Reparameterized Multi-scale Transformer for Deformable Retinal Image Registration

Qiushi Nie[1†], Xiaoqing Zhang[2,1†], Chuan Chen[1], Zhixuan Zhang[1], Yan Hu[1] and Jiang Liu[1,3,4,5*]

[1]Research Institute of Trustworthy Autonomous Systems and Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China.

[2]Center for High Performance Computing and Shenzhen Key Laboratory of Intelligent Bioinformatics, Shenzhen institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.

[3]Guangdong Provincial Key Laboratory of Brain-inspired Intelligent Computation, Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China.

[4]Singapore Eye Research Institute, 169856, Singapore.

[5]State Key Laboratory of Ophthalmology, Optometry and Visual Science, Eye Hospital, Wenzhou Medical University, Wenzhou 325027, China.

*Corresponding author(s). E-mail(s): liuj@sustech.edu.cn;
†These authors contributed equally to this work.

Abstract

Deformable retinal image registration is crucial in clinical diagnosis and longitudinal studies of retinal diseases. Most existing deep deformable retinal image registration methods focus on fully convolutional network (FCN) architecture design, which fail to model long-range dependencies among pixels—a significant factor in deformable retinal image registration. Transformers based on the self-attention mechanism, can capture global context dependencies, complementing local convolution. However, multi-scale spatial feature fusion and pixel-wise position selection are also crucial for the deformable retinal image registration, are often ignored by both FCNs and transformers. To fully leverage the merits of FCNs, multi-scale spatial attention, and transformers, we propose a hierarchical hybrid architecture, Reparameterized Multi-scale Transformer (RMFormer), for deformable retinal image registration. In RMFormer, we specifically develop a reparameterized multi-scale spatial attention to adaptively fuse multi-scale spatial features, with the assistance of the re-parameterizing technique, thereby highlighting informative pixel-wise positions in a lightweight manner. Experimental results on two publicly available datasets demonstrate the superiority of our RMFormer over state-of-the-art methods and show it is data-efficient in a limited medical image regime. Additionally, we are the first to provide a visualization analysis to explain how our proposed method affects the deformable retinal image registration process. The source code of our work is available at https://github.com/Tloops/RMFormer.

Keywords: deformable registration, retinal, Reparameterized Multi-scale Transformer, re-parameterization, attention

1 Introduction

Retinal image registration is a fundamental task in ophthalmic image analysis [1]. Its goal is to find the correspondences between ophthalmic images taken from different viewpoints, time and even modalities by aligning their information, which is beneficial to clinical diagnosis [2] and longitudinal studies [3] of retinal diseases, including age-related macular degeneration (AMD), diabetic retinopathy (DR), and glaucoma. Deformable retinal image registration is one of the most common retinal image registration tasks, aiming to achieve accurate mapping and alignment of specific structures in the retina [4] via image alignment. Fig. 1 provides a representative example of deformable retinal image registration based on fundus images, which can help the audience easily understand this task. Clinically, clinicians usually perform deformable retinal image registration, heavily relying on their professional knowledge and clinical training to find the corresponding blood vessels or lesions. Unfortunately, this mode is time-consuming and error-prone. Therefore, developing computer-aid registration techniques is necessary and significant to provide efficient and accurate deformable retinal image registration.

In the past years, deep learning techniques have dominated computer vision and medical image analysis fields [5–9]. Most of existing deep deformable image registration methods are based on fully convolutional networks (FCNs) and their variants [10–13], which can provide pixel-level predictions. However, these FCN-based methods have two common shortcomings: (1) Small convolution kernels (e.g., 3 × 3) restrict their receptive fields to a local region, inevitably losing sight of long-range dependencies among pixel-wise features, which are crucial for deformable image registration due to significant variations in lesion shape and size. (2) Most previous works use a single convolution kernel size to

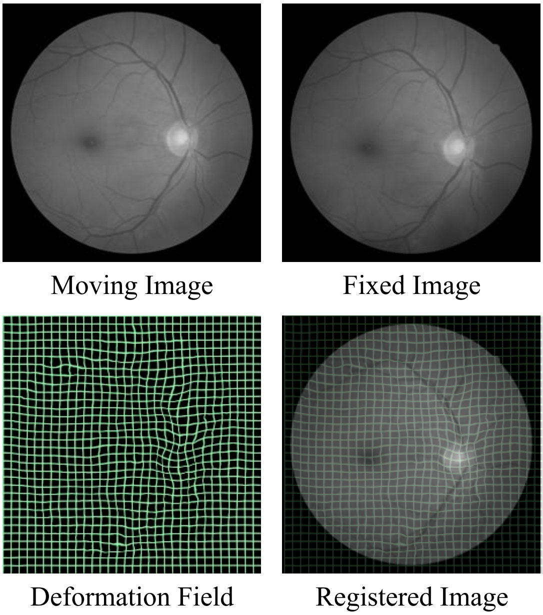


Fig. 1 Deformable retinal image registration warps the moving image via the predicted deformation field to align the moving image to the fixed image. Each pixel of the deformation field represents the spatial displacement of that particular pixel in the moving image with respect to its corresponding location in the fixed image.

learn single-scale features in every convolutional layer, which cannot capture multi-scale features.

Recently, Transformers based on the self-attention mechanism have achieved competitive performance through comparisons to FCN-based counterparts in various learning tasks [14, 15]. In the context of its long-range dependency modeling mechanism and adaptive feature encoding ability, Transformers also have been applied to tackle medical image registration task. Chen *et al.* [16] pioneered the use of Transformer blocks on high-level features extracted from convolutional layers of moving and fixed images. Subsequently, TransMorph [17] used Swin Transformer [18] as the encoder and convolutional layers as the decoder, respectively. Pure Transformer architectures such as XMorpher [19] and Swin-VoxelMorph [20] have also been developed to address medical image registration tasks. Although long-range feature interaction can be learned well with Transformers, they still have the following limitations: (1) The Transformer-based methods have low inductive bias as demonstrated in [21], thereby they commonly require more data to achieve convergence during training. The additional position embedding may mitigate this issue, but its inherent nature restricts its ability to capture positional information accurately. (2) Transformers use fixed patch sizes and only interact between the embedded tokens, which limits multi-scale spatial feature learning. Consequently, the potential for designing hybrid CNN-Transformer architectures remains a compelling area for further exploration in deformable image registration.

According to our systematical analysis, we have gained the following sights: (1) FCN-based methods excel in capturing local features and have strong inductive bias [22]. (2) Transformers demonstrate its superiority in long-range dependency modeling. Motivated by the complementary roles of CNNs and Transformers, it is natural to design a hybrid network with medical image registration in mind. Previous studies [16, 23, 24] have combined CNNs with Transformers by simply adding Transformer blocks to higher-level layers of the network, but they primarily serve as feature enhancers without effectively exploring their merits. Additionally, they overlook the relative importance of multi-scale spatial features, which is crucial for medical image registration. Spatial attention mechanisms have been proven effective in various vision tasks because they are capable of dynamically weighing spatial features for highlighting informative pixel-wise positions [25]. The utilization of multi-scale spatial features in spatial attention blocks can improve the network's robustness and feature representation learning ability, which has been less studied before. Moreover, previous works have not exploited the hybrid architecture of FCNs and Transformers with multi-scale spatial attention blocks for deformable retinal image registration in limited ophthalmic image regimes. A question naturally arises: *how to fully leverage the advantages of FCNs, Transformers, and multi-scale spatial attention for deep deformable retinal image registration network design in a computation-efficient manner?*

To tackle this question, we rethink the deformable retinal image registration network design and develop a novel hybrid registration network architecture, Reparameterized Multi-scale Transformer (RMFormer), for deformable retinal image registration based on fundus images, which is an encoder-decoder structure, as shown in Fig. 2. In the RMFormer, we propose a novel hybrid re-parameterized transformer (HRT) module, as shown in Fig. 2(d), by integrating the merits of inductive bias of FCN, long-range dependency modeling of transformer block, multi-scale spatial feature exploitation in multi-scale spatial attention, as shown in Fig. 2. HRT is comprised of a multi-scale convolution block (MSCB) for learning local features and highlighting significant pixel-wise positions and a Swin Transformer block (STB), for capturing global context dependencies. In particular, we design a reparameterized multi-scale spatial attention block (R-MSSA) in the MSCB block, which dynamically fuses multi-scale spatial features to emphasize informative pixel-wise positions. To reduce the computational cost and parameters of the R-MSSA block at the inference time, we apply the re-parameterization technique [26, 27] to merge multi-scale kernels within R-MSSA into a singular kernel by transferring the parameters. We conduct extensive experiments on two 2D/3D deformable medical image registration tasks to verify the effectiveness of our proposed RMFormer. The main contributions of this paper are summarized as follows:

1. We proposed a novel hybrid deformable medical image registration network architecture, RMFormer, aiming to fully leverage the advantages of

local features of FCNs, global features of Transformers, and multi-scale spatial attention features simultaneously, enhancing the robustness and feature representation learning ability of our RMFormer.

