Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Non-rigid retinal image registration using an unsupervised structure-driven regression network

Beiji Zou [a,b,c], Zhiyou He [a], Rongchang Zhao [a,b,*], Chengzhang Zhu [b], Wangmin Liao [a], Shuo Li [d]

[a] School of Computer Science and Engineering, Central South University, Changsha, Hunan Province 410083, China
[b] Hunan Engineering Research Center of Machine Vision and Intelligent Medicine, Changsha, Hunan Province 410083, China
[c] Mobile Health Ministry of Education-China Mobile Joint Laboratory, Changsha, Hunan Province 410083, China
[d] Department of Medical Imaging, Western University, London, ON N6A 4V2, Canada

ARTICLE INFO

ABSTRACT

Retinal image registration is clinically significant to help clinicians obtain more complete details of the retinal structure by correlating the properties of the retina. However, existing methods suffer from great challenges due to time-consuming optimization and lack of ground truth. In this paper, we propose an unsupervised learning framework for non-rigid retinal image registration, which directly learns the mapping from a retinal image pair to their corresponding deformation field without any supervision such as ground truth registration fields. Specifically, we formulate the complex mapping as a parameterized deformation function, which can be represented and optimized by a deep neural network. Furthermore, the Structure-Driven Regression Network (SDRN) framework is applied to compute the multi-scale similarity combined with contextual structures (e.g., vessel distribution, optic disk appearance, and edge information) to guide the end-to-end learning procedure more effectively with unlabeled data. Given a new pair of images, our method can quickly register images by directly evaluating the parametric function using the learned parameters, which runs faster than traditional registration algorithms. Experimental results, performed on the public challenging dataset (FIRE), show that our method achieves an average Dice similarity coefficient (DSC) of 0.753 with short execution times (0.021 s), which is more accurate and robust than existing approaches and promises to significantly speed up retinal image analysis and processing.

1. Introduction

Retinal image registration is a clinically significant fundamental task in medical image analysis since it provides more complete details to assess and evaluate the development of eye-related diseases in clinical practice. By correlating and comparing details in retinal images taken at different time periods, clinicians can evaluate the progress of diseases and decide on the appropriate treatments to be taken. In the clinical procedure, retinal image registration contributes to diagnose, monitor and track many ocular pathologies including age-related macular degeneration (ARMD), diabetic retinopathy (DR), glaucoma and vasculitis. It can establish the anatomical correspondences between a pair of images, and thus ensures image data comparability to facilitate the subsequent analysis such as longitudinal studies [1]. Therefore, automated retinal image registration is required by clinical practice and has been a topic of active research [2–4]. Several image registration toolkits have been developed such as ITK [5], Elastix [6] and so on.

Although many studies are devoted to retinal image registration, nowadays, it is still a challenging task due to time-consuming optimization and lack of ground truth which obstructs clinical practice. Existing deformable registration methods [7–9] for retinal images suffer from the following aspects: (1) vascular structures of the retinal image are very complicated due to intensity variations and changed structures, which makes it difficult to align vasculature; (2) traditional deformable registration methods iteratively optimize the cost function, which limits the speed of registration; (3) collecting large-scale accurate pixel-level annotation for registration is time-consuming and challenging. All of these factors act as obstacles to the application in clinical practice.

In this paper, we present an unsupervised retinal image registration method that learns the non-linear spatial correspondence without any supervised information. We mainly focus on content

changes (e.g., local deformations of blood vessels) in image registration which are often non-rigid transformation. The proposed method formulates retinal image registration as a parameterized deformation function, which directly models the complex mapping from an image pair to their corresponding deformation field. The parameters of the deformation function are learned by a regression model, named as Structure-Driven Regression Network (SDRN). Specifically, the parameters of our SDRN (i.e., the convolutional kernel weights) are optimized by adopting the multi-scale deformation fields and structure-driven strategies, which participate in the computation of the loss function. By training a set of image pairs from the dataset of interest, the procedure learns a common representation that can align any new pair of images from the same distribution. Registration between a new test image pair is achieved by simply evaluating the learned function, resulting in faster registration speed than traditional registration methods. Throughout this paper, we use the example of registering retinal images. However, our method is broadly applicable to other medical image registration. Results show that the proposed method performs accurate retinal image registration with short execution times, which is more robust and accurate than existing excellent approaches. Images can now be registered under a second with a GPU. The contributions of our work are threefold:

- an unsupervised learning framework is proposed to directly regress the non-linear spatial correspondence between the retinal image pair without any supervision during training;
- a novel structure-driven regression network, trained with multi-scale deformation fields strategy, is proposed to facilitate robust and accurate learning;
- the proposed method achieves the best registration performance compared with state-of-the-art algorithms.

2. Related works

Comprehensive overviews of registration methods and the most recent advances in this domain can be found in [10–12]. In the following two subsections, the most important classes of existing registration method that we categorize are presented: (1) traditional registration methods and (2) learning-based registration methods. More recently, learning-based registration methods have been widely investigated to improve registration performance. Therefore, they become prevalent.

2.1. Traditional registration methods

The optimization based registration methods can be classified into two categories [13]: intensity-based [14,15] and feature-based [16–18]. These registration algorithms are often based on linear transformations like rigid and affine which intend to globally align the two images. They are designed to deal with changes of content appearance (e.g., due to different sensors imaging the same organs) whereas the deformable image registration (DIR) is used to generate the local deformations between two images. Recently, DIR is widely used in many medical image analysis and applications to deal with content changes. It has great potential to establish non-linear spatial correspondences between retinal images due to its efficiency to find out a non-rigid transformation. Several works are devoted to develop effective DIR algorithms for retinal image registration [7–9].

Most conventional deformable registration methods regard the registration process as a high-dimensional optimization problem. They iteratively optimize a transformation based on a cost function:

$$\widehat{\phi} = \arg\min_{\phi} L_{sim}\left(I_f, I_m(\phi)\right) + \lambda L_{smooth}(\phi) \qquad (1)$$

where I_f, I_m denote the fixed and moving images, respectively, the deformation field ϕ can be obtained by minimizing function L_{sim} measuring similarity between I_f and warped moving image $I_m(\phi)$, with regularization L_{smooth} imposes on ϕ, and λ is the regularization parameter.

There are common similarity metrics and regularization terms have been proposed to solve the optimization problem in Eq. (1). Often, ϕ is a displacement vector field. Commonly adopted similarity metrics include the mean squared difference (MSD) [19], mutual information (MI) [20], and normalized cross correlation (NCC) [1,21,22]. A regularization term is often required to minimize the bending energy and smooth deformation [23,24]. For each pair of images, traditional registration methods iteratively optimize the cost function, which neglects the inherent registration patterns shared across images from the same distribution. Therefore, these algorithms often require task-sensitive parameter tuning and time-consuming iterative optimization which is computationally intensive.

2.2. Learning-based registration methods

For learning-based registration methods, the correlation between the deformation field and images can be learned by machine learning techniques [25–27]. Kim et al. [25] train a deformation–appearance model through support vector regression (SVR) to bridge the intrinsic statistics of deformation fields and image appearances. Gutierrez et al. [27] propose to estimate deformation parameters via a supervised regression model using gradient-based optimization, instead of directly minimizing registration energy. These methods have demonstrated improved registration performances. But it cannot be ignored that this type of registration method is data-driven typically, which often requires large-scale pixel-level labels. However, large-scale pixel-level labels are difficult to collect.

