IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 6, 2019, accepted February 20, 2019, date of publication March 1, 2019, date of current version March 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2901580

Deep Group-Wise Registration for Multi-Spectral Images From Fundus Images

TONGTONG CHE[1], YUANJIE ZHENG[1,2], JINYU CONG[1], YANYUN JIANG[1], YI NIU[1], WANZHEN JIAO[3], BOJUN ZHAO[3], AND YANHUI DING[1]

[1]School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China
[2]Key Lab of Intelligent Computing and Information Security in Universities of Shandong, Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology, Institute of Biomedical Sciences, Shandong Normal University, Jinan 250358, China
[3]Department of Ophthalmology, Shandong Provincial Hospital Affiliated to Shandong University, Jinan 250021, China

Corresponding authors: Yuanjie Zheng (zhengyuanjie@gmail.com) and Yanhui Ding (yanhuiding@126.com)

This work was supported in part by the National Nature Science Foundation of China under Grant 61572300, Grant 81871508, and Grant 61773246, in part by the Taishan Scholar Program of Shandong Province of China under Grant TSHW201502038, in part by the Major Program of Shandong Province Natural Science Foundation under Grant ZR2018ZB0419, and in part by the Primary Research and Development Plan of Shandong Province under Grant 2017GGX10112.

ABSTRACT Multi-spectral imaging (MSI) is a novel non-invasive tool for visualizing the entire span of the eye, from the internal limiting membrane to the choroid. However, spatial misalignments can be frequently observed in sequential MSI images because the eye saccade movement is usually faster than the MSI image acquisition speed. Therefore, registering MSI images is necessary for computer-based analysis of retinal degeneration via MSI. In this paper, we propose an early deep learning framework for achieving an accurate registration of MSI images in a group-wise fashion. The framework contains three parts: a template construction based on principal component analysis, a deformation field calculation, and a spatial transformation. The framework is uniquely capable of resolving two key challenges, i.e., the ''multi-modal'' characteristics in MSI images for the acquisition with different spectra and the requirement of joint registration of the sequential images. Our experimental results demonstrate the superior performance of our framework compared to several representative state-of-the-art techniques in both speed and accuracy.

INDEX TERMS Multi-spectral images, group-wise registration, deep learning, mono/multi-modal images.

I. INTRODUCTION

Ocular diseases, such as diabetic retinopathy, glaucoma, and age-related macular degeneration, have long been considered the leading causes of vision impairment or blindness. Ophthalmic fundus imaging has evolved to become a fundamental diagnostic tool for understanding ocular diseases [1] to effectively prevent visual impairment. Furthermore, as an advanced retinal imaging technique, multi-spectral imaging has a great ability to help doctors identify, interpret, and diagnose disease processes earlier than conventional imaging modalities. This enhanced diagnostic capability is due to the production of a series of monochrome slices throughout the entire thickness of the retina through the use of multiple wavelengths of light, as shown in Fig. 1. These images include 11 wavelengths of green, yellow, amber, red, and infrared ranging from 550 nm to 850 nm. The combined image obtained by combining different wavelengths of monochromatic light slices is of great significance to the diagnostic capabilities of ophthalmologists. For example, the image obtained by the combination of red and green light (550 nm + 620 nm) approximates the color fundus image. In addition, due to the different information contained in different spectra, the joint consideration of intra-subject information from multiple monochromatic light slices by ophthalmologists has helped improve the diagnostic accuracy.

However, there are two major bottlenecks in practice: First, the spatial misalignment of a sequence of images occurs due to the eye saccade movement being faster than the MSI image acquisition speed. Second, to track the condition of the disease for diagnosis and treatment, the same subject usually has to be scanned many times at different times. Therefore, these fundus images acquired at different times will produce some dislocation in space.Because it is difficult to see the misalignment between multi-spectral slices with the naked eye, we use the blood vessel image labeled by

The associate editor coordinating the review of this manuscript and approving it for publication was Huimin Lu.

2169-3536 © 2019 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

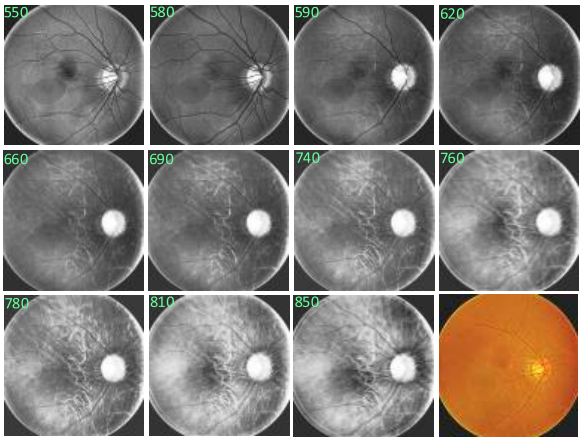


FIGURE 1. A representative sequence of MSI images from RHA arranged in order of wavelength. From left to right and top to bottom, the images are captured with short wavelengths of green (MSI-550), yellow (MSI-580) and amber (MSI-590), followed by 4 wavelengths of red (MSI-620, MSI-660, MSI-690 and MSI-740), 4 wavelengths of infrared (MSI-760, MSI-780, MSI-810 and MSI-850), and the combination of red and green light (550 nm + 620 nm).

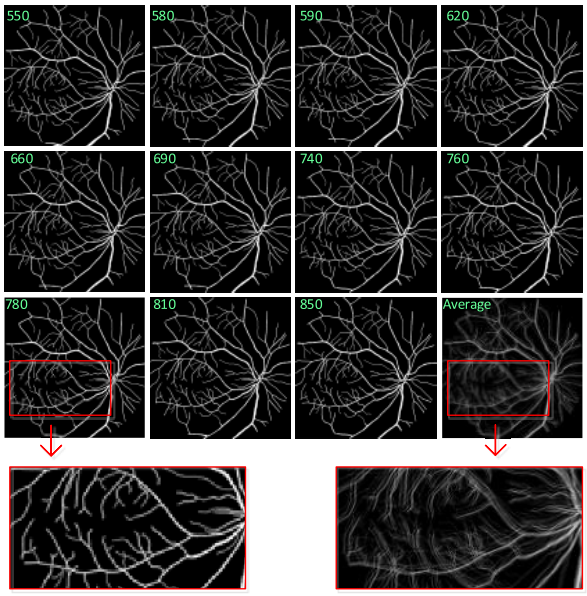


FIGURE 2. An example sequence of blood vessel images labeled by an ophthalmologist, in which the first 11 images are segmented blood vessel images corresponding to the first 11 images of Fig. 1, the last image in the third row is the average image calculated from the first 11 sequence images and the two images from left to right in the fourth row are an enlarged partial view of the MSI-740 nm blood vessel image and an enlarged partial view of the average image.

an ophthalmologist to calculate the average image of the MSI sequence images. As shown in Fig. 2, we can clearly see that spatial misalignment does exist in sequential MSI images. A computer-based algorithm of image registration has the ability to register all the above sequence images into a common space. Image registration is a key stage in image fusion [2], building a smart healthcare system [3], etc. The goal of image registration is to establish a one-to-one correspondence map between the source image and the target image. This map would greatly aid doctors in diagnosing ocular conditions by making full use of the benefits originating from MSI data.