2. We propose a reparameterized multi-scale spatial attention (R-MSSA) block to dynamically fuse multi-scale spatial features for highlighting significant pixel-wise positions, in which we employ a re-parameterization technique to reduce the complexity of R-MSSA at the inference stage.
3. The extensive experiments on a publicly available retinal image dataset and a 3D MRI dataset demonstrate the effectiveness of our proposed methods. Moreover, we are the first to provide a visual analysis of the intermediate feature maps to explain the inherent behaviors of our proposed method.

The rest of this paper is structured as follows: Section 2 presents a comprehensive review of the literature related to our work. Section 3 provides a detailed exposition of our proposed method. In Section 4, we present the results of our experimental evaluation, providing both a quantitative analysis of performance metrics and a qualitative appraisal of visual outputs. Finally, Section 6 concludes the key contributions of this work.

2 Related Work

2.1 FCN-based Deformable Registration Methods

Over the years, significant advancements have been made in deformable medical image registration, with many methods leveraging FCNs. Fan *et al.* [28] employed a hierarchical dual-supervised FCN to achieve brain MR registration. Balakrishnan *et al.* [11] introduced VoxelMorph, a deep learning framework for deformable medical image registration, utilizing an unsupervised learning strategy. Hu *et al.* [29] developed a dual-stream pyramid network that exploits multi-level contextual information and dual-stream feature representations from pairs of medical images. Kim *et al.* [13] applied two FCNs to generate forward and reverse deformation fields. Mok *et al.* [30] proposed an FCN designed to learn symmetric deformation fields, enhancing the invertibility of the transformation process. Additionally, cascaded FCN-based methods [31, 32] employ a series of stacked FCNs to perform coarse-to-fine registration.

For deformable retinal image registration, Zhang *et al.* [33] introduced a framework that combines joint vessel segmentation and deformable registration based on the U-Net architecture. Tian *et al.* [34] proposed a multi-scale U-Net specifically for deformable retinal image registration, using multi-scale fixed and moving images as inputs. Sui *et al.* [35] developed a multi-spectral image registration network, which employs a pyramid strategy to feed both the original image and ground truth vessel maps into each layer of the encoder. Benvenuto *et al.* [36] proposed a U-Net-based vessel registration network that uses segmented blood vessel maps as input. While these FCN-based methods excel at learning local features, they often struggle with capturing multi-scale spatial features and long-range global dependencies.

2.2 Transformer-based Deformable Registration Methods

Recently, Transformer-based methods have shown remarkable success in the medical image analysis domain, including tasks related to medical image registration. Chen *et al.* [16] ioneered the integration of a Transformer block into the higher layers of a UNet, enhancing feature representation by modeling long-range dependencies among features. Similarly, Song *et al.* [23] incorporated multiple Transformer blocks in the middle stage of the UNet architecture. Building on these developments, TransMorph [17] replaced the traditional encoder with Swin Transformer [18], shifting to a pure Transformer-based encoder for deformable medical image registration. Additionally, Shi *et al.* [19] introduced an X-shaped network that uses two cross-attention Transformer blocks to establish correlations between different feature representations. Zhu *et al.* [20] employed the Swin-UNet [37] for deformable medical image registration and introduced an extra inverse consistency constraint. Although these Transformer-based methods have proven effective in medical image registration, they often have low inductive bias and operate with a fixed patch size.

Moreover, U-Shape Transformer-based methods have achieved significant advancements in image segmentation tasks. For instance, U-Transformer [38] facilitates global information exchange between low-level features in the encoder and high-level features in the decoder through multi-head Transformers at various stages. Similarly, UCTransNet [39] introduces a Channel Transformer to replace traditional skip connections, effectively capturing multi-scale global features. Recently, LEFormer [40] employs dual encoder branches—one based on CNNs and the other on Transformers—and a cross-encoder fusion module to integrate local and global features. These U-shaped Transformer-based approaches have demonstrated their versatility and might also be applicable to medical image registration tasks, an area that previous research has largely overlooked.

In contrast to the previous methods, our method integrates the merits of local feature learning of FCNs, global feature modeling of Transformers, and multi-scale spatial feature fusion and selection in multi-scale spatial attention.

3 Methodology

Deformable image registration is a process that involves two input images, a moving image M and a fixed image F, both in an n-dimensional space. This process aims to maximize the similarity between the input image pair, and the output is a dense deformation field ϕ, where $\phi = Id + u$. In this equation, Id denotes the identity and u denotes a flow field of displacement vectors. To model this deformation field efficiently in retinal image registration, we propose a hybrid deformable medical image registration network, named Reparameterized Multi-scale Transformer (RMFormer), as shown in Fig. 2. A detailed description of the network architecture, the employed loss function, and regularization techniques are elaborated in the following sections.

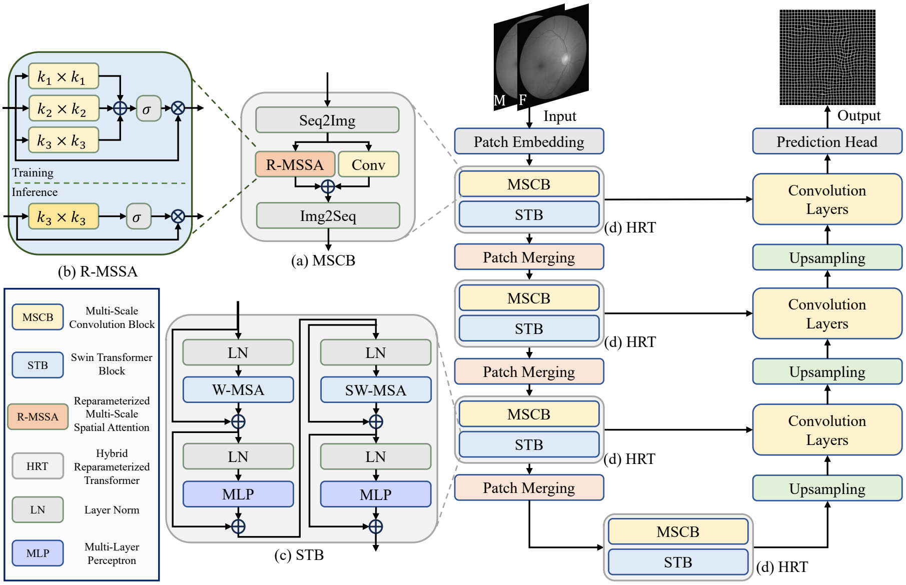


Fig. 2 The overall framework of the proposed hybrid deformable medical image registration network, Reparameterized Multi-scale Transformer (RMFormer), which fully leverages advantages of local features of FCNs, global features of transformers, and multi-scale feature utilization of multi-scale spatial attention block. First, we extract local features, multi-scale features, and long-range features by HRT blocks at four encoder stages. Then, we send the hierarchical multi-level features generated by three HRT blocks into convolutional layers of three decoder stages with corresponding multi-level skip connections. Finally, the last decoder generates the predicted deformation field.

3.1 Overall Architecture

The overall architecture of our proposed RMFormer is illustrated in Fig. 2, which follows the U-shape design philosophy. It consists of an encoder with four stages, a decoder with three stages, and three multi-level skip connections. The key novelty of our method comes from a hierarchical hybrid of convolution, reparameterized multi-scale spatial attention, and transformer feature encoding strategies. In the following section, we will describe RMFormer from the data pipeline.

Given the input image pair, denoted as $M \in \mathbb{R}^{H \times W \times C}$ and $F \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent the height, width, and the channel number of the image pair, respectively. Our proposed RMFormer first concatenates the input image pair along the channel axis and then split them into N non-overlapping 2D patches $X = \{x_i \in \mathbb{R}^{P \times P \times 2C}, i = 1, 2, ..., N\}$ with a convolutional layer and a max pooling layer, where P denotes the size of each patch that is typically set to 4 [14, 18]. The number of patches N is determined by the spatial resolution of the input image pair, where $N = \frac{H}{P} \times \frac{W}{P}$, with x_i representing the i-th patch. Next, each patch is embedded into an image token E of an arbitrary dimension C using a trainable linear projection. We omit the positional embedding in this paper because it is not significant for image registration [17].