Recently, deep neural network based methods are considered as novel solutions to medical image registration. Although the spatial transformer network (STN) proposed by Jaderberg et al. [28] is designed as a part classification task, whereas image alignment is not guaranteed, it is one of the first approaches that exploit CNN for image registration. STN gives neural networks the ability to actively spatially transform feature maps. STN is a differentiable module that can be inserted into existing CNNs architectures, making it possible to be applied to image registration. Recent papers [22,29–31] employ it to warp one image to another, enabling end-to-end training. Most deep learning-based methods [32–35] show accurate registration performance, but they are supervised. Miao et al. [32] use convolutional neural network regressors to directly predict the transformation parameters. Liao et al. [33] use an artificial agent and deep convolutional neural networks for iterative registration. Cao et al. [34] use a similarity-steered CNN model to predict the deformation fields even with small datasets. Mahapatra et al. [35] use generative adversarial networks (GANs) for deformable registration of retinal and cardiac MR images. Training examples are generated by synthetic data [32] or obtained by conventional registration methods [33–35], which is problem specific. Due to the limitations of supervised methods, a number of weakly-supervised methods consider extra information such as segmentations or weak labels during training [30,36]. Hu et al. [30] present a weakly-supervised method that learns to align segment labels. The authors demonstrate that a loss function based solely on segmentation labels can effectively train a cross-modality registration network. Lee et al. [36] introduce a generic framework, Image-and-Spatial Transformer Networks (ISTNs), aiming to extract and retain information about Structures-of-Interest (SoI) for structure-guided image registration. However, high-quality segmentation labels are quite difficult

to obtain. Alternatively, unsupervised deep learning methods have been applied to image registration [37,29,38–40]. Wu et al. [37] use a convolutional stacked auto-encoder (CAE) to learn the translational invariant feature representations between fixed and moving images, and use it in a conventional deformable registration framework. However, The CAE is separate from image registration tasks. So it does not necessarily extract the most descriptive features for image registration. Vos et al. [29] propose a deformable image registration network (DIRNet) which learns the mapping of input image pairs and displacement vector field by an unsupervised convolutional neural network. DIRNet has the advantages of non-iteratively outputting transformed image and training completely unsupervised. Unfortunately, directly optimizing a similarity metric between input image pairs cannot guarantee robustness and accuracy. Specifically, differences of image pairs may appear in the form of increased vessel tortuosity, microaneurysms, cotton-wool and spots. Furthermore, DIRNet has no regularization terms, so it cannot penalize undesired deformations. To address this problem, some registration methods [22,30,40] apply smoothing on deformation fields. Few recent approaches build unsupervised models to learn diffeomorphic representations [39–41]. For example, Dalca et al. [39] propose an unsupervised probabilistic generative model for diffeomorphic image registration that can offer uncertainty estimates. The proposed probabilistic framework can extend to anatomical surface alignment, which enables training the network using segmentation labels. To achieve accurate registration results, multi-scale approaches are employed. Vos et al. [38] present the Deep Learning Image Registration (DLIR) framework for unsupervised training of multi-stage ConvNets for hierarchical multi-resolution and multi-scale image registration, each ConvNet with a different B-spline grid spacing and images of different resolution as inputs, and Krebs et al. [40] propose an unsupervised multi-scale deformable registration approach, providing multi-scale estimations of velocities, deformation fields and warped images, leads to a more controlled training procedure compared to a single-scale approach.

3. Methodology

Taking a pair of fixed image I_f and moving image I_m as input, the proposed registration method directly models the mapping M from an image pair to the dense deformation field (DDF) ϕ. The DDF encodes displacement vectors between spatial coordinates of I_f and their counterparts in I_m. Specifically, our method is designed to formulate the complex mapping as a parameterized deformation function $M(I_f, I_m; \theta) = \phi$. Then we model the deformation function M using deep regression network (DRN), where the parameters θ are the learnable parameters of M. Therefore, the image registration is transformed as a problem to find the optimized parameters of the deformation function as

$$\widehat{\theta} = \arg\min_{\theta} L(V_f, V_m, M(I_f, I_m; \theta)) \qquad (2)$$

where $\widehat{\theta}$ is the learned parameters optimized by stochastic gradient descent. In order to bridge the large intensity variations and changed structure, we introduce the auxiliary images to facilitate robust local alignment learning in a structure-driven way, V_f and V_m in Eq. (2), which are vessel enhancement images generated by the enhancement algorithm [42] from the input image pair.

Our learning framework consists of three distinguishing parts: (1) a deep regression network for multi-scale DDF generation; (2) a spatial transformation function for warping images with the generated deformation fields; (3) a structure-driven loss function for optimizing the parameters of SDRN. As shown in Fig. 1, the learning framework takes I_f and I_m as input and estimates ϕ. Then we

warp V_m to $V_m(\phi)$ using a spatial transformation function based on [28], which contributes to evaluating the similarity of V_f and $V_m(\phi)$. Note that the similarity adopted by our method is NCC that is a widely-used similarity measure in a large number of registration methods [1,21,22]. Finally, the optimal parameters are founded by minimizing Eq. (2), similar to Eq. (1). During the testing phase, the deformation field and warped image of an unseen image pair can be obtained by directly evaluating function M. It is recommended to use a high dimensional scale deformation field to generate the warped image.

3.1. Deep regression network

DRN is a non-linear regressor to directly estimate the corresponding displacement vectors in DDF from input image pairs. We adopt a deep convolutional neural network to model the complex mapping of the deformation function, then train the network to optimize the parameters of the deformation function. Given a new pair of retinal images, DRN directly predicts the corresponding deformation field using the learned parameters in the test stage.

The parametrized deformation function M is based on a DRN inspired by VGG16 [43]. As shown in Fig. 2, the network takes the concatenation of the fixed image I_f and moving image I_m as input and aims to generate deformation fields ϕ. In our experiments, the input size is $256 \times 256 \times 2$. We apply a stack of convolutional layers with a small receptive field of 3×3 and stride 1. The stack of convolutional layers captures hierarchical features of the input image pairs which are necessary and close related to the corresponding deformation fields. Then ReLU activations and spatial pooling are followed after each convolutional layer. ReLU activations can enhance the nonlinearity and modeling capability. In pooling layers, we adopt max-pooling to obtain translation-invariant features and increase reception fields, as well as to reduce the number of parameters. Max-pooling is performed over a 2×2 pixel window, with stride 2 specifically. Finally, the feature maps, outputs of the last three pooling layers, execute 3×3 same convolutions with stride 1 to generate multi-scale deformation fields in registration layers. The number of kernels of these convolutions is determined by the dimensionality of the input images. For example, 2D images require 2 kernels for 2D displacement.

In our model, we propose a multi-scale deformation fields strategy to learn the robust and accurate deformation function at different scales. These deformation fields obtained from the DRN encode spatially varying deformations at three scale levels, and they enable the spatial function introduced in the next sections to generate warped images for computing image similarity from three branches in Eq. (9). This image registration strategy makes DRN have the capability of deep supervision without ground truth deformation fields or landmarks. Therefore, our method can perform complex registration tasks in a scale-adaptive manner, further improving the precision of registration results working on multiple levels of detail, which is demonstrated in Table 1. It is worth noting that the process of image registration is still unsupervised. Specifically, the multi-scale deformation fields strategy is only used for driving training. In our experiments, we set up three-scale deformation fields with sizes of $8 \times 8 \times 2, 16 \times 16 \times 2$ and $32 \times 32 \times 2$, respectively. The size of the deformation field represents the spacing of the control point grid, as shown in Fig. 2.

3.2. Spatial transformation function

The spatial transformation function consists of a grid generator and a sampler. It is constructed to compute V_{trans} by warping V_m since a similarity metric between V_{trans} and V_f is required when