A variety of registration algorithms have been proposed to deal with the spatial misalignment problem in different ways, including image registration in the ophthalmic field and general image registration. Existing retinal image registration methods [4], [5] display superior performance by employing point correspondences. Additionally, some vessel-based methods have been proposed [6], [7] that utilize retinal vascular features as a basis to achieve image registration. Unfortunately, these algorithms are only aimed at a pair of retinal images of the same modality, and the result of registration relies on the accuracy of vascular segmentation. In recent years, there has been an increasing interest in simultaneously aligning more than two images using a group-wise registration method to provide more useful information. Given a group of moving images and a common template image, this alignment involves mapping the moving images into a common reference space where the coordinate space is that of the template image. For group-wise registration in MSI, a feature-point-matching-based framework [8] was proposed to implement the joint alignment of multi-spectral images. However, the method uses traditional machine learning methods, which are very expensive and time consuming compared to deep learning methods.

The standard group-wise registration algorithms can be further subdivided into two classes: mono-modal sequence image-based methods and multi-modal sequence image-based methods. For the registration of mono-modal sequence images, some methods guided by templates have been proposed to deal with the spatial misalignment problem. Several template selection-based approaches [9], [10] have been demonstrated to be effective in the group-wise registration of brain images. These methods aim to select a real image in the group as a reference template, the target image is warped, and a series of intermediate templates may be traversed until the final template is obtained. There are also some approaches to take one image in the group as a template and register all other images to this template in a pair-wise manner. However, there is no standard for the selection of a template, and the resulting transformations and subsequent data analysis are easily biased towards the selected reference. Especially for MSI images, the template selection based on a single image will lose the function or structure information from other wavelength slices because different spectral slices (penetrating different light-absorbing species) may be associated with different anatomical structures. Template generation-based methods have also been presented in [11] and [12] to construct simulated template images that are more similar to the target images. The approach to mono-modal inter-subject data [13] iteratively calculates the group mean image and normalizes mutual information [14] as a pair-wise similarity metric to compare every image in the group to the average image. In [11], principal component analysis (PCA)

is exploited to establish a statistical model of the simulated deformation fields and generate multiple intermediate templates. These methods are not appropriate for multi-modal data, as multi-modal data have different scales, ranges, and contrast intensities.

Two recent works [15], [16] present effective strategies to address the problem of group-wise registration of multi-modal images by a hierarchical intensity-space subdivision scheme, and the approach of conditional template entropy, respectively. A significant drawback is that the performance of these strategies relies on the intensity information from a collection of images with manually identified landmarks. In the present work, we model the registration function using an encoder-decoder for extracting features from the template and moving images without annotation information in order to output deformation fields directly.

Because manual landmark and feature extraction in traditional method-based image registration is tedious and time consuming, many deep learning-based methods have been proposed that efficiently register images without iterative optimization or parameter tuning during testing. These methods include supervised-based approaches [17]–[19] and the unsupervised registration framework [20], [21], which uses a convolutional neural network to calculate the deformation field. Additionally, recent work has explored the unsupervised-based registration methods to learn similarity metrics from a pair of multi-modal images [22], showing their potential to outperform traditional methods in particular applications. However, these frameworks are limited to bi-modal images. That is, existing methods can only solve pair-wise registration. Considering a set of mono/multi-modal images, defining an effective similarity metric to guide global matching across modalities with these approaches is hard. In view of the successful applications of deep learning in computer vision [23], [24]. The aim of our study is to use deep learning to address the group-wise registration problem of MSI that simultaneously aligns more than two images.

In this paper, we propose a deep group-wise registration network based on the principle image with an unsupervised CNN, which is suitable for the group-wise registration of mono/multi-modal sequential fundus images. The proposed network takes a group of moving images and an unbiased template image based on PCA as inputs and outputs warped moving images. Joint optimization is employed to handle similarity measures corresponding to all images in a group. The experimental results show that this approach achieves good performance in group-wise registration. The main contributions can be summarized as follows:

1) Instead of pair-wise registration, we employ joint optimization by using a deep learning-based architecture to solve the problem of group-wise registration.

2) To speed up the convergence during training and address the bias of the template, we utilize an iteratively updated representative template image based on PCA.

3) The proposed network is used to align multiple mono-modal MSI images corresponding to the same wavelength scanned at different times and to align multi-modal images corresponding to different wavelengths.

4) To the best of our knowledge, this is the first universal deep learning method for unsupervised deformable group-wise registration of mono/multi-modal MSI data. The framework is approximated as an encoder-decoder that improves the registration accuracy and efficiency over existing registration methods and has great potential to be applied in real applications.

II. METHODS

In group-wise registration, given a collection of moving images $M_1, M_2, \ldots, M_N$ and a template image T, the goal of group-wise registration is to register N moving images to the template image such that all deformed moving images are similar to the template image T. In this paper, we propose to train an unsupervised deep CNN for group-wise registration of mono/multi-modal MSI images. The proposed network, depicted in Fig. 3, consists of a convolutional neural network, template construction, and a spatial transformer. This approach is based on the paradigm in which the similarity of the group of images is measured with respect to an iteratively updated template image. The following subsections introduce these aspects and a possible architecture.

A. CONVOLUTIONAL ARCHITECTURE

As shown in Fig. 4, the convolutional architecture of the group-wise registration network is based on U-net [25]. The architecture for the deformation field (ϕ_i) consists of two parts: an encoder and a decoder. Moving images are sequentially paired with the template image as input to the encoder. The encoder consists of 3×3 convolutions and 2×2 downsampling layers that learn features from the moving/template images. We apply convolutions followed by rectified linear unit (ReLU) activations [26] and batch normalization (BN) [27] with a stride of two. Every two convolutional layers are followed by an average pooling layer that can retain the most information and reduces the number of network parameters during downsampling. In the decoding stage, each step has a 2×2 deconvolution layer and two 3×3 convolution layers followed by ReLUs and BN. Additionally, the registration is directly generated by concatenating skip connections to the deconvolution that concatenates the high pixel features extracted at the encoding stage to the new feature maps in the decoding stage, and 1×1 kernels are applied to the last convolutional layer. For the group-wise registration of multi-modal images, because the moving and template images have different modalities with significantly different spatial resolutions, it is not necessarily optimal to have shared convolutional weights. Thus, the convolutional weights applied to the moving and template images are not shared, which adds several additional free weights to optimize compared with the task of mono-modal image registration.

B. TEMPLATE CONSTRUCTION

Given N images to be registered in a group, to find a more representative template image, we adopt the idea of

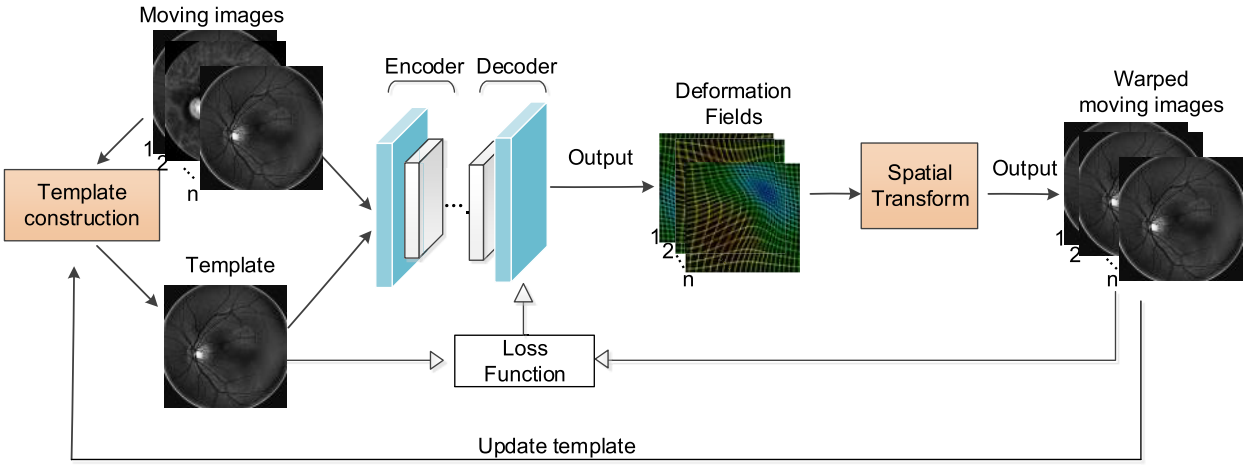


FIGURE 3. Overview of our proposed group-wise registration network. The input to the network consists of N moving images $M_1, M_2 \ldots, M_N$ in a group and the template image generated by PCA. The output of the network is N deformation fields, where each deformation field corresponds to a moving image. The subsequent spatial transform uses bilinear interpolation to obtain dense spatial transformations for registering 2D images by optimizing an image similarity metric.

