 for the LPBA and Mindboggle datasets, and 300 for the OASIS dataset.

D. Comparison Methods

We compare our method with several state-of-the-art registration methods: (1) SyN [7]: a classical traditional method, using the SyNOnly setting in ANTs. (2) VoxelMorph [8]: a popular single-stage registration network. (3) TransMorph [20]: a single-stage registration network with SwinTransformer [22] enhanced encoder. (4) PRNet++ [21]: a pyramid registration network using 3D correlation layer. (5) PIViT [31]: a pyramid method using a recurrent SwinTransformer in the decoder's lower layers. (6) ModeT [17]: a pyramid registration network utilizing a motion decomposition module in the decoder. (7) GroupMorph [24]: a pyramid registration network using a grouping module and a group-wise correlation in the decoder. (8) SLNet [16]: a pyramid registration network using a saliency feature enhancement module in the encoder and a similarity feature matching module in the decoder.

E. Results and Discussion

a) Comparison Results: Table I and II present the quantitative results of different methods on the LPBA, Mindboggle, and OASIS datasets. As shown in the tables, our method achieves the best performance in terms of DSC, ASSD, and HD95, while maintaining a low folding rate ($|J_\phi| \leq 0$).

TABLE I
NUMERICAL RESULTS OF DIFFERENT REGISTRATION METHODS ON THE LPBA AND MINDBOGGLE DATASETS. **BOLDED** VALUE INDICATES THE BEST.

Methods	LPBA				Mindboggle			
	DSC(%) ↑	ASSD ↓	HD95 ↓	$\lvert J_\phi\rvert \leq 0$ ↓	DSC(%) ↑	ASSD ↓	HD95 ↓	$\lvert J_\phi\rvert \leq 0$ ↓
SyN	70.4±1.8	1.46±0.13	6.43±0.58	**<0.00001**	53.2±1.9	1.15±0.13	6.33±0.56	**<0.00001**
VoxelMorph	67.3±2.8	1.63±0.19	6.99±0.69	5.43e-03	54.6±2.2	1.25±0.15	6.88±0.59	9.91e-03
TransMorph	68.8±2.8	1.56±0.20	6.87±0.72	4.56e-03	58.2±2.1	1.14±0.14	6.67±0.60	9.87e-03
PRNet++	69.5±2.2	1.51±0.17	6.73±0.68	1.13e-03	58.4±2.0	1.14±0.13	6.64±0.57	7.98e-03
PIViT	71.7±1.6	1.36±0.11	6.20±0.58	6.78e-04	57.9±1.7	1.04±0.13	6.15±0.57	1.49e-03
ModeT	72.1±1.4	1.34±0.11	6.12±0.57	7.40e-05	60.2±1.5	0.99±0.11	6.02±0.50	3.24e-04
GroupMorph	72.6±1.6	1.32±0.11	6.15±0.58	9.52e-04	61.1±1.6	0.99±0.12	6.07±0.53	2.55e-04
SLNet	73.3±1.4	1.28±0.11	6.05±0.55	3.97e-03	62.4±1.6	0.96±0.12	6.00±0.54	8.72e-03
MGNet	**74.1±1.3**	**1.24±0.09**	**5.96±0.53**	7.56e-05	**63.4±1.7**	**0.93±0.12**	**5.94±0.55**	3.06e-04

TABLE II
NUMERICAL RESULTS OF DIFFERENT REGISTRATION METHODS ON THE OASIS DATASET. **BOLDED** VALUE INDICATES THE BEST.

Methods	OASIS			
	DSC(%) ↑	ASSD ↓	HD95 ↓	$\lvert J_\phi\rvert \leq 0$ ↓
SyN	77.9±2.4	0.47±0.08	1.99±0.30	**<0.00001**
VoxelMorph	79.3±2.0	0.43±0.06	2.06±0.29	8.77e-03
TransMorph	81.2±2.0	0.37±0.06	1.88±0.27	8.30e-03
PRNet++	82.2±1.7	0.34±0.05	1.73±0.28	3.08e-03
PIViT	81.4±1.7	0.36±0.05	1.74±0.25	2.26e-03
ModeT	80.6±1.9	0.39±0.06	1.94±0.30	1.83e-04
GroupMorph	82.9±1.6	0.32±0.05	1.68±0.25	6.47e-04
SLNet	83.5±1.7	0.30±0.05	1.61±0.23	8.07e-03
MGNet	**84.0±1.6**	**0.29±0.04**	**1.54±0.22**	4.86e-04

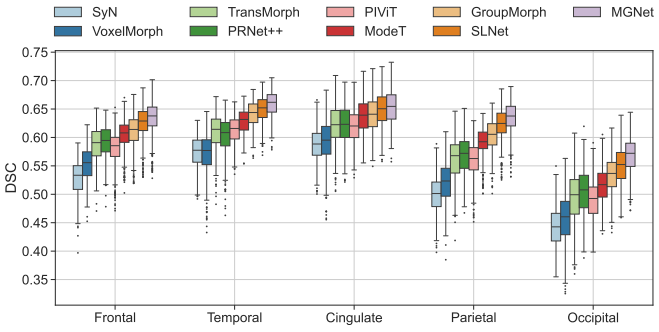

Fig. 4. Boxplots showing the distributions of DSC scores of five regions on the Mindboggle dataset produced by different registration methods.