In each encoder stage, we apply several hybrid re-parameterized transformer (HRT) modules and patch merging layers to capture multi-scale spatial information and long-range features for token selection. The HRT module generates the same number of tokens as the input, while the patch merging layers concatenate the features of each group of 2×2 neighboring tokens and expand the token dimensions between two stages, resulting in the reduction of the number of tokens by a factor of $2 \times 2 = 4$ (*e.g.*, $H \times W \times C \rightarrow \frac{H}{2} \times \frac{W}{2} \times 4C$). Then, a linear layer is applied to the 4C-dimensional concatenated features to produce features for each 2C-dimension. This paper employs four HRT modules and three patch merging layers in the encoder structure, resulting in a final output dimension of $\frac{H}{32} \times \frac{W}{32} \times 8C$.

It is followed by the decoder structure, we use pure convolutional layers to implement it. In each decoder stage, an upsampling layer and two convolution layers are used. The upsampling layer takes the feature map from the previous layer and outputs the feature maps upsampled to twice the size with bilinear interpolation. Then, these upsampled feature maps are concatenated with the corresponding feature maps from the encoder path via a skip connection. The concatenated feature maps are then sent into the convolutional block, which is composed of two consecutive 3×3 convolutions with batch normalizations and ReLU activation functions. After the last decoder stage, we obtain the feature maps with resolution $\frac{H}{4} \times \frac{W}{4}$. We use a prediction head that first upsamples the feature maps to resolution $\frac{H}{2} \times \frac{W}{2}$, then performs two consecutive 3×3 convolutions, and upsamples again to recover the original resolution of the input image. Finally, we adopt a 3×3 convolution to predict the deformation field with N channels, where N is the dimension of the input image.

3.2 Hybrid Re-parameterized Transformer Block

Considering the intrinsic locality of FCN-based methods with convolution operations, which can not learn long-range dependencies among pixels and ignore multi-scale spatial features. Additionally, Transformer based methods can capture global feature dependencies but also lose sight of multi-scale spatial features, which is significant for deformable retinal image registration. To integrate the merits of these three feature types, we develop a Hybrid Re-parameterized Transformer block for feature encoding, which consists of a multi-scale spatial convolution block and a Swin transformer block, as shown in Fig. 2.

3.2.1 Multi-scale Spatial Convolution Block

The structure of the proposed MSCB is illustrated in Fig. 2(a). Specifically, our proposed MSCB consists of a *Seq2Img* operation, a Convolutional Block, a Reparameterized Multi-scale Spatial Attention Block, and an *Img2Seq* operation. The output features of the Convolutional Block and the Reparameterized Multi-scale Spatial Attention Block, denoted as F_i^{conv} and F_i^{ms}, are obtained by passing the reshaped feature maps through them, respectively:

$$\begin{aligned} F_i^{\text{conv}} &= \text{ConvBlock}_i(\text{Seq2Img}(X)) \\ F_i^{\text{ms}} &= \text{R-MSSA}(\text{Seq2Img}(X)) \end{aligned} \tag{1}$$

where i denotes the i-th stage. Finally, the two features are added up and transformed back to tokens by the *Img2Seq* operation.

$$\text{MSCB}(X) = \text{Img2Seq}(F_i^{\text{conv}} + F_i^{\text{ms}}) \tag{2}$$

By integrating multi-scale spatial attention into convolutions, our proposed MSCB can enhance the features for the image registration task by capturing contextual dependencies across different scales and regions to better establish the correct correspondence.

Seq2Img: The *Seq2Img* operation reshapes the 1D tokens from Transformer to 2D feature maps which will be processed by the following convolutional block and Reparameterized Multi-scale Spatial Attention block, respectively.

Convolutional Block: To capture local information, the first jigsaw for our model is the convolution block. Specifically, given an input feature map F_{i-1}, the output of the convolutional stem is shown as follows:

$$F_i^{\text{conv}} = \text{ConvBlock}_i(F_{i-1}) \tag{3}$$

where i represents the i-th stage, F_i^{conv} is the output feature map with the same resolution as the input. We adopt the same convolution blocks as the four stages of ResNet [41] architecture. The detailed implementation will be elaborated in the next section.

Reparameterized Multi-scale Spatial Attention Block: The multi-scale and spatial relationships are very important for medical image registration when finding correspondences with different scales and shapes, and thus we propose multi-scale spatial attention to enable attention mechanism in a spatial perspective to capture multi-scale dependencies of tokens, which is a complement to token-wise self-attention in Transformer.

As shown in Fig. 2(b), our proposed reparameterized multi-scale spatial attention (R-MSSA) is implemented by convolutional layers with different kernels with a stride of 1. We use three different size of convolution kernels $k^{(1)} \in \mathbb{R}^{k_1 \times k_1}, k^{(2)} \in \mathbb{R}^{k_2 \times k_2}, k^{(3)} \in \mathbb{R}^{k_3 \times k_3}$ where $k_1 < k_2 < k_3$. The outputs of each convolution can be obtained by:

$$o_{is} = \text{BatchNorm}(F_{i-1} * k^{(s)}), s \in \{1, 2, 3\} \tag{4}$$

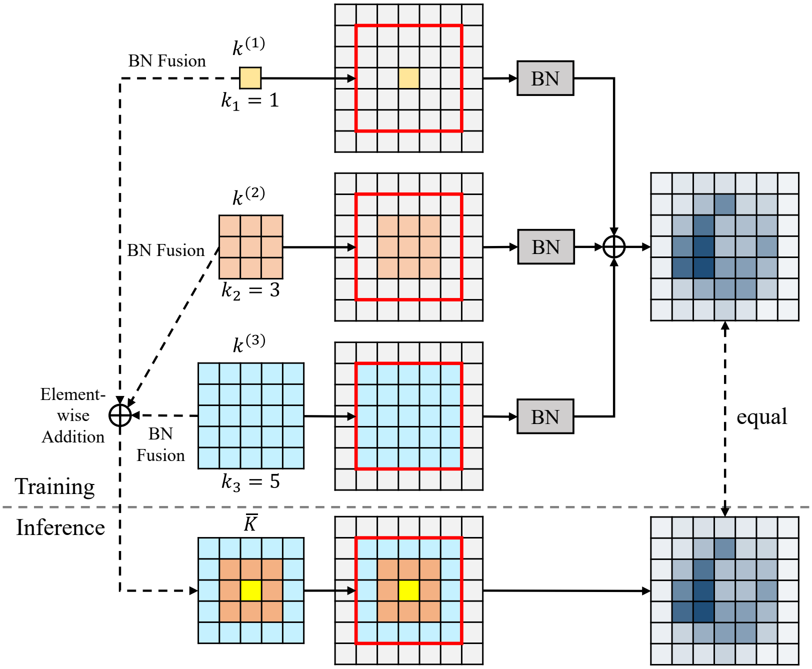


Fig. 3 A representative example of re-parameterization in our R-MSSA block. We assume the origin kernels $k^{(1)}$, $k^{(2)}$, and $k^{(3)}$ are of sizes 1 × 1, 3 × 3, and 5 × 5. After training, we fuse each kernel and the following batch normalization together and then use the fused kernel $\bar{K}$ of shape $\bar{k} = 5 \times 5$ to achieve more efficient inference.

where $o_{is} \in \mathbb{R}^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times 1}$. They are then summed up to get the feature attention map $\hat{F}_i^{\mathrm{ms}}$, which can be obtained by:

$$\hat{F}_i^{\mathrm{ms}} = \sigma(\sum_{s=1}^{3} o_{is}), i \in \{1, 2, 3, 4\} \tag{5}$$

where i denotes the i-th stage and σ is the Sigmoid activation function to get the weights that indicate the importance of each pixel of the feature map. Finally, the final calibrated multi-scale feature is calculated by the multiplication of attention map $\hat{F}_i^{\mathrm{ms}}$ and input feature F_{i-1} using spatial attention:

$$F_i^{\mathrm{ms}} = \hat{F}_i^{\mathrm{ms}} \otimes F_{i-1} \tag{6}$$

We seek to minimize the inference time of the MSSA block in a way that the multi-scale convolutions can be fused into a single standard convolution. We achieve this through structural re-parameterization, a technique previously explored in [26, 42, 43]. To elaborate, when multiple 2D kernels are applied to the same input with the same stride and produce outputs of the same resolution, their outputs can be summed up. By adding these kernels to corresponding positions, we can obtain an equivalent kernel that produces the same output. To be specific, we first fuse the j-th batch normalization and linear scaling transformation into the j-th convolutional kernels $k^{(j)}$ by:

$$K^{(j)} = \frac{\gamma_j}{\sigma_j} k^{(j)} - \frac{\mu_j \gamma_j}{\sigma_j} + \beta_j \tag{7}$$

where μ_j and σ_j are the values of the mean and standard deviation of batch normalization, γ_j and β_j are the learned scaling factor and bias term, respectively. After that, our multi-scale convolutions can be fused to a single convolution kernel $\bar{K} \in \mathbb{R}^{k_3 \times k_3}$ by:

$$\bar{K} = K^{(1)} \oplus K^{(2)} \oplus K^{(3)} \tag{8}$$

where $\oplus$ is the element-wise addition of the kernel parameters on the corresponding positions and $K^{(1)}$, $K^{(2)}$, and $K^{(3)}$ are batch normalization fused multi-scale 2D kernels. The element-wise addition of kernels with different sizes is shown in Fig. 3. Small-size kernels can be considered to have the same size as the largest kernel by padding zeros around them. As the strides of the kernels are the same, if we add the kernels on the corresponding positions, using the resulting kernel to operate on the original input will produce the same result, which can be easily verified. So at inference time, the feature attention map can be calculated simply by:

$$\hat{F}_i^{\text{ms}} = \sigma(F_{i-1} * \bar{K}), i \in \{1, 2, 3, 4\} \tag{9}$$

Img2Seq: The *Img2Seq* operation flattens the 2D feature maps from convolutions back to 1D tokens which serve as the input of the following Swin Transformer Block.