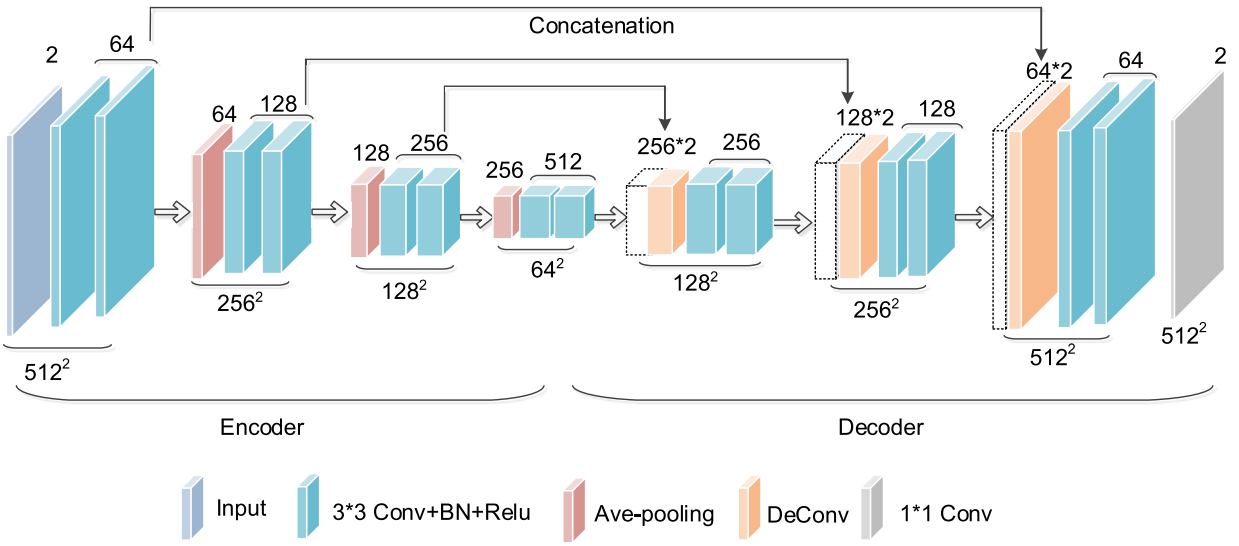


FIGURE 4. Proposed convolutional architecture of the group-wise registration network. Moving images are sequentially paired with the template image of size 512 × 512 as an input to the network. The number of channels is printed on the upper side of the cube, and the number below the cube indicates the size of the feature map. Skip connections are applied to concatenate the feature maps extracted at the encoding stage to the new feature maps. The output deformation field is the same size as the input image.

principal component analysis (PCA) combined with deep learning to generate a template image containing the principal components of the N images in the group, instead of using the average image as a template. PCA is regarded as linear dimensionality reduction using singular value decomposition of the data to project them to a lower dimensional space while retaining as many differences as possible. In our method, the pixel coordinate of each image is sampled as a separate observation, different images are taken as different variables corresponding to an N-dimensional space, and PCA is used to reduce the dimension to a one-dimensional subspace. The eigenvector V associated with the largest eigenvalue can serve as the weights w for the construction of the template image T.

$$V = (w_1, w_2, w_3 \ldots .w_N) \tag{1}$$

$$T(x) = \sum_{i=1}^{N} VM_i \circ \phi_i(x) \tag{2}$$

Here, V is a linear combination of the variables and contains all observations.

C. SPATIAL TRANSFORMATION FUNCTION

The optimal registration network parameters are found by minimizing the differences between $M_i \circ \phi_i$ and T. We deploy a differentiable spatial transformation operation based on spatial transformer networks [28] to compute $M_i \circ \phi_i$. We warp M_i with ϕ_i to $M_i \circ \phi_i$ using a spatial transformation function, enabling the model to evaluate the similarity of each warped moving image and template image. The spatial

transformation operation with bilinear interpolation is formulated as:

$$M_i' = M_i \circ \phi_i = \sum_{\substack{x' \in \{x+\phi(x)\} \\ y' \in \{y+\phi(y)\}}} M_i(x', y')(1 - |\phi(x) - x'|)(1 - |\phi(y) - y'|) \tag{3}$$

where M_i' is warped from M_i by ϕ_i, and for each pixel (x,y), we compute a pixel location $\phi(x)$ and $\phi(y)$ in M_i. $\{x + \phi(x)\}, \{y + \phi(y)\}$ denotes the 4-pixel neighborhood around the location$(x + \phi(x), y + \phi(y))$. Variables x and y indicate two directions in the 2D image space. To allow back-propagation of the loss during optimization, similar to [28], we compute the gradient of the spatial transformation with respect to the location (x,y) by taking the partial derivatives of Eq.(3).

D. LOSS FUNCTION

In our proposed method, our network is trained by minimizing the loss function to maximize the image similarity. The loss layer evaluates the registration loss between the template and each deformed moving image as formulated in Eq.(4).

$$L = \frac{1}{N} \sum_{i=1}^{N} (-L_{sim}(T^{pca}, M_i \circ \phi_i) + L_{smooth}(\phi_i)) \tag{4}$$

where the term L_{sim} measures the image similarity between the template image T^{pca} and each warped moving image M_i. Particularly, the template image T^{pca} is computed by principal component analysis of the current warped moving images in a group. L_{smooth} is the smoothness of the transformation estimated by the registration network. In our experiments, mutual information (MI) is used to calculate the similarity between the images of different modalities. L_{sim} is defined as:

$$L_{sim} = argmax_{\phi_i, M_i}(H(T) + H(M_i \circ \phi_i) - H(T, M_i \circ \phi_i)) \tag{5}$$

where the first two terms $H(T)$ and $H(M_i \circ \phi_i)$ are the marginal entropies of the marginal intensity distributions in the template and warped moving images, respectively. The final term denotes the joint entropy of the joint intensity distribution between the template image and warped moving image. Each warped moving image M_i' gradually approximates the iteratively optimized template image by maximizing the image similarity L_{sim}. The optimal deformation transformation is obtained by maximizing L_{sim}, and the optimal weight of the encoder-decoder network layers is obtaining by backpropagating the dissimilarity between the moving images and the template image using stochastic gradient descent. Our method iteratively updates the deformation fields based on the gradient of the energy function L. In this way, we register all images of the group in a common coordinate frame at the end. To ensure the continuity of deformation, we set the regularization term L_{smooth} by constraining the smoothness of the deformation field ϕ_i. The regularization is defined as:

$$L_{smooth} = \lambda_1 \|\nabla^2 \phi\|^2 + \lambda_2 \|\phi\|^2 \tag{6}$$

where ∇^2 represents the Laplacian operator, and λ_1 and λ_2 are the weighting parameters that control the balance between the image similarity measure and the regularization on the spatial transformation.