Structure-Driven Regression Network
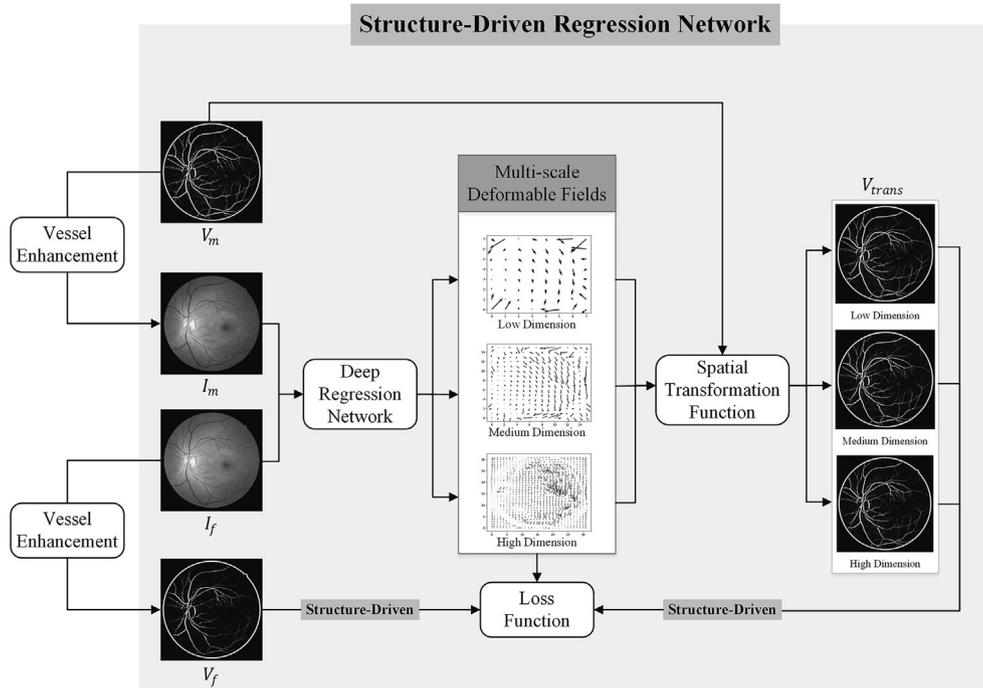
Fig. 1. Overview of the learning framework. The DRN takes a pair of fixed and moving images as input, and directly regresses the corresponding deformation fields at multiple scales. The spatial transformation function computes multiple warped images that enable the model from different scales to evaluate the image similarity. The parameters of the SDRN are optimized by a well-designed loss function taking into account the global similarity and local structure context.
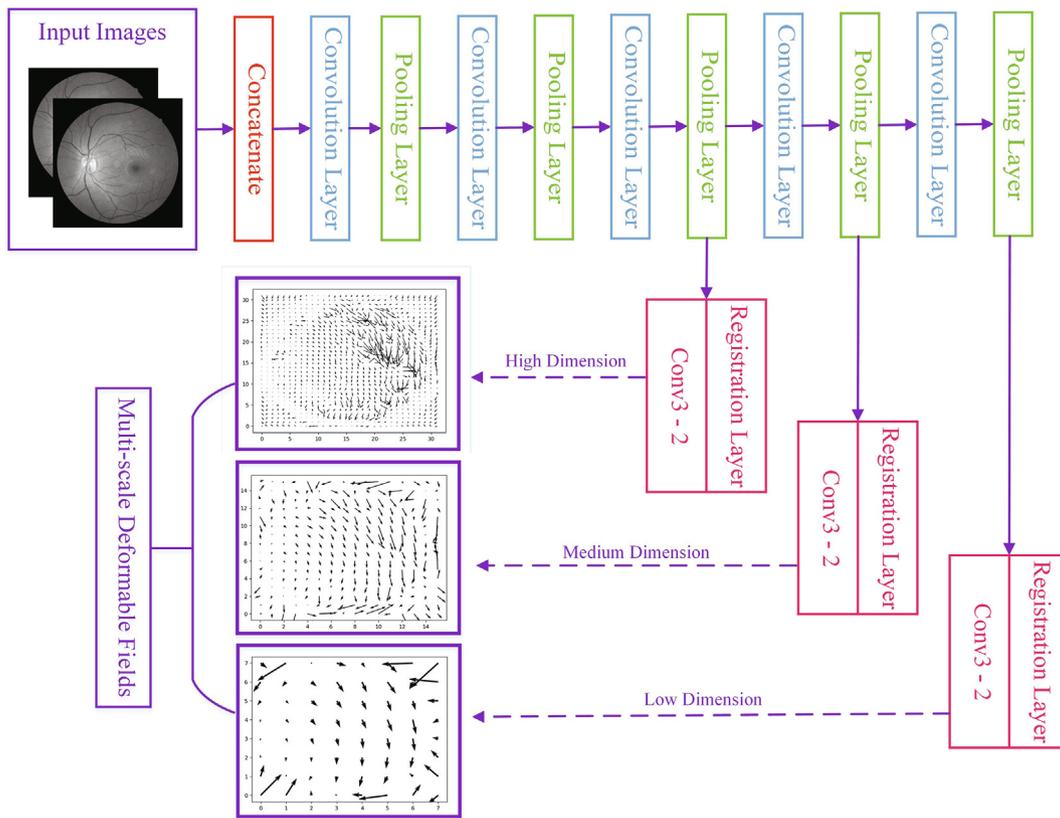
Fig. 2. Architecture details of DRN. The network analyzes pairs of fixed and moving images in concatenating pipelines and outputs a deformation field for each scale. The parameters of registration layers are denoted as "Conv⟨receptive field size⟩-⟨ number of filters⟩.".

Table 1
Registration results using different settings of deformation fields. In our method, the setting of three scales is adopted.

Settings	DSC	PA
Single scale	0.724	0.719
Two scales	0.752	0.748
Three scales	0.753	0.750

the network exploits standard gradient-based methods to learning the optimal parameters of DRN. Details can be seen in Fig. 3.

The deformation field estimated by DRN is used to transform the regular spatial grid G into a sampling grid $D_\theta(G)$. And for that, the deformation field D_θ should be interpolated to the same size as the input feature map (i.e., the same size as the fixed or moving image) via bicubic interpolation for pixel level calculation. The pointwise transformation is

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = D_\theta(G_i) = \begin{pmatrix} \theta_i^1 \\ \theta_i^2 \end{pmatrix} + \begin{pmatrix} x_i^r \\ y_i^r \end{pmatrix} \quad (3)$$

where (x_i^r, y_i^r) are the coordinates of G, (x_i^s, y_i^s) are the coordinates of $D_\theta(G)$, and $(\theta_i^1, \theta_i^2)^T$ is the displacement vector in D_θ.

Then, the enhanced moving image $V_m \in \mathbb{R}^{H \times W \times C}$ with height H, width W, and C channels and $D_\theta(G)$ are taken as inputs to the sampler to produce V_{trans} at the grid points. Bilinear interpolation is adopted during the sampling. Each (x_i^s, y_i^s) coordinates in $D_\theta(G)$ defines the spatial location where a sampling kernel is applied to get the value at a particular pixel in V_{trans}. This can be written as

$$V_{trans}^c(i) = \sum_x^H \sum_y^W V_m^c(x, y) k(x_i^s, y_i^s, x, y; \delta) \quad (4)$$

we use a bilinear sampling kernel k in our method, giving

$$k = \max(0, 1 - |x_i^s - x|) \max(0, 1 - |y_i^s - y|) \quad (5)$$

where i is the location of a pixel in V_{trans}, and c is the channel. Note that every channel is transformed in an identical way. Because the operations are differentiable, the DRN is able to be trained end-to-end.

3.3. Loss function

In order to model the complex mapping from image pairs to corresponding DDFs more robustly and accurately, we introduce a structure-driven strategy to assist DRN to keep aware of structure matching. This strategy uses the enhanced input image pair,

which encodes definitive local edges and context of the retinal structure, to build our novel loss function, further driving robust DRN training. Whereas, each pair of fixed and moving images are used only as input to DRN to learn informative image feature representations and deformation field between images without directly contributing to the loss function. The training process is unsupervised because it does not rely on external registration labels. This way, accurate and robust image alignment is ensured.

Instead, we assume that the loss function directly computed based on the similarity measures between the fixed and warped moving images. In fact, several CNN-based deformable registration methods do likewise recently [22,29]. The registration results will not perform well due to illumination, grayscale and texture changes. As a result, the deformation fields learned are often inaccurate, and the registered images are distorted. This assumption is confirmed in Section 5.2. To summarize, the structure-driven strategy provides local information to help structural alignment. As the optimization of the loss function proceeds, the parameters of DRN will be updated effectively. In this way, the registration results are more accurate and robust.

The loss function consists of two parts. The one part denoted as L_{sim} penalizes differences between the enhanced fixed image and the warped image generated by applying a deformation field to the enhanced moving image. Another part denoted as L_{smooth} is a regularization term that is calculated by enforcing smoothness constraints on the deformation field.

Let ϕ_l, ϕ_m, and ϕ_h denote low, medium, and high scale deformations fields, respectively. We set L_{sim} to the negative NCC for back-propagating dissimilarity. NCC between V_f and warped V_m is formulated as

$$NCC(\phi, V_f, V_m) = \frac{\sum_{x \in \Omega_{V_f}} (V_f(x) - \overline{V_f})(V_m(\phi \circ x) - \overline{V_m})}{\sqrt{\sum_{x \in \Omega_{V_f}} (V_f(x) - \overline{V_f})^2 \sum_{x \in \Omega_{V_f}} (V_m(\phi \circ x) - \overline{V_m})^2}} \quad (6)$$