3.2.2 Swin Transformer Block

In this work, we adopt Swin Transformer [18] block to capture long-range dependencies among features. The structure of swin transformer block (STB) is depicted in Fig. 2(c) and consists of a window-based multi-head self attention (W-MSA), a shifted window-based multi-head self attention (SW-MSA), LayerNorm (LN) layers, and Multi-layer Perceptrons (MLPs). Specifically, the Swin Transformer employs a window partition mechanism to compute self-attention exclusively within each rectangular window. Additionally, the windows are shifted by half of the window size along each dimension to facilitate computing self-attention within the shifted windows. Based on this approach, our STB can be formulated as:

$$\begin{aligned} \hat{z}_l &= \text{W-MSA}(\text{LN}(z_{l-1})) + z_{l-1} \\ z_l &= \text{MLP}(\text{LN}(\hat{z}_l)) + \hat{z}_l \\ \hat{z}_{l+1} &= \text{SW-MSA}(\text{LN}(z_l)) + z_l \\ z_{l+1} &= \text{MLP}(\text{LN}(\hat{z}_{l+1})) + \hat{z}_{l+1} \end{aligned} \tag{10}$$

where $\hat{z}_l$ and z_l denote the output features of the (S)W-MSA and the MLP module of the l^{th} block, respectively. The self-attention is computed as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}} + B)V \tag{11}$$

where $Q, K, V \in \mathbb{R}^{M^2 \times d}$ are *query*, *key*, and *value* matrices, M^2 denotes the number of tokens in a window, d denotes the dimension of *query* and *key*, and B represents the relative position of tokens in each window.

3.3 Loss

In the training process of retinal image registration, the loss function of our proposed network can be defined as:

$$\mathcal{L}(M, F, \phi) = \mathcal{L}_{sim}(M \circ \phi, F) + \lambda \mathcal{L}_{smooth}(\phi) \tag{12}$$

where M and F are the moving image and the fixed image respectively, and λ is the regularization parameter. $\mathcal{L}_{sim}(\cdot)$ is used to measure the similarity between F and $M \circ \phi$. In our work, we use the Structure Similarity Index Measure (SSIM) for retinal image registration. SSIM is designed to mimic human perception by considering structural information, luminance, and contrast. The SSIM between two images can be defined as

$$\begin{aligned} \text{SSIM}(x, y) &= L(x, y) * C(x, y) * S(x, y) \\ &= \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \end{aligned} \tag{13}$$

where L, C, and S are luminance, contrast, and structural similarity, respectively, μ_x, μ_y, σ_x, σ_y, and σ_{xy} represent the means, standard deviations, and cross-covariance of the pixel values in x and y, and c_1 and c_2 are constants to stabilize the division in case the denominator becomes zero. As we want to maximize the similarity, the SSIM loss is defined as

$$\mathcal{L}_{ssim}(x, y) = 1 - \text{SSIM}(x, y) \tag{14}$$

The function of $\mathcal{L}_{smooth}(\cdot)$ is to make the deformation field realistic, known as the smooth regularization term. It is defined as:

$$\mathcal{L}_{smooth}(\phi) = \sum_{p \in \Omega} \nabla \phi(p)^2 \tag{15}$$

4 Experiments

4.1 Datasets

FIRE Dataset [44]. The publicly available Fundus Image Registration (FIRE) dataset is intended for the registration tasks. The dataset consists of 129 retinal images forming 134 image pairs with a resolution of 2912 × 2912 pixels from 39 patients. These image pairs are categorized into three groups based on their inherent characteristics. For our experimental purposes, we have exclusively selected category S, 71 image pairs, and category A, 14 image pairs, to form our experiment dataset. The remaining category, category P, contains images with small overlaps, which is unsuitable for deformable image registration. The dataset is divided into 67 and 18 for training and testing sets.

OASIS Dataset [45, 46]. The Open Access Series of Imaging Studies (OASIS) dataset, sourced from the 2021 Learn2Reg challenge for inter-patient registration, comprises 451 brain T2 MRI images. Within this dataset, 394 images are designated for training purposes, 19 images for validation, and 38 images for testing. Notably, the dataset includes segmentation labels for 35 cortical and subcortical brain structures, providing additional information for analysis and evaluation.

4.2 Implementation Details

The proposed RMFormer is implemented using PyTorch [47] and is trained/tested on an NVIDIA RTX 2080Ti GPU and an NVIDIA A100-SXM4-40GB GPU. The Adam optimizer is employed during training for both datasets.

For the FIRE dataset, a learning rate of 0.001 and a batch size of 8 were utilized, with a maximum of 400 training epochs. The regularization parameter λ was set to 1 for all experiments. The number of *ResBlock* in each convolutional block of MSCB was set to {2,2,2,2}, following the ResNet-18 architecture. We use a window sizes of {8,8}, an embed dimension of 64, Swin Transformer block numbers {2,2,4,2}, and head numbers {4,4,8,8}.

For OASIS, a learning rate of 0.0001 and a batch size of 1 were employed during training, with a maximum of 500 training epochs. The regularization parameter λ was set to 1, and the segmentation weighting parameter γ was also set to 1. The number of *ResBlock* in each convolutional block of MSCB was adjusted to {3,4,6,3} following the ResNet-34 architecture. We use a window sizes of {5,6,7}, an embed dimension of 128, Swin Transformer block numbers {2,2,12,2}, and head numbers {4,4,8,16}.

4.3 Evaluation Metrics

For the FIRE dataset, we use the similarity metric (SSIM) to measure the deformable retinal registration performance of our RMFormer and comparable methods due to unavailable blood vessel segmentation labels. To evaluate the quality of the deformation field generated by methods, we utilize the Jacobian determinant J at each point (i, j) of the deformation field ϕ, which can be formulated as follows:

$$\det(J_\phi(i,j)) = \begin{vmatrix} \frac{\partial i}{\partial x} & \frac{\partial j}{\partial x} \\ \frac{\partial i}{\partial y} & \frac{\partial j}{\partial y} \end{vmatrix} \tag{16}$$

Specifically, we examine the percentage of negative terms of each point's Jacobian determinants for the output deformation field ϕ, denoted as $|J_\phi| < 0$, representing the deformation field's folding rate. Lower occurrences of folded pixels (or voxels) indicate a superior deformation field.

Conversely, we adopt the Dice score to measure the registration performance of methods on the publicly available OASIS dataset, where segmentation labels of anatomical structures are available. For the quality evaluation of the deformation field, we use the 3D Jacobian determinant of each point (i, j, k). It can be similarly defined as:

$$\det(J_\phi(i,j,k)) = \begin{vmatrix} \frac{\partial i}{\partial x} & \frac{\partial j}{\partial x} & \frac{\partial k}{\partial x} \\ \frac{\partial i}{\partial y} & \frac{\partial j}{\partial y} & \frac{\partial k}{\partial y} \\ \frac{\partial i}{\partial z} & \frac{\partial j}{\partial z} & \frac{\partial k}{\partial z} \end{vmatrix} \tag{17}$$

4.4 Ablation Studies

4.4.1 Effect of different kernel size selections in R-MSSA block

The upper part of Table 1 shows the registration performance of RMFormer by using different kernel sizes in R-MSSA on the FIRE dataset. Here, we select six kernel sizes: 1×1, 3×3, 5×5, 7×7, 9×9, 11×11. According to Table 1, we conclude as follows: (1) Our method achieves the highest similarity result with kernel sizes option 1×1, 5×5, and 9×9 than others kernel sizes with a competitive percentage of pixels with a negative Jacobian determinant. Hence, we adopt 1×1, 5×5, and 9×9 as the final kernel size for the following ablation and comparison experiment. (2) When only a single kernel is used, both too small (1×1) and too big (11×11) kernel sizes perform poorly. (3) When using multiple kernels, we see that the larger the size difference between the kernels used, the more diverse the features extracted, and the better the effect of the network. (4) Different kernel sizes are beneficial to enhancing diversities of feature maps by multi-scale learning, which conduces to performance improvement.