III. EXPERIMENTS

A. DATASET

We demonstrate our method on the task of mono-modal group-wise registration and multi-modal group-wise registration of MSI images. The experimental dataset is collected from an Annidis RHA (Annidis Health Systems Corp Ottawa, Canada). RHA is based on multi-spectral imaging. Monochromatic LED lights produce 11 monochrome slices for a comprehensive evaluation of the retina from shallow to deep (RPE) and to the choroid. These images are of the oculus dexter (OD) and oculus sinister (OS) of 27 healthy subjects and 73 patients with fundus lesions. Thirty patients were scanned for at least 4 sets of MSI images taken at different times. They are provided in the dicom format with a bit depth of 16 and a size of 2048 × 2048. All images were resized to 512 × 512. All datasets were augmented by adding three rotated (90°, 180°, 270°) and two flipped (left-right, up-down) variants for each image. For the group-wise registration of multi-modal images, the dataset consists of 990 sets of MSI sequence images. We split our dataset into 890, 50 and 50 sets for training, validation, and testing, respectively. Each set of sequence images comes from multi-spectral images of 11 different wavelengths. For the group-wise registration of mono-modal images, the dataset contains a total of 2475 sets of sequence images, which are divided into 2275, 100, and 100 sets for training, validation, and testing, respectively. Each set of sequence images is produced from multiple MSI images of the same wavelength scanned at different times.

B. IMPLEMENTATION DETAILS

The network is implemented in PyTorch [29], and Adam optimization [30] was used to train the network, with a learning rate of 0.001. We train networks to optimize the evaluation results on the validation set and report the results on the test set that we retain. The network is trained on two E5-2630U4 CPUs and four NVIDIA Tesla V100 GPUs. Our proposed network is trained separately on the mono/multi-modal MSI datasets. For the group-wise registration of multi-modal images from MSI, monochromatic slices of 11 different modes are used as a set of moving images. Additionally, monochrome slices of the same wavelength acquired at different times (in our experiments, scanned 4 times) were used as the moving images for the group-wise registration of mono-modal images. Their template images are a linear

combination of moving images in a group that was obtained by PCA. In training, the moving images are sequentially paired with the template image as input to the network. These moving images anatomically correspond to slices from the same subject but acquired at different depths of the retina at different times. We train the group-wise registration network with different weight values through backpropagation until convergence. Furthermore, the value of regularization parameter λ affects the performance of the network when using different λ to train the network separately. In our implementation, λ_1 and λ_2 are empirical values set to 0.5 and 0.01, respectively. We train the network for 50 epoch.

IV. RESULTS

A. EVALUATION METRIC

We evaluate our registration performance using the Dice similarity coefficient [31], the ratio of the labeled points with the correct alignment over the threshold value [8] and the closest point distance [32].

1) DICE SIMILARITY COEFFICIENT

For all test samples, we exploit expert-labeled anatomical segmentations (vascular and optic), and the annotations in the test scan are used for final quantitative evaluation only. Each Dice score is computed using any two images in the group, and the final average Dice score of each structure is calculated over all subjects. We expect that the regions of any two images after registration M_i', M_j' in a group correspond well to the same anatomical structure overlap. Let $P^r_{M_i'}$ and $P^r_{M_j'}$ represent the set of pixels of an anatomic structure. We denote the sets in M_i' and M_j' as $P^r_{M_i'}$ and $P^r_{M_j'}$, respectively. The Dice score of two structures is defined as:

$$Dice = 2 \times \frac{P^r_{M_i'} \bigcap P^r_{M_j'}}{|P^r_{M_i'}| + |P^r_{M_j'}|} \tag{7}$$

The closer the Dice value is to 1, the better the overlap of the two structures, that is, the better the registration performance. This evaluation measure evaluates the degree of matching to individual anatomical regions as well as the total image volume.

2) CLOSEST POINT DISTANCE

To avoid the use of the DSC of the vessel tree segmentation, which will have an impact on the assessment results due to the sensitivity of vascular segmentation, we also evaluate the accuracy of the registration by measuring the closest point distance. First, the ophthalmologist manually labels the obvious vessel intersections based on all of the previously labeled vessel tree images. Then, for each intersection in each MSI image after registration, the closest point distance algorithm searches for the nearest neighbor among the other corresponding images in the group. Finally, the ratio of points that are correctly matched is calculated through comparison with the ground-truth.

3) RATIO OF REGISTRATION

For each test set of sequence images, the pathologist manually picks 15 points in each MSI image and then annotates them based on MRIcron [33]. We calculate the distance between the manually labeled points in each MSI image after registration and the corresponding points in the other images in the group. The registration is considered successful if the distance divided by the radius of the retinal image after preprocessing (257 pixels in our experiment) is less than the set threshold (t). Then, the ratio of registration is obtained based on the number of manually labeled points with correct alignment over the threshold values. The ratio of registration is defined as:

$$Ratio = \frac{Q(d(M_i'(S_k)), (M_j'(S_k)) < t)}{15C_N^2} \tag{8}$$

where Q denotes the number of points where the distance between the corresponding points between any two images is less than the threshold, M_i', and M_j' are any two images after registration, $\{S_k | k \in [1, 15]\}$ represents the set of manually labeled points and N is the number of images in a group.

B. TEMPLATE EVALUATION

To illustrate the importance of selecting representative template images, especially the template of a set of multi-modal images, two experiments are performed on MSI images of different wavelengths. First, the average image and the PCA image are separately used as a template for comparison of sharpness changes during training. Second, the effect of template images on the registration performance is studied with respect to a single image, the average image, and the PCA image in the group.

Fig. 5 shows the process of changing the template image during the training of a group-wise registration experiment. Conventional methods of group-wise registration use a very fuzzy group mean image as the template, as shown in the top row of Fig. 5. In contrast, our method begins with a clear PCA image, as shown in the second row of Fig. 5, which is close to the population center. The evolutions of the template image using a conventional method and our proposed method are provided in the top row and second row of Fig. 5, respectively. At 30 epochs, the sharpness of the PCA template image is already very high and far exceeds that of the average template image. Clearly, the technique of combining PCA template images with deep networks converges faster than that using an average template image. The main reason for this fast convergence is that the PCA-based template image jointly considers the principal component information of a set of MSI sequence images, making the template image a more representative image.