$\overline{V_f}$ and $\overline{V_m}$ are defined as Eq. (7) and Eq. (8)

$$\overline{V_f} = \frac{1}{|\Omega_{V_f}|} \sum_{x \in \Omega_{V_f}} V_f(x) \quad (7)$$

$$\overline{V_m} = \frac{1}{|\Omega_{V_f}|} \sum_{x \in \Omega_{V_f}} V_m(\phi \circ x) \quad (8)$$

where ϕ is a deformation field, x is spatial coordinates of a pixel in V_f, Ω_{V_f} is the domain of V_f, $|\Omega_{V_f}|$ is the number of pixels, and $\phi \circ x$ is
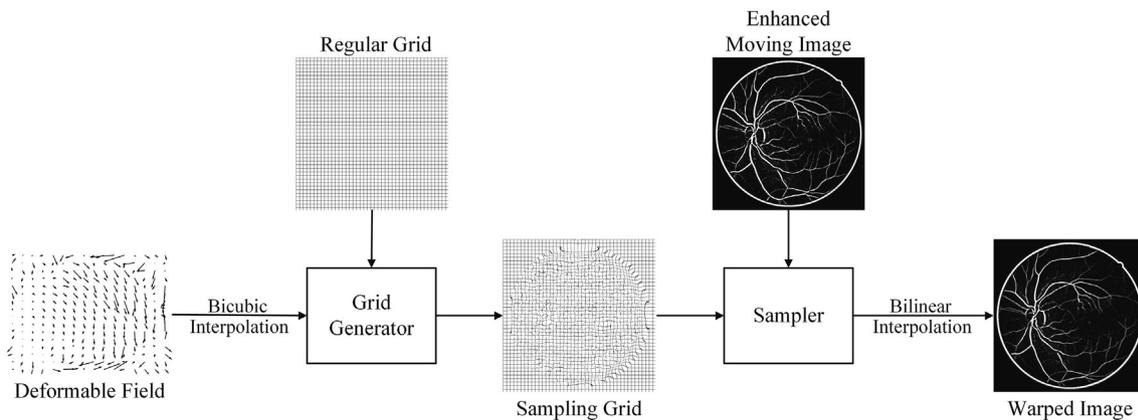
Fig. 3. The working principle of the spatial transformation function.

deformed spatial coordinates of a pixel by ϕ in V_m. Through the above analysis, we can easily get the formula of L_{sim}. It is given by

$$L_{sim} = -\alpha_l NCC(\phi_l, V_f, V_m) - \alpha_m NCC(\phi_m, V_f, V_m)$$
$$- \alpha_h NCC(\phi_h, V_f, V_m) \qquad (9)$$

where α_l, α_m, and α_h are the parameters that control the weights at different scales.

A regularization term L_{smooth} is needed to encourage smooth deformation fields. The deformation fields are regularized by square L2-norm of their spatial gradients. As shown in Eq. (10).

$$L_{smooth} = \beta_l L_{grad}(\phi_l) + \beta_m L_{grad}(\phi_m) + \beta_h L_{grad}(\phi_h) \qquad (10)$$

with

$$L_{grad}(\phi) = \sum_{x \in \Omega} \|\nabla \phi(x)\|_2^2 \qquad (11)$$

where ϕ is a deformation field, x is spatial coordinates of a pixel of ϕ, and Ω is the domain of ϕ. β_l, β_m, and β_h are regularization parameters.

To summarize, the total loss function is $L = L_{sim} + L_{smooth}$.

4. Experimental setup

We evaluate the effectiveness of SDRN on the task of fundus image registration. Detailed data description and experimental set up are provided as follows.

4.1. Dataset

We perform experiments on the publicly available Fundus Image Registration (FIRE) dataset [44] to validate the effectiveness of our method. The dataset consists of 129 retinal images forming 134 image pairs with a resolution of 2912×2912 pixels. These image pairs are classified into three different categories according to their characteristics. Category P contains 49 image pairs with such small overlaps that the data cannot be guaranteed to be affinely aligned. Therefore, this category is not suitable for deformable registration to evaluate our proposed algorithm. In our experiment, we choose category S with 71 image pairs and category A with 14 image pairs as our experimental dataset. It contains 85 image pairs and is divided into 59 training image pairs and 26 test image pairs. We augment the training images to a total of 10030 image pairs by rigid, affine and elastic transformations. Note that the fundus images are resampled to 256×256 and converted to grayscale. We exploit the vessel enhancement method [42] to enhance vascular structures using 2D multiscale enhancement filters. We explain that we do not use supervised information such as ground truth deformation fields or landmarks.

4.2. Evaluation metrics

It is not well-defined to obtain ground truth deformation fields because many deformation fields can generate similar looking deformed images. Hence, we evaluate SDRN using vascular structures segmentations. We extract the retinal blood vessels based on line operators [45]. As in the previous method, we use the following evaluation metrics for our experimental dataset to evaluate the performance of our method.

Pixel Accuracy (PA) Pixel accuracy is a simple metric that measures the ratio of the same pixels in both segmentation masks. Here, we adopt PA to measure the performance of registration algorithms by judging the alignment of the segmented retinal vessels since it will gain a higher value when the retinal vessels are well aligned. PA is equal to the number of identical pixels in the

two anatomical segmentation images divided by the total number of pixels.

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \qquad (12)$$

where p_{ii} represents the number of true positives, whereas p_{ij} and p_{ji} are usually interpreted as false positives and false negatives respectively.

Dice Similarity Coefficient (DSC) Dice similarity coefficient [46] measures the overlap between vasculature segmentations. If a registration field represents accurate anatomical correspondences, we expect the regions in the fixed image and deformed moving image corresponding to the same anatomical structure. The DSC of the two segmentation masks is defined as

$$DSC(S_f, S_{trans}) = 2 \times \frac{|S_f \cap S_{trans}|}{|S_f| + |S_{trans}|} \qquad (13)$$

where S_f is the vessel segmentation of the fixed image and S_{trans} is the vessel segmentation of the deformed moving image. The higher DSC value indicates a better registration result.

Root Mean Square Error (RMSE) For landmarks, the registration error is determined as RMSE between points in the deformed image and its corresponding points in the fixed image.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \left[(x_i - x_i')^2 + (y_i - y_i')^2 \right]}{N}} \qquad (14)$$

where N is the number of corresponding points, (x_i, y_i) is a point of the fixed image and (x_i', y_i') on the deformed image.

4.3. Implementation details

Our network is implemented with Tensorflow [47]. We choose the Adam [48] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We set a large learning rate to 10^{-1} for the initial training stage to acquire a high convergence speed with a batch size of 5 pairs of retinal images. The parameters introduced in Section 3.3 are set to $\alpha_l = \beta_l = 0.3, \alpha_m = \beta_m = 0.6$ and $\alpha_h = \beta_h = 0.9$. The registration performance is not very sensitive to parameter tuning. Our SDRN and DIRNet are conducted on a Nvidia GeForce GTX 1080 Ti GPU. We select conventional iterative registrations (SimpleITK [49] and SimpleElastix [50]) for comparison. Experiments are implemented in Python and performed with an Intel Xeon E5-2683 v3 2.00 GHz CPU.

5. Results

We conduct several ablation studies to investigate the effectiveness of our SDRN. In addition, several state-of-the-art registration tools and frameworks are adopted as the baselines of our work to compare with each other. Ablation studies and comparison experiments indicate that our SDRN can achieve an accurate result with short execution times.

5.1. Evaluation for the contribution of the multi-scale deformation fields strategy

In this section, we investigate the influence of the multi-scale strategy on retinal image registration. For this purpose, we set up an ablation experiment that we fix this model and evaluate the

registration performance by using deformation fields with three different settings. As shown in Table 1, without using this strategy, the registration performance drops on average. Notably, the best performance is achieved using our proposed way three-scale deformation field. DSC and PA increase by approximately 4% in a three-scale way compared to the single-scale way. Registration speeds for all these settings are approximately 0.021 s. Furthermore, we can observe these changes in the registration performance of each image pair in Fig. 4. Over half of the test data achieves the highest values. The result demonstrates that this strategy can optimize the network using deep supervision, thus further improving the performance via working on multiple levels of detail. The reason can be explained by the fact that the size of the deformation field represents the spacing of the control point grid, as we mentioned before. Thus, the lower dimension of the deformation field is able to capture the larger deformations between the image pairs globally. In contrast, the higher dimension of the deformation field can restore smaller deformations locally.

5.2. Evaluation for the contribution of the structure-driven strategy

To validate the effectiveness of the structure-driven strategy, we set up an ablation experiment that uses the input images, a pair of fixed and moving images, to directly compute the loss function instead of using the auxiliary images, which has been discussed before in Section 3.3. The effect can be visually appreciated in Fig. 5. As red arrows point out, Fig. 5(c) is distorted in the area where anatomical differences and grayscale change exist in the test image pairs. Whereas Fig. 5(d) still maintains good performance since we use the retinal vasculature structure, which is often considered to be the representative and robust feature, to solve the problem of lack of robustness during the training stage. That is, our method still accurately predicts the deformation field, when there exist differences in the image pairs such as increased vessel tortuosity and grayscale. The comparable evaluations for each test sample are shown in Fig. 6. Evaluations of almost all image pairs

are significantly improved by our proposed method. Obviously, the proposed strategy significantly improves the precision and robustness of registration. This demonstrates that the trained SDRN is accurate and the proposed registration method is well applicable. This strategy is useful for helping structure alignment and solving the problem of lack of robust similarity metrics for images.