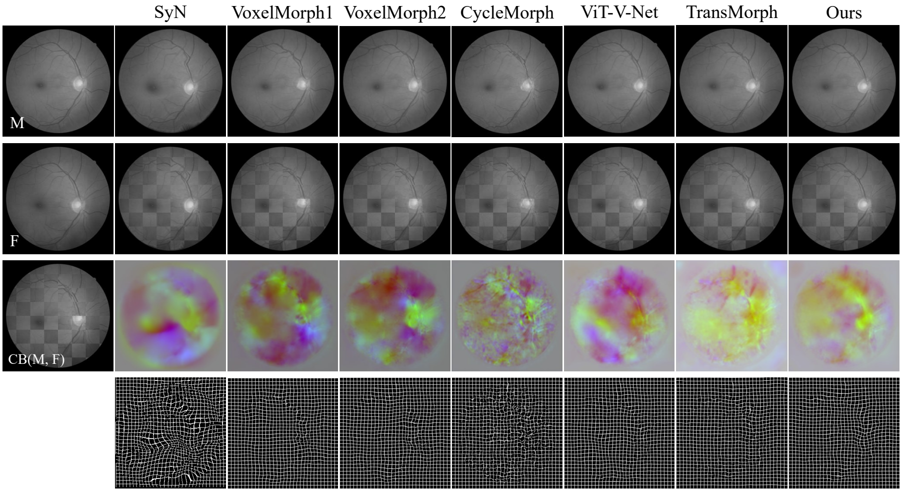


Fig. 4 Qualitative results of RMFormer (last column) and other comparison methods on FIRE dataset. The first column lists the moving image, fixed image, and the original checkerboard image CB(M, F). After that, each column represents a certain method. The first row shows the resulting warped moving image $M \circ \phi$. The second row shows the checkerboard of the fixed and the warped moving image CB$(M \circ \phi, F)$. The third row visualizes the output displacement fields, where spatial dimensions x and y are mapped to the red and green channels, respectively. The fourth row provides a visualization of the effect of the displacement field on the standard grid image.

4.4.2 Effects of each block in MSCB

The lower part of Table 1 shows the ablation study results of each submodule within our proposed MSCB. CB denotes the Convolutional Block mentioned in section 3.2.1; RC denotes adding a residual connection. For *w/o MSCB*, the MSCBs of all stages are removed. For *w/o R-MSSA* and *w/o ConvBlock*, we remove all the R-MSSA branches and all the convolutional blocks in the MSCBs, respectively. For *R-MSSA after CB*, we place the R-MSSA module after the convolutional block instead of the original parallel setting. Based on this, we further explored *(R-MSSA+RC) after CB*, in which we add a residual connection next to the R-MSSA block to change the purpose of the R-MSSA from feature selection to feature enhancement.

Although Atrous Spatial Pyramid Pooling (ASPP) module [48] is similar to our R-MSSA block, the key motivation between them is different: our R-MSSA is a spatial attention module, which captures multi-scale features for emphasizing informative pixel-wise positions; while ASPP belongs to pooling methods, which employs atrous convolution operators with dilation rates to expand the receptive field sizes without down-sampling operator for capturing multi-scale features. This paper summarizes the main difference between R-MSSA and ASPP as follows:

- Our R-MSSA block first utilizes different convolutional kernel sizes (1x1, 5x5, 9x9) to capture multi-scale feature maps, and then merges them into

Table 1 The first part shows the influence of the different sizes of kernels in R-MSSA. The second part shows the comparison of various designs of RMFormer on FIRE dataset. The **bolded** numbers denote the highest scores.

Setting	**SSIM** ↑	**% of** $\lvert J_\Phi \rvert < 0$ ↓
only 1x1	0.913 ± 0.038	8.48e-4 ± 3.59e-3
only 3x3	0.917 ± 0.039	1.36e-3 ± 6.36e-3
only 5x5	0.917 ± 0.039	**4.24e-5 ± 2.51e-4**
only 7x7	0.918 ± 0.039	3.60e-3 ± 2.13e-2
only 9x9	0.917 ± 0.039	**4.24e-5 ± 2.51e-4**
only 11x11	0.912 ± 0.038	9.62e-3 ± 1.32e-2
1x1, 3x3, 5x5	0.912 ± 0.038	6.95e-3 ± 1.44e-2
1x1, 3x3, 7x7	0.912 ± 0.038	5.21e-3 ± 9.69e-3
1x1, 3x3, 9x9	0.917 ± 0.039	8.05e-4 ± 3.02e-3
1x1, 3x3, 11x11	0.919 ± 0.038	3.14e-3 ± 9.95e-3
1x1, 5x5, 7x7	0.916 ± 0.039	1.53e-3 ± 5.69e-3
1x1, 5x5, 9x9	**0.920 ± 0.038**	8.48e-5 ± 5.02e-4
1x1, 5x5, 11x11	0.916 ± 0.039	5.09e-4 ± 1.53e-3
1x1, 7x7, 9x9	0.919 ± 0.038	3.60e-3 ± 1.02e-2
1x1, 7x7, 11x11	0.912 ± 0.038	1.03e-2 ± 1.73e-2
1x1, 9x9, 11x11	0.913 ± 0.038	7.21e-3 ± 1.26e-2
3x3, 5x5, 7x7	0.911 ± 0.039	1.86e-3 ± 5.19e-3
w/o MSCB	0.911 ± 0.040	4.24e-4 ± 1.78e-3
w/o R-MSSA	0.916 ± 0.038	1.19e-3 ± 3.30e-3
w/o ConvBlock	0.917 ± 0.038	4.24e-4 ± 1.89e-3
R-MSSA after CB	0.915 ± 0.039	7.63e-4 ± 3.81e-3
(R-MSSA+RC) after CB	0.917 ± 0.039	7.21e-4 ± 4.01e-3
ASPP replaces R-MSSA	0.917 ± 0.039	2.33e-3 ± 1.03e-2
RMFormer	**0.920 ± 0.038**	**8.48e-5 ± 5.02e-4**

a singular convolutional kernel size at the inference time with the reparameterization technique. Conversely, the ASPP module applies atrous convolution operators with different dilation rates (6, 12, 18, 24) to expand the receptive fields without downsampling.

- The R-MSSA block is fundamentally based on the spatial attention mechanism, where the spatial attention map is derived from the output feature maps of convolutions with different kernel sizes. On the other hand, our goal is to fully exploit the multi-scale features to highlight significant pixel-wise positions and suppress redundant ones effectively. In contrast, ASPP, based on atrous convolution operators, aims to replace downsampling operators to capture multi-scale feature information. It may lead to adjacent pixels being derived from independent subsets and local information loss, which is not optimal for pixel-level tasks like deformable registration.

Moreover, we use ASPP block to substitute our R-MSSA block to compare their performance, further proving their differences.

From the results in Table 1, we see that: (1) R-MSSA outperforms ASPP, indicating that the objectives of them in capturing multi-scale features, keeping with previous discussions. (2) Both the R-MSSA and ConvBlock branches in the MCCB have unique roles in capturing local features and multi-scale

Table 2 Quantitative evaluation results for the FIRE dataset. SSIM and the percentage of pixels with a negative Jacobian determinant (i.e., folded pixels) are evaluated for different methods. The **bolded** numbers denote the highest scores.

Methods	**SSIM** ↑	**% of** $\|J_\Phi\| < 0$ ↓
w/o registration	0.851 ± 0.041	-
SyN[49]	0.886 ± 0.052	3.81e-2 ± 9.57e-2
VoxelMorph-1[11]	0.904 ± 0.041	1.27e-4 ± 5.54e-4
VoxelMorph-2[11]	0.909 ± 0.039	1.53e-3 ± 4.65e-3
CycleMorph[13]	0.906 ± 0.037	7.63e-2 ± 1.09e-1
ViT-V-Net[16]	0.910 ± 0.039	4.20e-3 ± 9.39e-3
TransMorph[17]	0.911 ± 0.040	5.51e-4 ± 2.04e-3
Swin-UNet[37]	0.900 ± 0.045	6.36e-4 ± 2.84e-3
UCTransNet[39]	0.912 ± 0.041	**0** ± **0**
META-UNet[50]	0.912 ± 0.041	**0** ± **0**
RMFormer	**0.920** ± **0.038**	8.48e-5 ± 5.02e-4

features, respectively, bringing meaningful performance improvement. MSCB takes advantage of them, further boosting performance.