Additionally, to emphasize the effectiveness of PCA-based templates for group-wise registration, the proposed group-wise registration network was compared to 6 other methods based on single template selection and on an average image [34] in a multi-modal group-wise registration task. For the template selection based on a single image in a group,

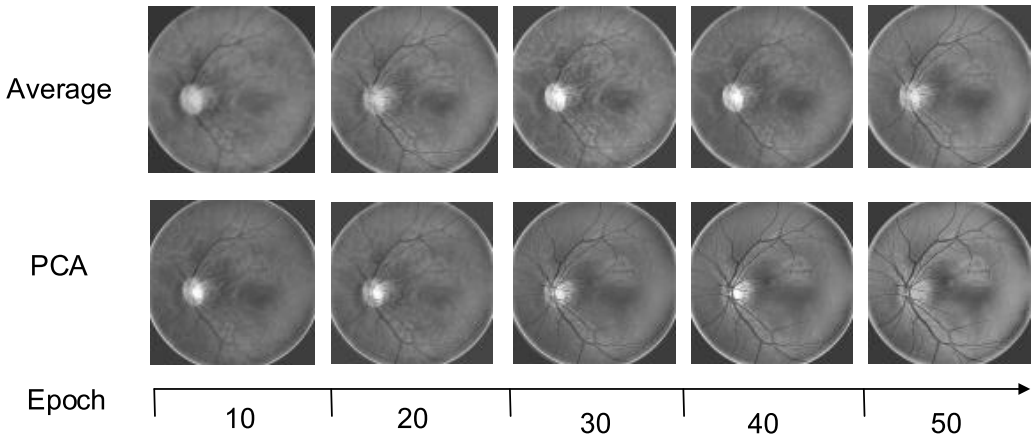


FIGURE 5. The process of changing the template image during training. The top row shows how the template image changes as the number of iterations increases when the proposed network uses the average image as a template. The next row indicates the change in the template image when the PCA image is used as a template.

TABLE 1. The overlap evaluation results for different template-based group-wise registrations of multi-modal images from MSI. Comparison of the average Dice scores for the approaches based on a single image (550 nm, 590 nm, 620 nm, 660 nm, and 740 nm), the average image, and the PCA image in the group.

Dice	Optic	Vascular
550 nm	0.913±0.047	0.711±0.042
590 nm	0.906±0.073	0.709±0.039
620 nm	0.919±0.053	0.714±0.036
660 nm	0.925±0.072	0.712±0.061
740 nm	0.893±0.045	0.703±0.034
Average	0.921±0.074	0.715±0.064
PCA	**0.967±0.046**	**0.732±0.051**

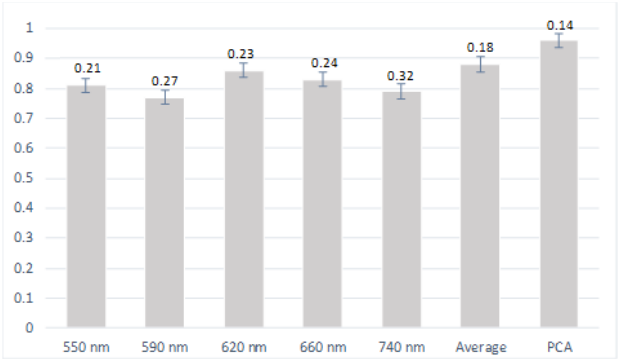


FIGURE 6. Accuracy of registration based on different templates determined by measuring the closest point distance of labeled vessel tree images. The x-axis denotes 7 different methods, while the y-axis is the ratio of the correct registration at the intersection. The variance is displayed above each blue rectangle.

we exploit the MSI images acquired at 550 nm, 590 nm, 620 nm, 660 nm, and 740 nm as the template image and register images acquired at other wavelengths to this template image. Note that the shorter the wavelength is in these wavelength-dependent MSI images, the higher the sharpness, i.e., the 550 nm MSI image is the clearest. Particularly, we delete the template construction from our PCA-based group-wise registration network for the experiments based on a single image and implement pair-wise registration.

By analyzing the average Dice scores shown in Table 1 and the registration correct rate distributions shown in Fig. 6, we can clearly see that the average Dice scores of the registration based on template selection (550 nm, 590 nm, 620 nm, 660 nm, and 740 nm) is lower than those based on template generation (average and PCA). At the same time, the accuracy produced by the clearest template image of 550 nm is inferior to that produced by the 620 nm and 660 nm images in the template selection-based registration. Thus, if we choose a template from an equally qualified MSI sequence image, even the clearest one may produce sub-optimal results. There is no standard to select the template image, which can cause biased transformations, and the subsequent data analysis can be easily biased towards the selected template. Additionally, the proposed network based on PCA achieved the best result because the PCA image considers the principal component information of a set of MSI sequence images fairly and comprehensively.

C. PERFORMANCE EVALUATION

Due to the superior performance of global affine alignment [35] and SyN [36] in traditional registration tasks, we compare the proposed method to these methods in registration experiments of MSI images. The proposed deep learning-based registration algorithm is improved in time and precision compared to traditional methods. We exploit DIRNet [20] with MI, implementing pair-wise registration of MSI images in a group for comparison with the proposed group-wise registration network. We implement pair-wise registration of MSI images based on affine alignment, SyN, and DIRNet. Affine transformation and SyN implementation depend on the ANTs [36] software package, with the MI similarity measure. The proposed network is also compared with the average mutual information (AMI) [34] in a group-wise registration task that uses the average image as a template image.

TABLE 2. Overlap evaluation results for the group-wise registration of MSI images, and comparison of the average Dice scores for global affine alignment, SyN, AMI, DIRNet and the proposed network, including the group-wise registration of multi-modal images from different wavelength fundus images and the group-wise registration of mono-modal images from different acquisition times at the same wavelength.

Dice		Affine	SyN	DIRNet	AMI	Proposed
Multi-modal	Optic	0.833±0.085	0.875±0.062	0.861±0.078	0.921±0.074	**0.967±0.046**
	Vascular	0.407±0.081	0.636±0.072	0.654±0.074	0.715±0.064	**0.732±0.051**
Mono-modal	Optic	0.851±0.052	0.882±0.058	0.893±0.062	0.953±0.053	**0.971±0.021**
	Vascular	0.502±0.064	0.653±0.060	0.644±0.062	0.736±0.054	**0.742±0.031**

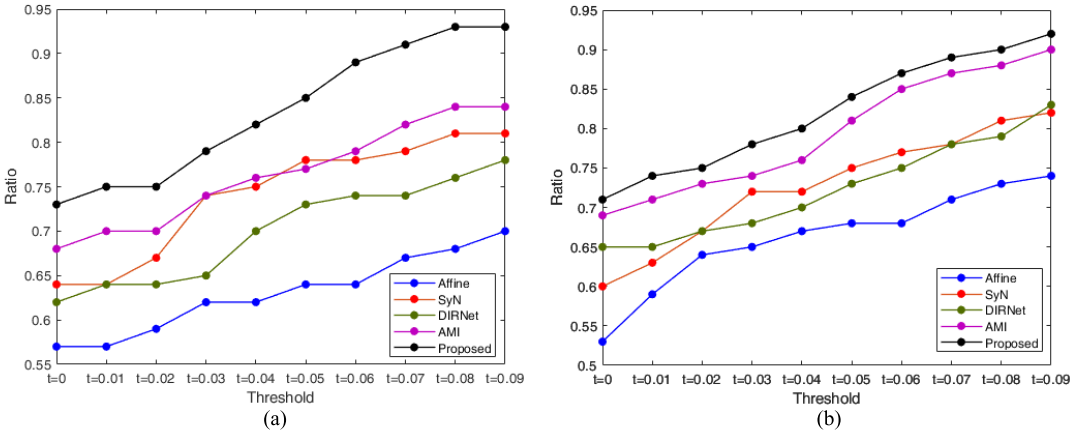


FIGURE 7. Performance curves for group-wise registration of MSI. Comparison of average ratio global affine alignment, SyN, DIRNet, AMI and our proposed network. The y-axis denotes the ratio of correct registration while the x-axis is the threshold. (a) The group-wise registration of multi-modal images. (b) The group-wise registration of mono-modal images.