5.3. Comparing with state-of-the-art

We compare our approach with the popular registration toolkits using SimpleITK [49] and SimpleElastix [50] with manually tuned parameters. SimpleITK is an easy-to-use interface to the ITK, intended to facilitate its use in rapid prototyping, education, and scientific activities via high-level programming languages like Python, etc. SimpleElastix brings Elastix multiple programming languages allow users to call Elastix. To compare with recent CNN-based registration approaches, we also test DIRNet the first unsupervised end-to-end method [29], VoxelMorph a state-of-the-art fast and accurate registration method [31] and a weakly-supervised method [30]. We apply a small learning rate of 10^{-4} for all learning-based methods while using more training epochs. For [30,31], we sweep the regularization parameters and set them to 0.5.

Table 2 shows the comparative evaluations of the previously discussed algorithms. Our proposed method achieves the highest DSC 0.753 and PA 0.750. At the same time, RMSE is significantly improved using our algorithm. It is worth noting that there are many anatomical differences in our experimental dataset, as shown in Fig. 7. As a result, features that are visible in one image, may be occluded in the other, leading to inaccurate registration results. It indicates that our method not only improves the registration accuracy, but also is robust to anatomical differences that may appear in the form of increased vessel tortuosity, microaneurysms, cotton-wool, spots, etc. We can also observe that, in
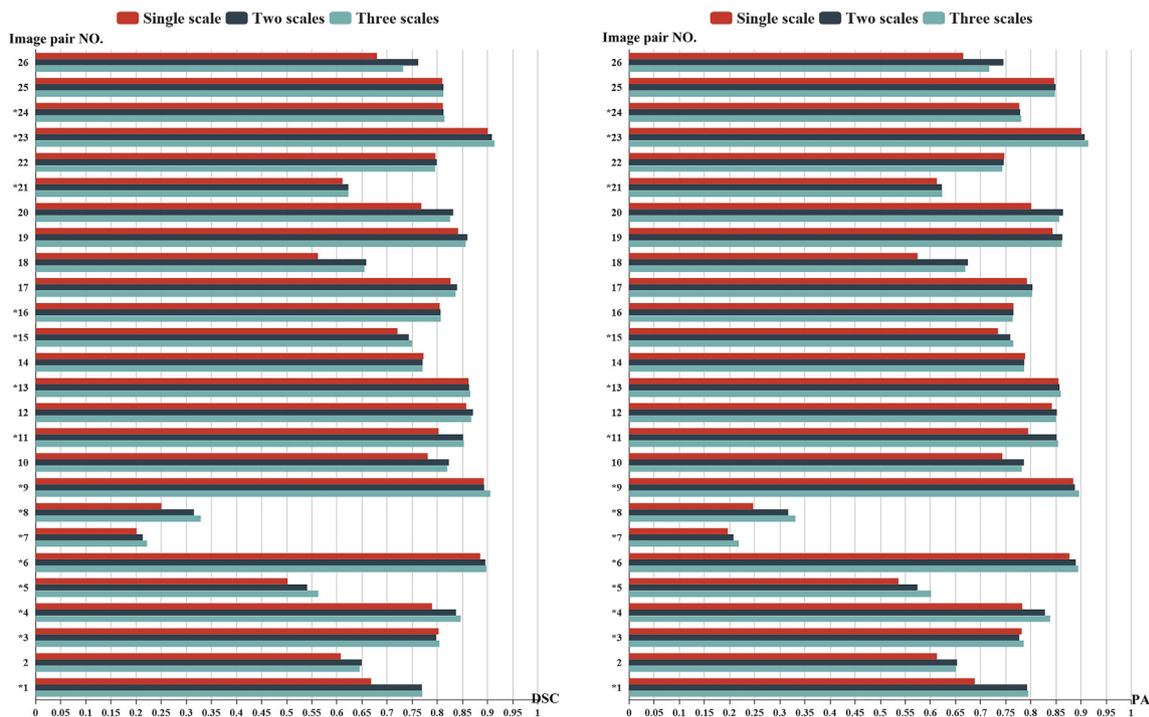
Fig. 4. DSC and PA using different settings of deformation fields. "*" means the cases where the setting of three scales achieves the highest values.
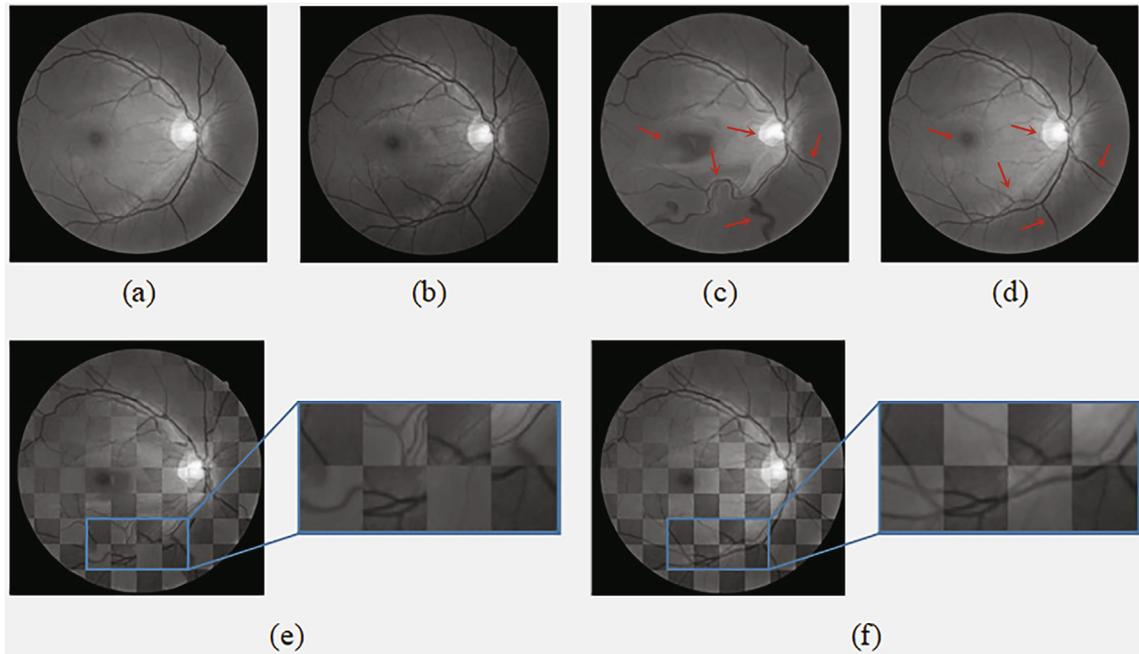
Fig. 5. Example results of the ablation experiment. (a) A fixed image, (b) a moving image. Registered images (i.e. warped images) of (c) SDRN-C and (d) SDRN. (e), (f) respectively denote the registration accuracy test results using SDRN-C and SDRN.
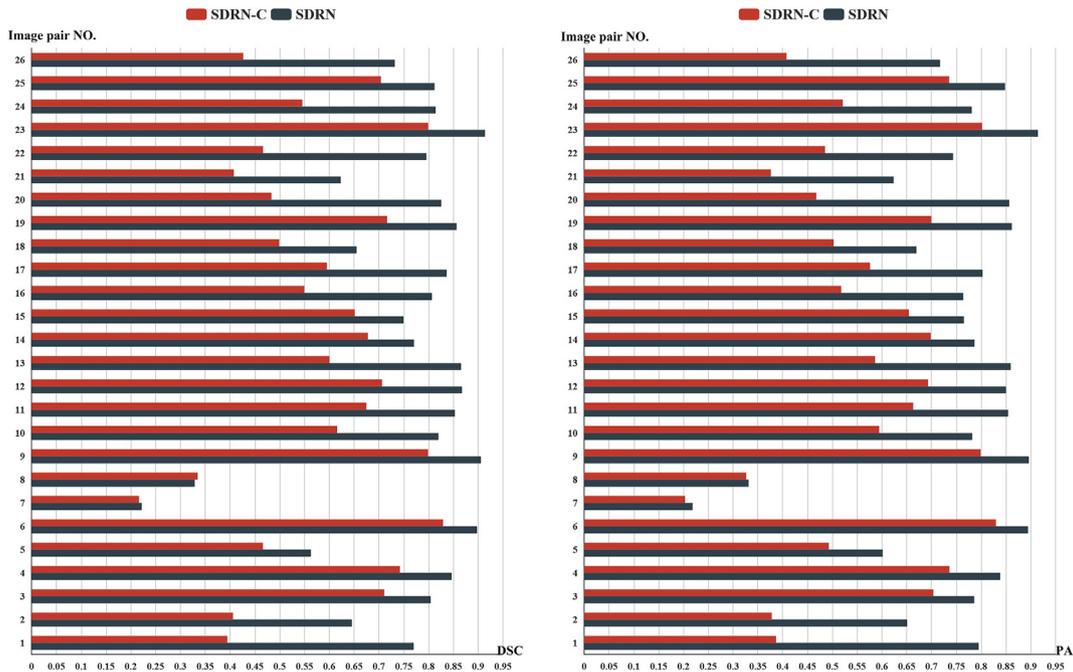
Fig. 6. DSC and PA for each pair of test images from FIRE after performing the deformable registration by SDRN-C (a controlled experiment for SDRN without structure-driven strategy) and SDRN.