Table III presents the inference time (Time), floating-point operations (FLOPs), and model parameters (Params) of different methods on the LPBA ($160 \times 192 \times 160$) and OASIS ($160 \times 192 \times 224$) datasets.

On the Mindboggle dataset, MGNet improves DSC by 8.8% and 5.2% compared to the single-stage networks VoxelMorph and TransMorph, respectively. Compared with the pyramid networks PRNet++ and PIViT, MGNet improves DSC by 4.6% and 2.4% on the LPBA dataset, respectively. For similarity matching, compared to ModeT and SLNet, our method improves DSC by 3.4% and 0.5% on the OASIS dataset, respectively. Moreover, MGNet improves DSC on the Mindboggle dataset by 2.3% compared to GroupMorph, which incorporates intra-layer multi-scale deformation modeling. In addition to registration accuracy, MGNet also maintains a low folding rate ($\lvert J_\phi\rvert \leq 0$), indicating that the proposed model produces deformation fields with high smoothness and continuity.

TABLE III
MODEL EFFICIENCY OF DIFFERENT REGISTRATION METHODS ON THE LPBA AND OASIS DATASETS. **BOLDED** VALUE INDICATES THE BEST.

Methods	LPBA			OASIS		
	Time (s) ↓	FLOPs (G) ↓	Params (M) ↓	Time (s) ↓	FLOPs (G) ↓	Params (M) ↓
SyN	110.76	-	-	140.896	-	-
VoxelMorph	0.05	164.32	**0.40**	**0.13**	230.05	**0.40**
TransMorph	0.18	521.17	46.77	0.37	726.04	46.77
PRNet++	0.56	1271.16	1.24	0.82	1779.58	1.24
PIViT	**0.03**	**56.36**	0.66	0.20	**78.45**	0.66
ModeT	0.60	61.05	1.03	1.05	85.46	1.03
GroupMorph	0.33	170.32	1.37	0.61	238.45	1.37
SLNet	0.46	409.44	9.89	0.77	573.22	9.89
MGNet	0.53	473.30	12.48	0.82	662.61	12.48

Fig. 4 illustrates the distribution of DSC scores for specific anatomical regions on the Mindboggle dataset, including the frontal, parietal, occipital, temporal, and cingulate regions. As shown in the boxplots, the DSC distribution of the proposed method is consistently located at a higher position than those of the compared methods. Fig. 5 presents visualizations of the registered images and the corresponding deformation fields produced by different methods on the Mindboggle dataset.

b) Ablation Results: To verify the effectiveness of each proposed component, we conducted a series of ablation experiments on the LPBA and Mindboggle datasets. As shown in Table IV, using the base pyramid structure as a baseline achieves a DSC of 61.8% on the Mindboggle dataset. Individually adding the MASM or GCA improves the DSC to 63.0% and 62.6%, respectively. The full MGNet achieves the best performance (63.4%), a 1.6% improvement over the baseline, confirming that the two components synergistically enhance registration accuracy.

Next, we evaluate MASM variants that use only local or global axes (Table V). On the Mindboggle dataset, the local-only (62.8%) and global-only (63.0%) variants exhibit lower DSC than the full MASM (63.4%). These results demonstrate