For the registration of multi-modal images from MSI, the average Dice value before registration is particularly low (0.236 ± 0.034) because of the structural specificity of the fundus image. In particular, the distribution of the vascular structure is not uniform. In Table 2, we can see from the results of the two experiments that our proposed approach achieved the best accuracy. This result indicates that the proposed network outperforms the pair-wise registration in a joint optimal manner. Furthermore, the proposed group-wise registration network PCA-based method is superior to the other baseline methods. Regarding AMI with the average image as a template for the group-wise registration, AMI performs comparably to the proposed network for mono-modal fundus images. The template image in our method interpreted as the PCA image usually yields a more accurate registration result for multi-modal images.

In our experiments, the DSC values based on vessel segmentation are generally low due to the complexity of the vascular structure and the sensitivity of segmentation. For the group-wise registration of multi-modal images, we further evaluate the registration performance of the algorithm by measuring the closest point distance of labeled vessel tree images. From Fig. 8, it is not difficult to find a clear distinction between the registration accuracies for significant points in the vessel tree obtained by different methods. Additionally, we have two findings. First, group-wise registration outperforms pair-wise registration. The major reason is that different spectral slices may be associated with different anatomical structures. However, pair-wise registration relies on the independent processes of pair-wise images rather than on the joint information of MSI images, possibly leading to only a fraction of the total information available within the group of images being used in each pair-wise registration. Second, the proposed PCA-based group-wise registration network is far superior to AMI because of the large differences in the appearance of multi-modal images. The group average image is very blurry compared to PCA images.

We plot the corresponding point-match ratio measure against different methods for the group-wise registration of mono-modal images and multi-modal images, as shown in Fig.7(a) and Fig.7(b), respectively. We can see that the registration performance will gradually increase with a larger threshold. However, our algorithm shows an improved accuracy result that is better than those of all the other methods. This implies that the proposed PCA-based group-wise registration network improves the algorithm performance in registration. This improvement is conferred by the deep learning-based joint optimization that iteratively updates the template image.

D. COMPUTATIONAL COST EVALUATION

Because of the different experimental environments and programming languages of the different methods, we roughly compare the computational time of our method to those of the other 4 classic registration methods in the testing stage, as reported in Table 3. The runtime results were produced on two CPUs and four GPUs. Pair-wise estimates of mutual information, such as global affine alignment and SyN, exhibit

TABLE 3. Computational cost of 5 different registration methods in the training and testing stages.

Time		Affine	SyN	DIRNet	AMI	Proposed
Multi-modal	CPU(min)	121.4±29.1	127.9±25.7	73.9±16.2	53.0±6.2	**44.1±5.8**
	GPU(min)	–	–	41.9±6.7	7.3±0.8	**3.2±1.1**
Mono-modal	CPU(min)	73.2±9.5	69.8±2.3	37.9±5.2	22.3±4.6	**20.1±4.2**
	GPU(min)	–	–	28.2±3.2	4.1±1.7	**1.4±0.6**

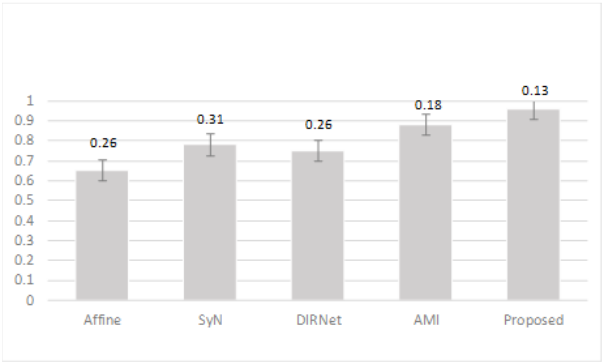


FIGURE 8. Accuracy of the registration of multi-modal images measured using the closest point distance of labeled vessel tree images. The x-axis denotes 5 different methods, while the y-axis is the ratio of the correct registration at the intersection. The variance is displayed above each blue rectangle.

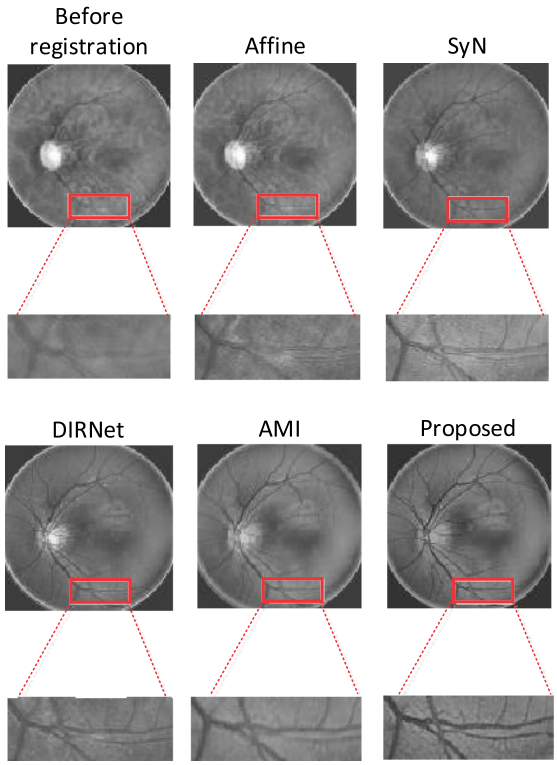


FIGURE 9. MSI registration results achieved by 7 different methods, where the upper-left image is the image before registration, and the rest of the images are images after registration by different methods. The top row shows the mean image of all subjects in a group, and the boxes represent magnifications of the local region.

a higher calculation time because all possible pairs of images in the group need to be calculated. The deep learning-based pair-wise registration algorithm DIRNet consistently outperforms the traditional affine transformation and SyN in registration accuracy and computation time. The extension to group-wise registration leads to a dramatic performance drop because more iterations are needed to optimize the registration results. As for AMI, it takes much time to converge due to the ambiguity of the average template in the group-wise registration of mono/multi-modal images. We can see that our method requires approximately 3.2 minutes for the multi-modal images and 1.4 minutes for the mono-modal images to register a group of images. These times are faster than those of the other methods used during testing. In the two sets of experiments on mono-modal image and multi-modal image registration, the registration time of the mono-modal images is less due to the different numbers of moving images (mono: 4 images; multi: 11 images).

E. VISUALIZATION RESULTS

Additionally, the results of MSI registration are visually illustrated. Taking the group-wise registration of multi-modal images from MSI as an example, Fig. 9 shows the mean image before and after registration by 5 different methods. We find that the mean of the images before registration is blurry. The mean of the registered images obtained by proposed group-wise registration network is sharper than those obtained by the other methods. This result means that the image can be best aligned by our method.

As an example of result visualization, we randomly select 3 different wavelengths (580 nm, 620 nm, and 740 nm) of MSI images with optic disc annotations to obtain the overlap of the discs. The comparison of the optic disc segmentations before and after registration for each method is shown in Fig. 10.