the test stage, CNN-based methods can register an unseen image pair within 1 s with a GPU, which indicates our method and other deep learning-based methods effectively solve the problem of the traditional deformation registration methods with time-consuming iterative optimization. CNN-based methods using GPUs have been becoming more common for improving computational efficiency. We visualize deformation fields for each method using

deformed images with a deformation grid in Fig. 8. Compared to the other algorithms, our approach produces smoother and more regular deformations. As discussed before, the deformation field generated by DIRNet has many undesired deformations because DIRNet does not smooth it. Generally, conventional algorithms produce more irregular deformation fields compared to CNN-based algorithms. To illustrate the effectiveness of the proposed

Table 2
Comparative average performance of different methods. Time denotes average run time in seconds taken to register an image pair.

Method	DSC	PA	RMSE	Time(s)
Before registration	0.437	0.438	4.174	\
SimpleITK(Demons) [49]	0.446	0.441	3.867	30.685
SimpleITK(BSplines) [49]	0.526	0.522	2.362	14.451
SimpleElastix [50]	0.599	0.594	2.302	71.005
Vos et al., 2017 [29]	0.575	0.567	2.808	0.006
Balakrishnan et al., 2019 [31]	0.747	0.732	1.540	0.004
Hu et al., 2018 [30]	0.745	0.735	1.355	0.011
Our method	0.753	0.750	0.915	0.021
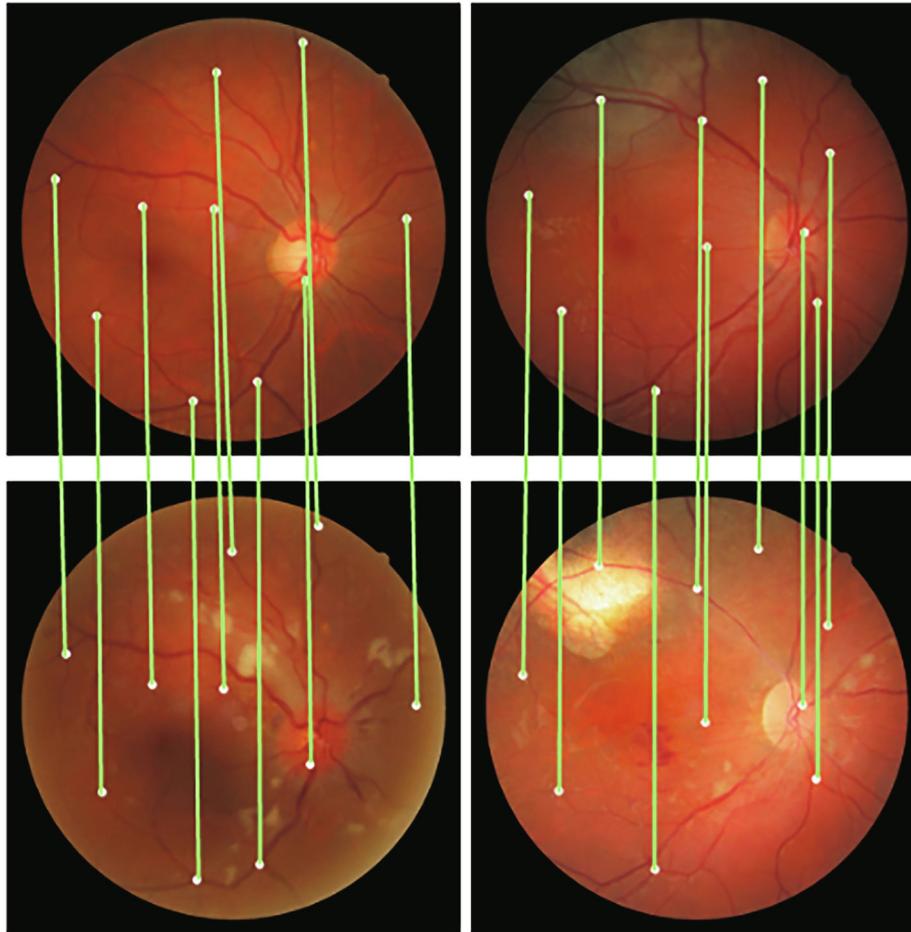
Fig. 7. Examples of input image pairs with corresponding points.
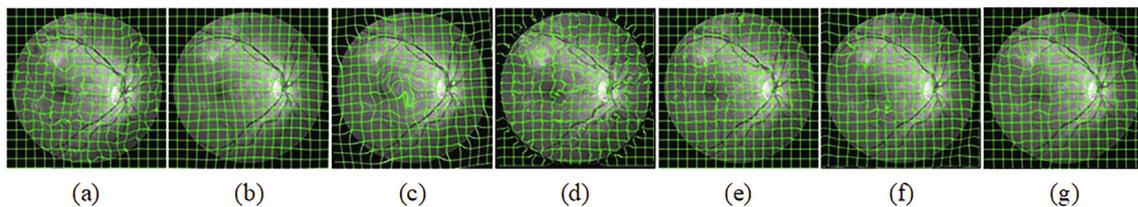
Fig. 8. Examples of deformed images with a deformation grid. (a) SimpleITK(Demons), (b) SimpleITK(BSplines), (c) SimpleElastix, (d) Vos et al. [29], (e) Balakrishnan et al. [31], (f) Hu et al. [30], (g) our method.

method, we also provide visual inspection in Fig. 9. From the overlapped images and combinations of structures using different methods, we can see that our method is able to generate the most realistic deformations among comparison algorithms. In particular,