4.5 Comparisons with Superior Methods

4.5.1 Result on retinal image dataset

We conducted a comparative analysis of our method against several previous approaches: Symmetric Normalization (SyN) [49], FCN-based unsupervised methods VoxelMorph [11], CycleMorph [13], and Transformer-based methods Vit-V-Net [16], TransMorph [17]. The SyN algorithm was implemented and tuned using the ANTs [51] package. In this comparison, VoxelMorph-1 refers to the original architecture, whereas VoxelMorph-2 denotes an enhanced variant with twice the number of convolutional filters. Additionally, we compare our method with the latest U-shape Transformer-based methods: Swin-UNet [37], UCTransNet [39], and META-UNet [50]. For all these methods, we use the official online implementation and maintenance code and follow the optimal parameter settings. We train them from scratch to get the best performance. For RMFormer, the kernels in R-MSSA are set to the size of {1, 5, 9}, as this configuration was found to deliver the best results.

Table 2 shows the average SSIM and percentage of pixels with negative Jacobian determinants over all subjects and structures for different methods on the FIRE dataset. RMFormer achieves the overall best performance regarding average SSIM while producing smooth registration fields (less non-positive Jacobian voxels) through comparisons to superior methods. Compared with the current SOTA registration methods and SOTA U-shape transformer-based methods, RMFormer can achieve the best results, which shows the novelty of our proposed Hybrid Reparameterized Transformer module, which fully integrates local, global, and multi-scale features.

4.5.2 Result on 3D MRI dataset

For MRI registration, we follow the previous works [17] and use the local normalized cross-correlation (LNCC) as the similarity metric. The LNCC loss is defined as

$$\mathcal{L}_{lncc}(x,y) = \sum_{p\in\Omega} \frac{\left(\sum_{p_i}(x(p_i)-\bar{x}(p))(y(p_i)-\bar{y}(p))\right)^2}{\left(\sum_{p_i}(x(p_i)-\bar{x}(p))^2\right)\left(\sum_{p_i}((y(p_i)-\bar{y}(p))^2)\right)} \tag{18}$$

where Ω denotes the cuboid on which input images are defined, $\bar{x}$ and $\bar{y}$ denotes the local mean in the window of size n^3 around voxel p. p_i iterates over the volume, with $n = 9$ in our work.

In the context where the segmentation of the input image pair S_M and S_F is provided, we can exploit this supplementary information during the training process to enhance the anatomical correspondence between $M \circ \phi$ and F. To achieve this, we incorporate a loss function denoted as $\mathcal{L}_{seg}$, which evaluates the degree of overlap between the segmentations. This loss function is included as a component of the overall loss function:

$$\begin{aligned}\mathcal{L}(M,F,\phi) = \mathcal{L}_{sim}(M\circ\phi,F) + \lambda\mathcal{L}_{smooth}(\phi)\\ +\gamma\mathcal{L}_{seg}(S_M\circ\phi,S_F)\end{aligned} \tag{19}$$

where S_M and S_F represent, respectively, the organ segmentation of M and F. Additionally, the hyperparameter γ is utilized to control the weighting of the segmentation. In the domain of image registration, the Dice score is frequently employed as a metric to evaluate the quality of registration. Consequently, we directly minimize the Dice loss between the segmentations S_M^k and S_F^k, where the subscript k represents the kth structure/organ:

$$\mathcal{L}_{Dice}(S_x,S_y) = 1-\frac{1}{K}\sum_k \frac{2\sum_{p\in\Omega}S_x^k(p)S_y^k(p)}{\sum_{p\in\Omega}\left(S_x^k(p)\right)^2+\sum_{p\in\Omega}\left(S_y^k(p)\right)^2} \tag{20}$$

Table 3 shows the average Dice score, the 95% Hausdorff distance, and the standard deviation of the logarithm of the Jacobian determinant of the validation set of the challenge. The scores of various methods are obtained from the leaderboard of the challenge. RMFormer achieved the overall best performance regarding both Dice and HD95, which demonstrates the effectiveness of our method in the 3D dataset.

Table 3 Quantitative evaluation results for the OASIS dataset from the 2021 Learn2Reg challenge task 3. The results came from the challenge's leaderboard. The **bolded** numbers denote the highest scores.

Methods	DSC ↑	HD95 ↓	SDlogJ ↓
Lv *et al.* [52]	0.827 ± 0.013	1.722 ± 0.318	0.121 ± 0.015
nnU-Net [53]	0.846 ± 0.016	1.500 ± 0.304	**0.067 ± 0.005**
LapIRN [54]	0.861 ± 0.015	1.514 ± 0.337	0.072 ± 0.007
VoxelMorph-huge	0.847 ± 0.014	1.546 ± 0.306	0.133 ± 0.021
TransMorph [17]	0.858 ± 0.014	1.494 ± 0.288	0.118 ± 0.019
TransMorph-Large [17]	0.862 ± 0.014	1.431 ± 0.282	0.128 ± 0.021
Fourier-Net-Large [55]	0.860 ± 0.013	1.374 ± 0.279	0.478 ± 0.113
RMFormer	**0.872 ± 0.016**	**1.338 ± 0.280**	0.177 ± 0.040

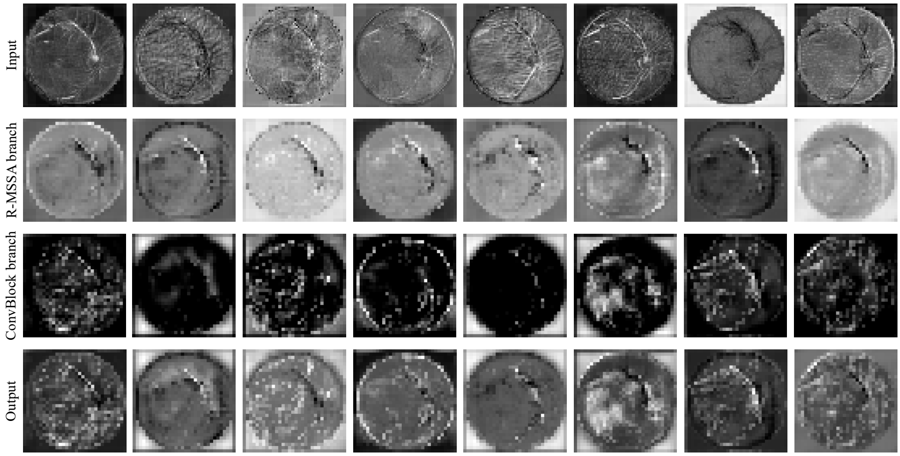


Fig. 5 Representative intermediate feature maps from the MSCB in the second stage. The first row shows the input feature maps of MSCB. The second and the third rows show the output of the R-MSSA branch and the convolutional block branch, respectively. The fourth row shows the final output feature map of MSCB.

4.6 Visual Analysis and Explanation

4.6.1 Qualitative analysis of deformation field

Fig. 4 shows the qualitative deformable retinal image registration results of SyN, VoxelMorph, CycleMorph, ViT-V-Net, TransMorph, and RMFormer based on the FIRE dataset. The first column shows the original image pair M and F, and the checkerboard (CB) image between them. The following columns are the visual results of our proposed RMFormer and the comparable methods. The first row shows the warped images using the predicted deformation field. The second row illustrates the checkerboards between the warped and fixed images. The third row visualizes the predicted displacement field, where the spatial dimensions x and y in the displacement field are mapped to the red and green channels. The fourth row shows the effect of performing the predicted displacement field on a standard grid image.

From the first row, we can see that for SyN, VoxelMorph, and CycleMorph, the blood vessels in the upper part of the warped images have large unreasonable deformations. These methods pay much attention to local vessel shape alignment but ignore the significance of global vessel shape alignment. ViT-V-Net and TransMorph can produce reasonable deformation; however, from the second row, we can easily see that they failed to align the blood vessels. Different from these methods, RMFormer can provide reasonable deformation while achieving the best alignment effects of local and global vessel shapes. We also find that RMFormer can produce the smoothest deformation field from the third and fourth rows. Overall, our proposed RMFormer not only focuses on global shape features but also guarantees the quality of local information.