V. DISCUSSION

A. COMPARISON WITH PREVIOUS STUDIES

In this work, we propose a novel registration method suitable for mono/multi-modal group-wise registration based on a principal image with an unsupervised CNN. Compared with previous deep learning-based pair-wise registration approaches [37], [38] that supervise iterative training, the proposed method does not require a manually annotated ground-truth. Compared with traditional group-wise registration approaches [12], [39], our method models an end-to-end network that can automatically perform feature extraction and deformation field calculations. Furthermore, compared with the latest deep learning-based multi-modal registration approaches [22] that use image synthesis to transform multi-modal registration into mono-modal registration tasks, our framework does not depend on other processes that simulta-

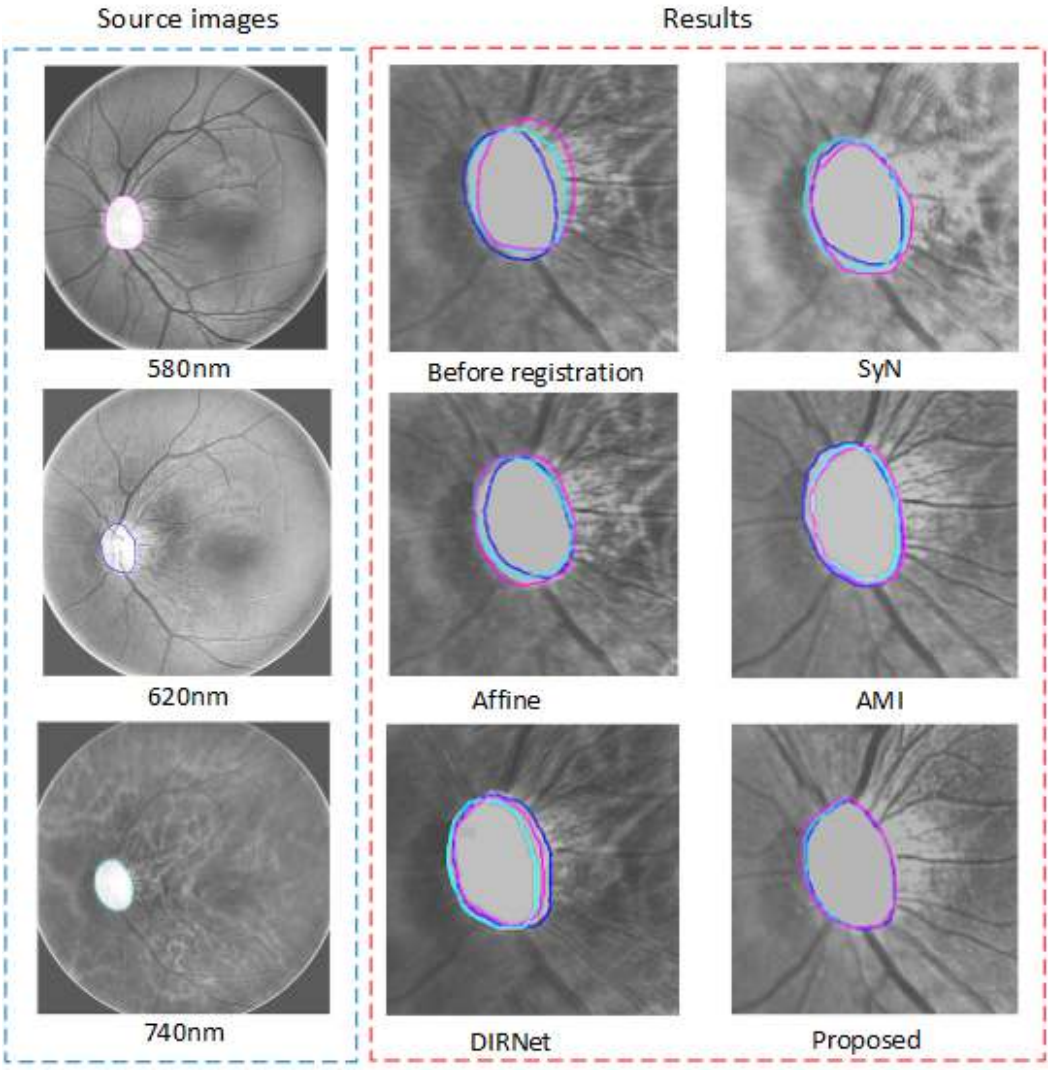


FIGURE 10. Visualization of the registration results. The source images (MSI images acquired at 580 nm, 620 nm, and 740 nm) marked with the disc area are displayed in the blue box. Registration results obtained using Syn, affine, AMI, DIRNet and the proposed network are shown in the red box. In particular, the overlap of the optic discs before registration is shown at the upper left of the red box.

neously align multiple multi-modal images. The experimental results demonstrate the effectiveness of our method.

B. LIMITATIONS AND FUTURE WORK

Although our method achieves promising results, there are still several limitations. First, some subjects could not be included due to the incompleteness of the MSI sequence data. To improve this, we may need more clinical data for training. Additionally, the current network framework can only simultaneously compute two images during the feature extraction phase. We will jointly learn the characteristics of all images by considering the spatial and timing relationships among different modalities.

VI. CONCLUSION

In this work, we propose a novel registration method suitable for the group-wise registration of mono/multi-modal images from MSI based on a principal image with an unsupervised CNN. Our proposed group-wise registration network achieves accurate registration of MSI images in a group-wise fashion, with a typical feedforward and backpropagation-based deep learning setting. All images are well aligned through the iterative updating of the representative template image. Furthermore, our method models an end-to-end network that can automatically perform feature extraction and deformation field calculations. By using such a novel deep network implementing group-wise registration of mono/multi-modal MSI images, the results outperform those of the conventional learning-based group-wise registration and deep learning-based pair-wise registration methods. Such a model has very high potential for real-world applications.

REFERENCES

[1] C. Zimmer, D. Kahn, R. Clayton, P. Dugel, and K. B. Freund, "Innovation in diagnostic retinal imaging: multispectral imaging," *Retina Today*, vol. 9, no. 7, pp. 94–99, 2014.