our method also aligns tiny blood vessels better than [30,31]. In summary, the statistical result demonstrates that the results of our method outperform traditional methods and other CNN-based methods.
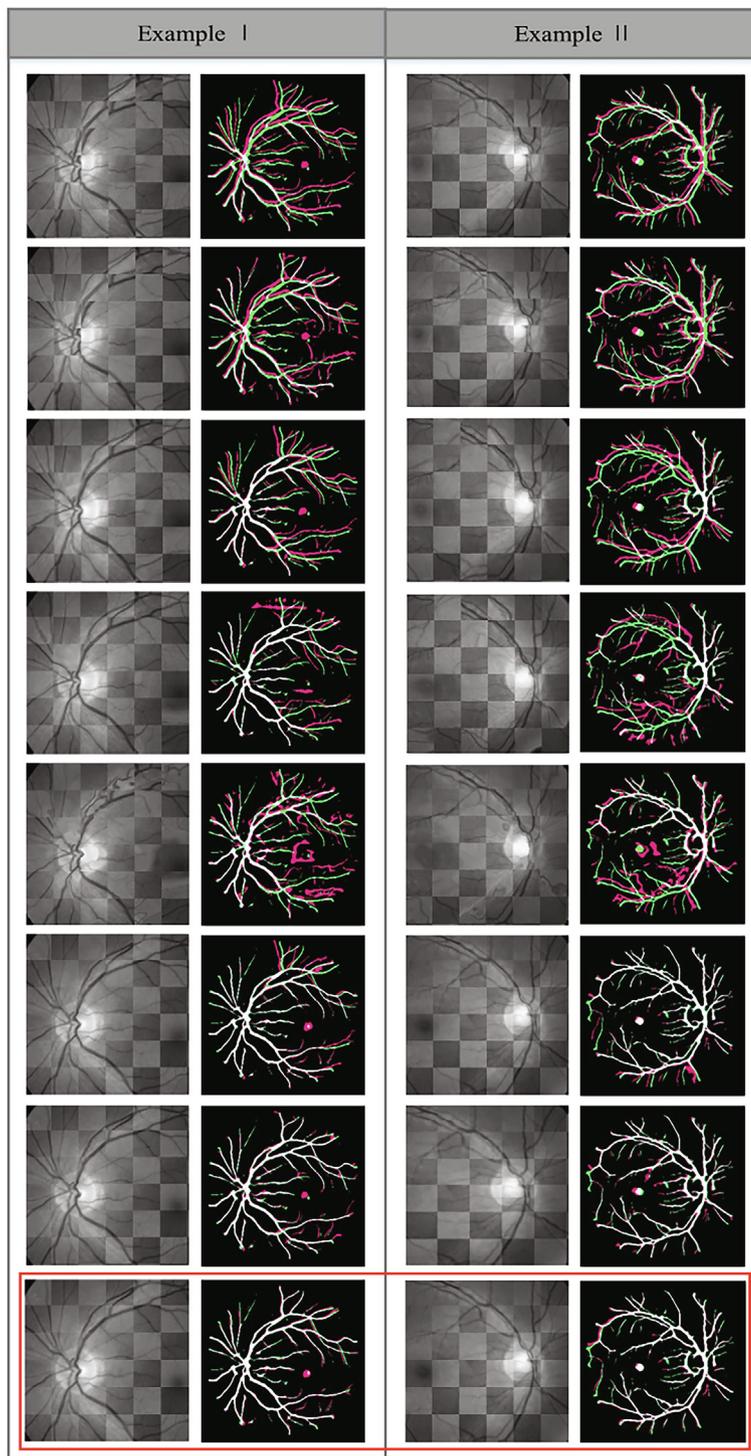
Fig. 9. Examples of registration results for fundus images. Visual comparisons can be illustrated from checkerboards of the same region and combinations of retinal vasculature structure by different methods after registration.

6. Conclusion

In this paper, we have proposed a convolutional neural network algorithm for end-to-end deformable registration in an unsupervised manner. The proposed SDRN directly learns the mapping from an image pair to the corresponding deformation field. This complex mapping is modeled by combining the novel multi-scale deformation fields and structure-driven strategies to effectively guide the training stage, which significantly improves the registration precision and robustness. Given an unseen pair of images, the method can perform registration quickly within one second. Experimental results demonstrate that the proposed method performs better than traditional iterative registration methods and CNN-based methods. Investigating the generalization of the proposed method to 3D images and a wider range of applications would be an important future research direction.

CRediT authorship contribution statement

Beiji Zou: Conceptualization, Resources. **Zhiyou He:** Methodology, Software, Writing - original draft. **Rongchang Zhao:** Conceptualization, Writing - review & editing. **Chengzhang Zhu:** Visualization, Data curation. **Wangmin Liao:** Visualization, Data curation. **Shuo Li:** Writing - review & editing.

References

[1] X. Cao, J. Yang, J. Zhang, Q. Wang, P. Yap, D. Shen, Deformable image registration using a cue-aware deep regression network, IEEE Trans. Biomed. Eng. 65 (9) (2018) 1900–1911, https://doi.org/10.1109/TBME.2018.2822826.

[2] L. Wang, C. Pan, Nonrigid medical image registration with locally linear reconstruction, Neurocomputing 145 (2014) 303–315, https://doi.org/10.1016/j.neucom.2014.05.030, URL: http://www.sciencedirect.com/science/article/pii/S0925231214006511.

[3] M. Chen, A. Lang, H.S. Ying, P.A. Calabresi, J.L. Prince, A. Carass, Analysis of macular OCT images using deformable registration, Biomed. Opt. Express 5 (7) (2014) 2196–2214, https://doi.org/10.1364/BOE.5.002196.

[4] N.S. Alexander, G. Palczewska, P. Stremplewski, M. Wojtkowski, T.S. Kern, K. Palczewski, Image registration and averaging of low laser power two-photon fluorescence images of mouse retina, Biomed. Opt. Express 7 (7) (2016) 2671–2691, https://doi.org/10.1364/BOE.7.002671.

[5] L. Ibanez, W. Schroeder, L. Ng, J. Cates, The ITK Software Guide.http://hdl.handle.net/1926/388..

[6] S. Klein, M. Staring, K. Murphy, M.A. Viergever, J.P. Pluim, Elastix: a toolbox for intensity-based medical image registration, IEEE Trans. Med. Imaging 29 (1) (2010) 196–205, https://doi.org/10.1109/TMI.2009.2035616.

[7] J. Nunes, Y. Bouaoune, E. Delechelle, P. Bunel, A multiscale elastic registration scheme for retinal angiograms, Comput. Vis. Image Underst. 95 (2) (2004) 129–149, https://doi.org/10.1016/j.cviu.2004.03.007.

[8] A. Myronenko, X. Song, Image registration by minimization of residual complexity, IEEE Conference on Computer Vision and Pattern Recognition 2009 (2009) 49–56, https://doi.org/10.1109/CVPR.2009.5206571.

[9] A. Perez-Rovira, E. Trucco, P. Wilson, J. Liu, Deformable registration of retinal fluorescein angiogram sequences using vasculature structures, Annual International Conference of the IEEE Engineering in Medicine and Biology 2010 (2010) 4383–4386, https://doi.org/10.1109/IEMBS.2010.5627094.

[10] A. Sotiras, C. Davatzikos, N. Paragios, Deformable medical image registration: a survey, IEEE Trans. Med. Imaging 32 (7) (2013) 1153–1190, https://doi.org/10.1109/TMI.2013.2265603.

[11] M. Holden, A review of geometric transformations for nonrigid body registration, IEEE Trans. Med. Imaging 27 (1) (2008) 111–128, https://doi.org/10.1109/TMI.2007.904691.

[12] M.A. Viergever, J.A. Maintz, S. Klein, K. Murphy, M. Staring, J.P. Pluim, A survey of medical image registration – under review, Med. Image Anal. 33 (2016) 140–144, https://doi.org/10.1016/j.media.2016.06.030, 20th anniversary of the Medical Image Analysis journal (MedIA),http://www.sciencedirect.com/science/article/pii/S1361841516301074.

[13] F.P. Oliveira, J.M.R. Tavares, Medical image registration: a review, Comput. Methods Biomech. Biomed. Eng. 17 (2) (2014) 73–93, https://doi.org/10.1080/10255842.2012.670855, pMID: 22435355.

[14] N. Ritter, R. Owens, J. Cooper, R.H. Eikelboom, P.P. Van Saarloos, Registration of stereo and temporal images of the retina, IEEE Trans. Med. Imaging 18 (5) (1999) 404–0418, https://doi.org/10.1109/42.774168.

[15] A. Myronenko, X. Song, Intensity-based image registration by minimizing residual complexity, IEEE Trans. Med. Imaging 29 (11) (2010) 1882–1891, https://doi.org/10.1109/TMI.2010.2053043.

[16] M. Vaillant, C. Davatzikos, Hierarchical matching of cortical features for deformable brain image registration, in: A. Kuba, M. Šáamal, A. Todd-Pokropek (Eds.), Information Processing in Medical Imaging, Springer, Berlin Heidelberg, Berlin, Heidelberg, 1999, pp. 182–195.

[17] S. Kahaki, M.J. Nordin, A.H. Ashtari, S.J. Zahra, Deformation invariant image matching based on dissimilarity of spatial features, Neurocomputing 175 (2016) 1009–1018, https://doi.org/10.1016/j.neucom.2015.09.106, URL: http://www.sciencedirect.com/science/article/pii/S0925231215015957.

[18] G. Yang, C.V. Stewart, M. Sofka, C. Tsai, Registration of challenging image pairs: Initialization, estimation, and decision, IEEE Trans. Pattern Anal. Mach. Intell. 29 (11) (2007) 1973–1989, https://doi.org/10.1109/TPAMI.2007.1116.

[19] J. Kybic, M. Unser, Fast parametric elastic image registration, IEEE Trans. Image Process. 12 (11) (2003) 1427–1442, https://doi.org/10.1109/TIP.2003.813139.

[20] J.P.W. Pluim, J.B.A. Maintz, M.A. Viergever, Image registration by maximization of combined mutual information and gradient information, in: S.L. Delp, A.M. DiGoia, B. Jaramaz (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2000, Springer, Berlin Heidelberg, Berlin, Heidelberg, 2000, pp. 452–461.