4.6.2 Intermediate feature representation visualization of MSCB

To investigate how the R-MSSA block and the MSCB blocks affect the deformable retinal image registration process, we visualize the internal feature maps produced by the proposed MSCB. Fig. 5 provides the representative internal feature maps of each subblock from the MSCB in the second stage. We can summarize as follows: (1) The R-MSSA branch pays more attention to specific regions than the convolutional block branch, agreeing with our expectations; (2) The convolutional block branch pays more attention to extracting local features with diversity; (3) The final output merges the merits of the multi-scale features in highlighting informative regions, and local features in improving the diversities of feature representations with different kernels through convolution operations.

4.6.3 Intermediate feature representation visualization of FCN-based method, Transformer-based method, and hybrid method

To investigate difference of the inductive bias in FCN-based method and Transformer-based method, we visualize the intermediate feature maps in the second layer of VoxelMorph (FCN-based) and Swin-UNet (Transformer-based), together with our proposed RMFormer (hybrid), as shown in Fig. 6. VoxelMorph, with its strong inductive bias, extracts feature that focus on specific sparse areas. In contrast, Swin-UNet, which has a lower inductive bias but facilitates global information interaction, generates feature maps where the activation values of adjacent positions are similar. Our RMFormer, by integrating both structures and leveraging spatial attention, produces feature maps that complement each other, capturing both specific structures and global context.

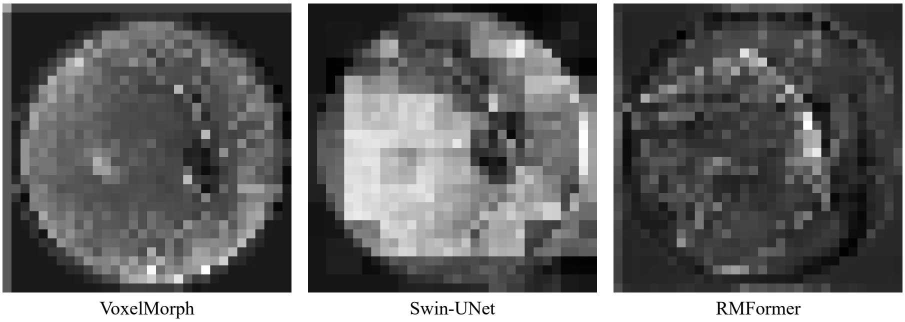


Fig. 6 Representative intermediate feature maps from VoxelMorph, Swin-UNet, and RMFormer.

5 Discussion and Limitations

This paper argues that the complementary roles of FCN-based and Transformer-based methods have not been fully explored for medical image registration tasks. In seeking solutions to this problem, we find that current FCN-based and Transformer-based methods often ignore the multi-scale feature capturing. Consequently, we propose a hybrid deformable medical image registration network to incorporate the merits of FCNs and Transformers as well as add a novel multi-scale spatial attention for multi-scale spatial feature fusion and enhancement. The results show that our proposed method performs better than recent methods on 2D/3D deformable medical image registration tasks. Moreover, this paper provides a visual explanation of intermediate feature maps from the feature representation learning perspective. There are still some limitations in our proposed method, as follows:

- We utilize an R-MSSA block to capture multi-scale spatial features for highlighting significant pixel-wise positions in RMFormer. However, the multi-scale spatial feature fusion and pixel position selection mechanisms need to be further improved, which may be beneficial to both improving the explanation and performance of deformable medical image registration.
- This paper only provides the deformable registration for retinal images, and an extended version of affine registration can be further explored.
- We only test the effectiveness of RMFormer on limited retinal image data due to the scarcity of retinal image images.

To address the above limitations, we will design improved multi-scale feature extraction methods and expand our method to provide affine registration on a larger retinal image registration dataset.

6 Conclusion

This paper proposes a Reparameterized Multi-scale Transformer (RMFormer) for deformable retinal image registration by integrating the advantages of local features of FCNs, global features of Transformers, and multi-scale spatial attention features. Moreover, we apply a re-parameterizing technique to reduce the parameters and computational cost of our proposed R-MSSA block at the inference time. The extensive experiments on both the 2D retinal image registration task and 3D MRI image registration task demonstrate the effectiveness and generalization ability of our method through comparisons to state-of-the-art methods in a limited medical image regime. Furthermore, we provide the internal feature representation visualization to explain how our proposed MSCB work affects the deformable registration process. In the future, we plan to improve multi-scale feature extraction and selection strategies and provide affine registration for retinal images.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No.82272086), the Leading Goose Program of Zhejiang (2023C03079), and Shenzhen Natural Science Fund (No. JCYJ20200109140820699).

Declarations of Conflict of Interest

The authors declared that they have no conflicts of interest to this work.

References

[1] Saha, S.K., Xiao, D., Bhuiyan, A., Wong, T.Y., Kanagasingam, Y.: Color fundus image registration techniques and applications for automated analysis of diabetic retinopathy progression: A review. Biomedical Signal Processing and Control **47**, 288–302 (2019)

[2] Hernandez-Matas, C., Zabulis, X., Argyros, A.A.: Retinal image registration as a tool for supporting clinical applications. Computer Methods and Programs in Biomedicine **199**, 105900 (2021)

[3] Liu, S., Datta, A., Ho, D., Dwelle, J., Wang, D., Milner, T.E., Rylander, H.G., Markey, M.K.: Effect of image registration on longitudinal analysis of retinal nerve fiber layer thickness of non-human primates using optical coherence tomography (oct). Eye and Vision **2**(1), 1–12 (2015)

[4] Zou, J., Gao, B., Song, Y., Qin, J.: A review of deep learning-based deformable medical image registration. Frontiers in Oncology **12**, 1047215 (2022)

[5] Zhang, X., Xiao, Z., Yang, B., Wu, X., Higashita, R., Liu, J.: Regional context-based recalibration network for cataract recognition in as-oct. Pattern Recognition **147**, 110069 (2024)

[6] Zhang, X., Xiao, Z., Fu, H., Hu, Y., Yuan, J., Xu, Y., Higashita, R., Liu, J.: Attention to region: Region-based integration-and-recalibration

networks for nuclear cataract classification using as-oct images. Medical Image Analysis, 102499 (2022)

[7] Li, S., Xiong, M., Yang, B., Zhang, X., Higashita, R., Liu, J.: Oct image blind despeckling based on gradient guided filter with speckle statistical prior. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023). IEEE

[8] Li, H., Liu, H., Hu, Y., Fu, H., Zhao, Y., Miao, H., Liu, J.: An annotation-free restoration network for cataractous fundus images. IEEE Transactions on Medical Imaging **41**(7), 1699–1710 (2022)

[9] Fang, Q., Yang, Y., Wang, H., Sun, H., Chen, J., Chen, Z., Pu, T., Zhang, X., Liu, F.: Lcrnet: local cross-channel recalibration network for liver cancer classification based on ct images. Health Information Science and Systems **12**(1), 5 (2023)

[10] Zhang, J.: Inverse-consistent deep networks for unsupervised deformable image registration. arXiv preprint arXiv:1809.03443 (2018)

[11] 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)

[12] Mok, T.C., Chung, A.C.: 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, pp. 211–221 (2020). Springer

[13] Kim, B., Kim, D.H., Park, S.H., Kim, J., Lee, J.-G., Ye, J.C.: Cyclemorph: cycle consistent unsupervised deformable image registration. Medical image analysis **71**, 102036 (2021)

[14] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[15] Shen, J., Hu, Y., Zhang, X., Qiu, Z., Deng, T., Xu, Y., Liu, J.: Interaction-oriented feature decomposition for medical image lesion detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 324–333 (2022). Springer

[16] Chen, J., He, Y., Frey, E.C., Li, Y., Du, Y.: Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. arXiv preprint arXiv:2104.06468 (2021)

[17] 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)

[18] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)

[19] Shi, J., He, Y., Kong, Y., Coatrieux, J.-L., Shu, H., Yang, G., Li, S.: Xmorpher: Full transformer for deformable medical image registration via cross attention. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 217–226 (2022). Springer

[20] Zhu, Y., Lu, S.: Swin-voxelmorph: A symmetric unsupervised learning model for deformable medical image registration using swin transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 78–87 (2022). Springer

[21] Xu, Y., Zhang, Q., Zhang, J., Tao, D.: Vitae: Vision transformer advanced by exploring intrinsic inductive bias. Advances in neural information processing systems **34**, 28522–28535 (2021)

[22] He, A., Wang, K., Li, T., Du, C., Xia, S., Fu, H.: H2former: An efficient hierarchical hybrid transformer for medical image segmentation. IEEE Transactions on Medical Imaging **42**(9), 2763–2775 (2023)