[2] H. Lu, L. Zhang, and S. Serikawa, "Maximum local energy: An effective approach for multisensor image fusion in beyond wavelet transform domain," *Comput. Math. Appl.*, vol. 64, no. 5, pp. 996–1003, 2012.
[3] Z. Yin, R. Gravina, H. Lu, M. Villari, and G. Fortino, "PEA: Parallel electrocardiogram-based authentication for smart healthcare systems," *J. Netw. Comput. Appl.*, vol. 117, pp. 10–16, Sep. 2018.
[4] C. Hernandez-Matas, X. Zabulis, A. Triantafyllou, P. Anyfanti, and A. A. Argyros, "Retinal image registration under the assumption of a spherical eye," *Computerized Med. Imag. Graph.*, vol. 55, pp. 95–105, Jan. 2017.
[5] R. Estrada, C. Tomasi, M. T. Cabrera, D. K. Wallace, S. F. Freedman, and S. Farsiu, "Enhanced video indirect ophthalmoscopy (VIO) via robust mosaicing," *Biomed. Opt. Express*, vol. 2, no. 10, pp. 2871–2887, 2011.
[6] B. Fang and Y. Y. Tang, "Elastic registration for retinal images based on reconstructed vascular trees," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 6, pp. 1183–1187, Jun. 2006.
[7] C. V. Stewart, C.-L. Tsai, and B. Roysam, "The dual-bootstrap iterative closest point algorithm with application to retinal image registration," *IEEE Trans. Med. Imag.*, vol. 22, no. 11, pp. 1379–1394, Nov. 2003.
[8] Y. Zheng *et al.*, "Joint alignment of multispectral images via semidefinite programming," *Biomed. Opt. Express*, vol. 8, no. 2, pp. 890–901, 2017.
[9] H. Jia, P. T. Yap, G. Wu, Q. Wang, and D. Shen, "Intermediate templates guided groupwise registration of diffusion tensor images," *Neuroimage*, vol. 54, no. 2, pp. 928–939, 2011.
[10] B. C. Munsell, A. Temlyakov, and S. Wang, "Fast multiple shape correspondence by pre-organizing shape instances," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 840–847.
[11] S. Tang, Y. Fan, G. Wu, M. Kim, and D. Shen, "RABBIT: Rapid alignment of brains by building intermediate templates," *Neuroimage*, vol. 47, no. 4, pp. 1277–1287, 2009.
[12] S. Joshi, B. Davis, M. Jomier, and G. Gerig, "Unbiased diffeomorphic atlas construction for computational anatomy," *Neuroimage*, vol. 23, no. 1, pp. S151–S160, 2004.
[13] K. K. Bhatia, J. Hajnal, A. Hammers, and D. Rueckert, *Similarity Metrics for Groupwise Non-Rigid Registration*. Berlin, Germany: Springer, 2007.
[14] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3D medical image alignment," *Pattern Recognit.*, vol. 32, no. 1, pp. 71–86, 1999.
[15] Ž. Špiclin, B. Likar, and F. Pernuš, "Groupwise registration of multimodal images by an efficient joint entropy minimization scheme," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2546–2558, May 2012.
[16] M. Polfliet, S. Klein, W. Huizinga, M. M. Paulides, W. J. Niessen, and J. Vandemeulebroucke, "Intrasubject multimodal groupwise registration with the conditional template entropy," *Med. Image Anal.*, vol. 46, pp. 15–25, May 2018.
[17] J. Wulff and M. J. Black, "Efficient sparse-to-dense optical flow estimation using a learned basis and layers," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 120–130.
[18] R. Liao *et al.*, "An artificial agent for robust image registration," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2016.
[19] X. Yang, R. Kwitt, M. Styner, and M. Niethammer, "Quicksilver: Fast predictive image registration—A deep learning approach," *Neuroimage*, vol. 158, pp. 378–396, Sep. 2017.
[20] B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum, "End-to-end unsupervised deformable image registration with a convolutional neural network," in *Proc. Int. Workshop Deep Learn. Med. Image Anal.*, 2017, pp. 204–212.
[21] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca. (2018). "An unsupervised learning model for deformable medical image registration." [Online]. Available: https://arxiv.org/abs/1802.02604
[22] X. Cao, J. Yang, L. Wang, Z. Xue, Q. Wang, and D. Shen. (2018). "Deep learning based inter-modality image registration supervised by intra-modality similarity." [Online]. Available: https://arxiv.org/abs/1804.10735
[23] H. Lu *et al.*, "Wound intensity correction and segmentation with convolutional neural networks," *Concurrency Comput. Pract. Exper.*, vol. 29, no. 6, p. e3927, 2016.
[24] S. Hou, J. Lin, S. Zhou, M. Qin, W. Jia, and Y. Zheng, "Deep hierarchical representation from classifying logo-405," *Complexity*, vol. 29, no. 6, p. e3927, 2017.
[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
[26] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.
[27] S. Ioffe and C. Szegedy. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." [Online]. Available: https://arxiv.org/abs/1502.03167
[28] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. (2015). "Spatial transformer networks." [Online]. Available: https://arxiv.org/abs/1506.02025
[29] N. Ketkar, "Introduction to PyTorch," in *Deep Learning With Python*. Berkeley, CA, USA: Apress, 2017, pp. 195–208.
[30] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980
[31] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.
[32] K. Deng, J. Tian, J. Zheng, X. Zhang, X. Dai, and M. Xu, "Retinal fundus image registration via vascular structure graph matching," *Int. J. Biomed. Imag.*, vol. 2010, Jul. 2010, Art. no. 906067.
[33] C. Rorden and M. Brett, "Stereotaxic display of brain lesions," *Behavioural Neurol.*, vol. 12, no. 4, pp. 191–200, 2000.
[34] J. Ceranka, M. Polfliet, F. Lecouvet, N. Michoux, J. de Mey, and J. Vandemeulebroucke, "Registration strategies for multi-modal whole-body MRI mosaicing," *Magn. Reson. Med.*, vol. 79, no. 3, pp. 1684–1695, 2017.
[35] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," *Med. Image Anal.*, vol. 5, no. 2, pp. 143–156, Jun. 2001.
[36] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ANTs similarity metric performance in brain image registration," *NeuroImage*, vol. 54, no. 3, pp. 2033–2044, 2011.
[37] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec, "SVF-Net: Learning deformable image registration using shape matching," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2017, pp. 266–274.
[38] J. Krebs *et al.*, "Robust non-rigid registration through agent-based action learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2017, pp. 344–352.
[39] C. Wachinger, W. Wein, and N. Navab, "Three-dimensional ultrasound mosaicing," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2007, pp. 327–335.

TONGTONG CHE was born in Shandong, China, in 1995. She is currently pursuing the master's degree with the School of Information Science and Technology, Shandong Normal University of China. Her research interests include medical image processing and deep learning.

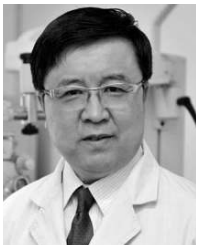
YUANJIE ZHENG was a Senior Research Investigator with the Perelman School of Medicine, University of Pennsylvania. He is also serving as the Vice Dean of the School of Information Science and Technology and the Institute of Life Sciences, Shandong Normal University. He is currently a Professor with the School of Information Science and Engineering, Shandong Normal University, and also a Taishan Scholar of People's Government of Shandong Province, China. His research interests include medical image analysis, translational medicine, computer vision, computational photography, patient care by creating algorithms for automatically quantifying and generalizing the information latent in various medical images for tasks, such as disease analysis and surgical planning through the applications of computer vision and machine learning approaches to medical image analysis tasks, and development of strategies for image-guided intervention/surgery.

JINYU CONG was born in Shandong, China, in 1991. She received the B.E. and M.E. degrees from the Shandong University of Traditional Chinese Medicine, Jinan, Shandong, China. She is currently pursuing the Ph.D. degree with Shandong Normal University, Jinan, Shandong, China. Her research interests include the medical image processing and machine learning.

YANYUN JIANG was born in Shandong, China, in 1993. She is currently pursuing the master's degree with the School of Information Science and Engineering, Shandong Normal University, Jinan, Shandong, China. Her research interests include machine learning and medical image analysis.

YI NIU is currently a Lecturer with the School of Information Science and Engineering, Shandong Normal University. Her research interests include picture analysis and nonlinear functional analysis. She devoted herself to the study of variational theory of nonlinear development equations, and explored the practical application of nonlinear model in image processing.

WANZHEN JIAO received the Ph.D. degree in ophthalmology from Shandong University, Jinan, in 2014. She is currently an Ophthalmologist with Shandong Provincial Hospital Affiliated to Shandong University. Her research interests include laser treatment of diabetic retinopathy, age-related macular degeneration, retinal vascular occlusion, and choroidal neovascular disease.



BOJUN ZHAO is currently a Chief Physician of ophthalmology with Shandong Provincial Hospital Affiliated to Shandong University, also the Head of the Fundus Group, Chinese Eye Microcirculation Society, and also a Doctoral Tutor with Shandong University. He has worked in many medical and research institutions in U.K. His research interests include the diagnosis and treatment of fundus diseases and optic nerve diseases.

YANHUI DING is currently an Associate Professor with the School of Information Science and Technology, Shandong Normal University. His research interests include medical image analysis and data analysis.