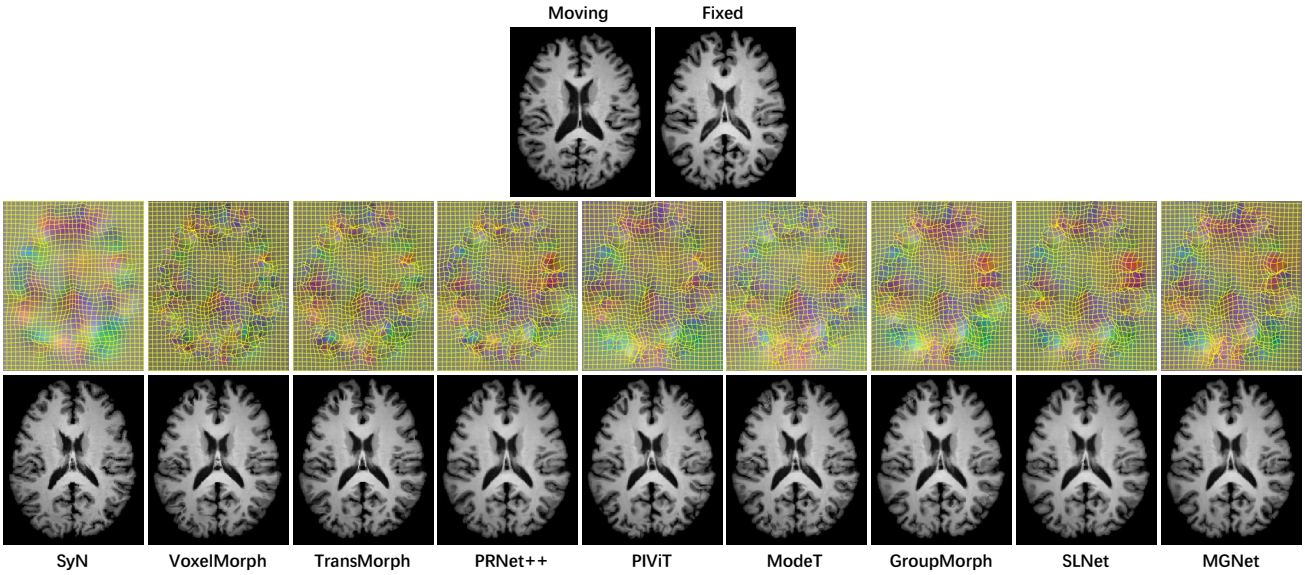

Fig. 5. The visualization of the registration results and the corresponding deformation fields from different methods on the Mindboggle dataset.

TABLE IV
ABLATION STUDIES OF NETWORK COMPONENTS ON THE LPBA AND MINDBOGGLE DATASETS. **BOLDED** VALUE INDICATES THE BEST.

Components		LPBA		Mindboggle	
MASM	GCA	DSC(%) ↑	$\lvert J_\phi\rvert \leq 0$ ↓	DSC(%) ↑	$\lvert J_\phi\rvert \leq 0$ ↓
-	-	71.2±1.8	2.91e-02	61.8±1.9	2.51e-03
✓	-	73.5±1.4	1.42e-04	63.0±1.6	**2.63e-04**
-	✓	73.3±1.5	1.47e-04	62.6±1.7	2.68e-04
✓	✓	**74.1±1.3**	**7.56e-05**	**63.4±1.7**	3.06e-04

TABLE V
ABLATION STUDIES OF MASM COMPONENTS ON THE LPBA AND MINDBOGGLE DATASETS. **BOLDED** VALUE INDICATES THE BEST.

Components		LPBA		Mindboggle	
Global	Local	DSC(%) ↑	$\lvert J_\phi\rvert \leq 0$ ↓	DSC(%) ↑	$\lvert J_\phi\rvert \leq 0$ ↓
✓	-	73.7±1.5	9.87e-05	63.0±1.6	3.26e-04
-	✓	73.5±1.6	1.10e-04	62.8±1.7	**2.60e-04**
✓	✓	**74.1±1.3**	**7.56e-05**	**63.4±1.7**	3.06e-04

that the fine-grained structures captured by the local axis and the long-range dependencies modeled by the global axis are complementary for accurate correspondence estimation. Subsequently, we investigated the number of groups k in the GCA (Table VI). On the LPBA dataset, the DSC improves from 73.5% ($k = 1$) and 73.7% ($k = 2$) to 74.1% ($k = 4$). These results confirm the effectiveness of the proposed grouping strategy and the multi-scale receptive fields constructed by the progressive stacked convolution.

V. CONCLUSION

In this paper, we propose MGNet, a Multi-Axis Similarity Matching and Grouping Contextual Attention Pyramid Network for Deformable Medical Image Registration. To address the limitations of existing methods, the Multi-Axis Similarity Matching Module (MASM) extends explicit matching beyond local windows by performing explicit similarity matching along local and global axes and fusing the resulting correspondences. Meanwhile, the Grouping Contextual Attention Module (GCA) enables refined characterization of complex deformations through contextual modeling of grouped features at multiple scales. Extensive experiments on the LPBA, Mindboggle, and OASIS datasets demonstrate that MGNet achieves state-of-the-art registration performance while maintaining a low folding rate. Furthermore, ablation experiments confirm the effectiveness of the proposed model and components.

TABLE VI
ABLATION STUDIES OF THE NUMBER OF GROUPS (k) ON THE LPBA AND MINDBOGGLE DATASETS. **BOLDED** VALUE INDICATES THE BEST.