[23] Song, L., Liu, G., Ma, M.: Td-net: unsupervised medical image registration network based on transformer and cnn. Applied Intelligence **52**(15), 18201–18209 (2022)

[24] Wang, Y., Qian, W., Li, M., Zhang, X.: A transformer-based network for deformable medical image registration. In: Artificial Intelligence: Second CAAI International Conference, CICAI 2022, Beijing, China, August 27–28, 2022, Revised Selected Papers, Part I, pp. 502–513 (2022). Springer

[25] Zhang, X., Xiao, Z., Wu, X., Chen, Y., Zhao, J., Hu, Y., Liu, J.: Pyramid pixel context adaption network for medical image classification with supervised contrastive learning. IEEE Transactions on Neural Networks and Learning Systems, 1–14 (2024). https://doi.org/10.1109/TNNLS.2024.3399164

[26] Ding, X., Guo, Y., Ding, G., Han, J.: Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1911–1920 (2019)

[27] Zhang, X., Wu, X., Xiao, Z., Hu, L., Qiu, Z., Sun, Q., Higashita, R., Liu, J.: Mixed-decomposed convolutional network: A lightweight yet efficient convolutional neural network for ocular disease recognition. CAAI Transactions on Intelligence Technology **9**(2), 319–332 (2024)

[28] Fan, J., Cao, X., Yap, P.-T., Shen, D.: Birnet: Brain image registration using dual-supervised fully convolutional networks. Medical Image Analysis **54**, 193–206 (2019). https://doi.org/10.1016/j.media.2019.03.006

[29] Hu, X., Kang, M., Huang, W., Scott, M.R., Wiest, R., Reyes, M.: Dual-stream pyramid registration network. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II, pp. 382–390 (2019). Springer

[30] Mok, T.C., Chung, A.: Fast symmetric diffeomorphic image registration with convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4644–4653 (2020)

[31] De Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Išgum, I.: A deep learning framework for unsupervised affine and deformable image registration. Medical image analysis **52**, 128–143 (2019)

[32] Zhao, S., Lau, T., Luo, J., Chang, E.I.-C., Xu, Y.: Unsupervised 3d end-to-end medical image registration with volume tweening network. IEEE Journal of Biomedical and Health Informatics **24**(5), 1394–1404 (2020). https://doi.org/10.1109/JBHI.2019.2951024

[33] Zhang, J., An, C., Dai, J., Amador, M., Bartsch, D.-U., Borooah, S., Freeman, W.R., Nguyen, T.Q.: Joint vessel segmentation and deformable registration on multi-modal retinal images based on style transfer. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 839–843 (2019). IEEE

[34] Tian, Y., Hu, Y., Ma, Y., Hao, H., Mou, L., Yang, J., Zhao, Y., Liu, J.: Multi-scale u-net with edge guidance for multimodal retinal image deformable registration. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 1360–1363 (2020). IEEE

[35] Sui, X., Zheng, Y., Jiang, Y., Jiao, W., Ding, Y.: Deep multispectral image registration network. Computerized Medical Imaging and Graphics **87**, 101815 (2021)

[36] Benvenuto, G.A., Colnago, M., Casaca, W.: Unsupervised deep learning

network for deformable fundus image registration. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1281–1285 (2022). IEEE

[37] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision, pp. 205–218 (2022). Springer

[38] Petit, O., Thome, N., Rambour, C., Themyr, L., Collins, T., Soler, L.: U-net transformer: Self and cross attention for medical image segmentation. In: Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12, pp. 267–276 (2021). Springer

[39] Wang, H., Cao, P., Wang, J., Zaiane, O.R.: Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 2441–2449 (2022)

[40] Chen, B., Zou, X., Zhang, Y., Li, J., Li, K., Xing, J., Tao, P.: Leformer: A hybrid cnn-transformer architecture for accurate lake extraction from remote sensing imagery. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5710–5714 (2024). IEEE

[41] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[42] Ding, X., Zhang, X., Han, J., Ding, G.: Diverse branch block: Building a convolution as an inception-like unit. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10886–10895 (2021)

[43] Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J.: Repvgg: Making vgg-style convnets great again. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13733–13742 (2021)

[44] Hernandez-Matas, C., Zabulis, X., Triantafyllou, A., Anyfanti, P., Douma, S., Argyros, A.A.: FIRE: Fundus Image Registration Dataset. https://doi.org/10.35119/maio.v1i4.42

[45] Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. Journal of Cognitive Neuroscience **19**(9), 1498–1507 (2007).

https://doi.org/10.1162/jocn.2007.19.9.1498

[46] Hering, A., Hansen, L., Mok, T.C., Chung, A.C., Siebert, H., Häger, S., Lange, A., Kuckertz, S., Heldmann, S., Shao, W., *et al.*: Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. IEEE Transactions on Medical Imaging **42**(3), 697–712 (2022)

[47] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)

[48] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)

[49] 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)

[50] Wu, H., Zhao, Z., Wang, Z.: Meta-unet: Multi-scale efficient transformer attention unet for fast and high-accuracy polyp segmentation. IEEE Transactions on Automation Science and Engineering (2023)

[51] Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C.: A reproducible evaluation of ants similarity metric performance in brain image registration. Neuroimage **54**(3), 2033–2044 (2011)

[52] Lv, J., Wang, Z., Shi, H., Zhang, H., Wang, S., Wang, Y., Li, Q.: Joint progressive and coarse-to-fine registration of brain mri via deformation field integration and non-rigid feature fusion. IEEE Transactions on Medical Imaging **41**(10), 2788–2802 (2022)

[53] Siebert, H., Hansen, L., Heinrich, M.P.: Fast 3d registration with accurate optimisation and little learning for learn2reg 2021. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 174–179 (2021). Springer

[54] Mok, T.C., Chung, A.C.: Conditional deformable image registration with convolutional neural network. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24, pp. 35–45 (2021). Springer

[55] Jia, X., Bartlett, J., Chen, W., Song, S., Zhang, T., Cheng, X., Lu, W., Qiu, Z., Duan, J.: Fourier-net: Fast image registration with band-limited deformation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 1015–1023 (2023)

Qiushi Nie received the B.Eng. degree in computer science from the Southern University of Science and Technology (SUSTech), Shenzhen, China, in 2022. He is currently pursuing the master's degree with Department of Computer Science and Engineering in SUSTech.

His research interests include deep learning and medical image registration.

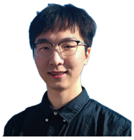
E-mail: 12232413@mail.sustech.edu.cn

ORCID iD: 0009-0003-8414-0843

Xiaoqing Zhang received the B.S. degree in water conservancy and hydropower engineering from South China Agricultural University in 2016, the M.S. degree in computer technology from Zhengzhou University in 2019, and the Ph.D. degree in mechanics from the Southern University of Science and Technology (SUSTech) in 2023. He is currently a postdoctoral fellow in the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences.

His research interests include feature disentanglement, interpretable artificial intelligence, long-tailed learning, and medical image processing.

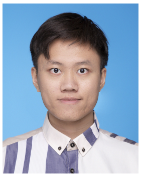
E-mail: xq.zhang2@siat.ac.cn



Chuan Chen received the B.Sc. degree in Computer Science and Technology from the Southern University of Science and Technology (SUSTech), Shenzhen, China, in 2024. He is currently pursuing the master's degree with Department of Computer Science in Johns Hopkins University.

His research interests include deep learning, medical image registration, depth estimation and NLP.

E-mail: cchen307@jh.edu



Zhixuan Zhang obtained the B.Eng. degree in computer science from the Southern University of Science and Technology (SUSTech) in Shenzhen, China, in 2024. Currently, he is pursuing the master's degree in the Department of Computer Science and Engineering at SUSTech.

His research interests focus on medical image segmentation.

E-mail: 12432725@mail.sustech.edu.cn

Yan Hu received the Ph.D. degree from the Department of Information Science and Technology, the University of Tokyo, Japan. She is working now in the Southern University of Science and Technology, China.

Her research interests include medical image analysis, surgery video processing, and computer-aided surgery.

E-mail: huy3@sustech.edu.cn

Jiang Liu received the B.Sc. degree in computer science from University of Science and Technology of China, China in 1988, the M.Sc. and Ph.D. degrees from National University of Singapore, Singapore in 1992 and 2004, respectively. He founded the Intelligent Medical Imaging Research Team which was once the world's largest ophthalmic medical image processing team, focusing on ophthalmic artificial intelligence research. Currently, he is a professor in Department of Computer Science and Engineering, Southern University of Science and Technology, China.

His research interests include artificial intelligence, eye-brain research, precision medicine, and surgical robots.

E-mail: liuj@sustech.edu.cn (Corresponding author)

ORCID iD: 0000-0001-6281-6505