Number of Groups (k)	LPBA		Mindboggle	
	DSC(%) ↑	$\lvert J_\phi\rvert \leq 0$ ↓	DSC(%) ↑	$\lvert J_\phi\rvert \leq 0$ ↓
1	73.5±1.6	1.30e-04	63.0±1.7	3.03e-04
2	73.7±1.4	1.12e-04	63.1±1.8	**2.62e-04**
4	**74.1±1.3**	**7.56e-05**	**63.4±1.7**	3.06e-04

ACKNOWLEDGMENT

This research was supported by the Guangxi Natural Science Foundation under Grant 2026GXNSFHA00640030 and 2025GXNSFAA069359.

REFERENCES

[1] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE transactions on medical imaging*, vol. 32, no. 7, pp. 1153–1190, 2013.
[2] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: a survey," *Machine Vision and Applications*, vol. 31, no. 1, p. 8, 2020.
[3] E. Ferrante and N. Paragios, "Slice-to-volume medical image registration: A survey," *Medical image analysis*, vol. 39, pp. 101–123, 2017.

[4] J.-P. Thirion, "Image matching as a diffusion process: an analogy with maxwell's demons," *Medical image analysis*, vol. 2, no. 3, pp. 243–260, 1998.

[5] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast mr images," *IEEE transactions on medical imaging*, vol. 18, no. 8, pp. 712–721, 2002.

[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*, vol. 61, no. 2, pp. 139–157, 2005.

[7] 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*, vol. 12, no. 1, pp. 26–41, 2008.

[8] 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*, vol. 38, no. 8, pp. 1788–1800, 2019.

[9] T. C. Mok and A. C. Chung, "Large deformation diffeomorphic image registration with laplacian pyramid networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 211–221.

[10] Y. Shu, H. Wang, B. Xiao, X. Bi, and W. Li, "Medical image registration based on uncoupled learning and accumulative enhancement," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 3–13.

[11] J. Lv, Z. Wang, H. Shi, H. Zhang, S. Wang, Y. Wang, and Q. Li, "Joint progressive and coarse-to-fine registration of brain mri via deformation field integration and non-rigid feature fusion," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2788–2802, 2022.

[12] S. Zhou, B. Hu, Z. Xiong, and F. Wu, "Self-distilled hierarchical network for unsupervised deformable image registration," *IEEE Transactions on Medical Imaging*, vol. 42, no. 8, pp. 2162–2175, 2023.

[13] M. Meng, L. Bi, D. Feng, and J. Kim, "Non-iterative coarse-to-fine registration based on single-pass deep cumulative learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 88–97.

[14] H. Wang, D. Ni, and Y. Wang, "Recursive deformable pyramid network for unsupervised medical image registration," *IEEE Transactions on Medical Imaging*, vol. 43, no. 6, pp. 2229–2240, 2024.

[15] T. Ma, S. Zhang, J. Li, and Y. 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*, 2024, pp. 11 546–11 555.

[16] Y. Chang, Z. Li, and Z. Xu, "A similarity learning network for unsupervised deformable brain mri registration," *Knowledge-Based Systems*, vol. 315, p. 113291, 2025.

[17] 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, pp. 740–749.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[19] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces," *Medical image analysis*, vol. 57, pp. 226–236, 2019.

[20] 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*, vol. 82, p. 102615, 2022.

[21] M. Kang, X. Hu, W. Huang, M. R. Scott, and M. Reyes, "Dual-stream pyramid registration network," *Medical image analysis*, vol. 78, p. 102379, 2022.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[23] Y. Li, T. Yao, Y. Pan, and T. Mei, "Contextual transformer networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 2, pp. 1489–1500, 2022.

[24] Z. Tan, L. Zhang, Y. Lv, Y. Ma, and H. Lu, "Groupmorph: medical image registration via grouping network with contextual fusion," *IEEE Transactions on Medical Imaging*, vol. 43, no. 11, pp. 3807–3819, 2024.

[25] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[26] H. K. Cheng, Y.-W. Tai, and C.-K. Tang, "Rethinking space-time networks with improved memory coverage for efficient video object segmentation," *Advances in neural information processing systems*, vol. 34, pp. 11 781–11 794, 2021.

[27] Y. R. Rao, N. Prathapani, and E. Nagabhooshanam, "Application of normalized cross correlation to image registration," *International Journal of Research in Engineering and Technology*, vol. 3, no. 5, pp. 12–16, 2014.

[28] D. W. Shattuck, M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K. L. Narr, R. A. Poldrack, R. M. Bilder, and A. W. Toga, "Construction of a 3d probabilistic atlas of human cortical structures," *Neuroimage*, vol. 39, no. 3, pp. 1064–1080, 2008.

[29] A. Klein and J. Tourville, "101 labeled brain images and a consistent human cortical labeling protocol," *Frontiers in neuroscience*, vol. 6, p. 171, 2012.

[30] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of cognitive neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[31] 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, pp. 602–612.