[21] D.L. Collins, A.C. Evans, Animal: Validation and applications of nonlinear registration-based segmentation, Int. J. Pattern Recognit Artif Intell. 11 (08) (1997) 1271–1294, https://doi.org/10.1142/S0218001497000597.

[22] G. Balakrishnan, A. Zhao, M.R. Sabuncu, J. Guttag, A.V. Dalca, An unsupervised learning model for deformable medical image registration, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9252–9260.

[23] T. Vercauteren, X. Pennec, A. Perchant, N. Ayache, Diffeomorphic demons: efficient non-parametric image registration, NeuroImage 45 (Supplement 1) (2009) S61–S72, https://doi.org/10.1016/j.neuroimage.2008.10.040, mathematics in Brain Imaging.http://www.sciencedirect.com/science/article/pii/S1053811908011683.

[24] M. Staring, S. Klein, J.P.W. Pluim, A rigidity penalty term for nonrigid registration, Med. Phys. 34 (11) (2007) 4098–4108, https://doi.org/10.1118/1.2776236, URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.2776236.

[25] M. Kim, G. Wu, P. Yap, D. Shen, A general fast registration framework by learning deformation-appearance correlation, IEEE Trans. Image Process. 21 (4) (2012) 1823–1833, https://doi.org/10.1109/TIP.2011.2170698.

[26] Q. Wang, M. Kim, Y. Shi, G. Wu, D. Shen, Predict brain MR image registration via sparse learning of appearance and transformation, Med. Image Anal. 20 (1) (2015) 61–75, https://doi.org/10.1016/j.media.2014.10.007.

[27] B. Gutierrez-Becker, D. Mateus, L. Peter, N. Navab, Guiding multimodal registration with learned optimization updates, Med .Image Anal. 41 (2017) 2–17, https://doi.org/10.1016/j.media.2017.05.002.

[28] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks, in: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems 28, Curran Associates Inc, 2015, pp. 2017–2025.http://papers.nips.cc/paper/5854-spatial-transformer-networks.pdf..

[29] B.D. de Vos, F.F. Berendsen, M.A. Viergever, M. Staring, I. Išgum, End-to-end unsupervised deformable image registration with a convolutional neural network, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer International Publishing, Cham, 2017, pp. 204–212.https://doi.org/10.1007/978-3-319-67558-9_24..

[30] Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, E. Bonmati, G. Wang, S. Bandula, C.M. Moore, M. Emberton, S. Ourselin, J.A. Noble, D.C. Barratt, T. Vercauteren, Weakly-supervised convolutional neural networks for multimodal image registration, Med. Image Anal. 49 (2018) 1–13, https://doi.org/10.1016/j.media.2018.07.002.

[31] G. Balakrishnan, A. Zhao, M.R. Sabuncu, J. Guttag, A.V. Dalca, Voxelmorph: a learning framework for deformable medical image registration, IEEE Trans. Med. Imaging 38 (8) (2019) 1788–1800, https://doi.org/10.1109/TMI.2019.2897538.

[32] S. Miao, Z.J. Wang, Y. Zheng, R. Liao, Real-time 2D/3D registration via CNN regression, 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI) (2016) 1430–1434, https://doi.org/10.1109/ISBI.2016.7493536.

[33] R. Liao, S. Miao, P. de Tournemire, S. Grbic, A. Kamen, T. Mansi, D. Comaniciu, An artificial agent for robust image registration., in: AAAI, 2017, pp. 4168–4175..

[34] X. Cao, J. Yang, J. Zhang, D. Nie, M. Kim, Q. Wang, D. Shen, Deformable image registration based on similarity-steered cnn regression, in: M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D.L. Collins, S. Duchesne (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2017, Springer International Publishing, Cham, 2017, pp. 300–308.

[35] D. Mahapatra, B. Antony, S. Sedai, R. Garnavi, Deformable medical image registration using generative adversarial networks, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, pp. 1449–1453.https://doi.org/10.1109/ISBI.2018.8363845..

[36] M.C.H. Lee, O. Oktay, A. Schuh, M. Schaap, B. Glocker, Image-and-spatial transformer networks for structure-guided image registration, in: D. Shen, T. Liu, T.M. Peters, L.H. Staib, C. Essert, S. Zhou, P.-T. Yap, A. Khan (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing, Cham, 2019, pp. 337–345.

[37] G. Wu, M. Kim, Q. Wang, B.C. Munsell, D. Shen, Scalable high-performance image registration framework by unsupervised deep feature representations learning, IEEE Trans. Biomed. Eng. 63 (7) (2016) 1505–1516, https://doi.org/10.1109/TBME.2015.2496253.

[38] B.D. de Vos, F.F. Berendsen, M.A. Viergever, H. Sokooti, M. Staring, I. Išgum, A deep learning framework for unsupervised affine and deformable image registration, Med. Image Anal. 52 (2019) 128–143, https://doi.org/10.1016/j.media.2018.11.010, URL: http://www.sciencedirect.com/science/article/pii/S1361841518300495.

[39] A.V. Dalca, G. Balakrishnan, J. Guttag, M.R. Sabuncu, Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces, Med. Image Anal. 57 (2019) 226–236, https://doi.org/10.1016/j.media.2019.07.006.

[40] J. Krebs, H. Delingette, B. Mailhé, N. Ayache, T. Mansi, Learning a probabilistic model for diffeomorphic registration, IEEE Trans. Med. Imaging 38 (9) (2019) 2165–2176, https://doi.org/10.1109/TMI.2019.2897112.

[41] J. Krebs, T. Mansi, B. Mailhé, N. Ayache, H. Delingette, Unsupervised probabilistic deformation modeling for robust diffeomorphic registration, in: D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J.M.R. Tavares, A. Bradley, J.P. Papa, V. Belagiannis, J.C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, A. Madabhushi (Eds.), Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer International Publishing, Cham, 2018, pp. 101–109.

[42] T. Jerman, F. Pernuš, B. Likar, Ž. Špiclin, Enhancement of vascular structures in 3D and 2D angiographic images, IEEE Trans. Med. Imaging 35 (9) (2016) 2107–2118, https://doi.org/10.1109/TMI.2016.2550102.

[43] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556..

[44] C. Hernandez-Matas, X. Zabulis, A. Triantafyllou, P. Anyfanti, S. Douma, A.A. Argyros, FIRE: fundus image registration dataset, J. Modeling Ophthalmol. 1 (4) (2017) 16–28.

[45] E. Ricci, R. Perfetti, Retinal blood vessel segmentation using line operators and support vector classification, IEEE Trans. Med. Imaging 26 (10) (2007) 1357–1365, https://doi.org/10.1109/TMI.2007.898551.

[46] L.R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (3) 297–302.https://doi.org/10.2307/1932409..

[47] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, TensorFlow: Large-scale machine learning on heterogeneous distributed systems..

[48] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980..

[49] B. Lowekamp, D. Chen, L. Ibanez, D. Blezek, The design of SimpleITK, Front. Neuroinformatics 7 (2013) 45, https://doi.org/10.3389/fninf.2013.00045.

[50] K. Marstal, F. Berendsen, M. Staring, S. Klein, SimpleElastix: a user-friendly, multi-lingual library for medical image registration, IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2016 (2016) 574–582, https://doi.org/10.1109/CVPRW.2016.78.

Beiji Zou received the B.S., M.S., and Ph.D. degrees from Zhejiang University in 1982, Qinghua University in 1984 and Hunan University in 2001 respectively. He is currently a Professor at the School of Computer Science and Engineering at Central South University. His research interests include computer graphics and image processing.

Zhiyou He is currently pursuing the M.E. degree with Central South University, Changsha, China. His research interests include computer vision and medical image analysis.

Rongchang Zhao now is an Associate Professor and Master Supervisor at the School of Computer Science, at Central South University. Before that, he received my Ph.D. and bachelor's degrees from Department of Electronic Information Science and Technology, Lanzhou University, China, in 2006 and 2011, respectively. Her research interests include Medical Image Analysis, Artificial Intelligence, Deep Learning, Computer Vision.

Chengzhang Zhu received (M.E. degree from Huazhong University of Science and Technology in 2006 and) the Ph.D degree from School of Information Science and Engineering, Central South University, Changsha, China, in 2016. Currently, she is a faculty of the College of Literature and Journalism, Central South University. Her research interests include medical image processing, computer vision and pattern recognition.

Wangmin Liao is currently pursuing the M.E. degree with Central South University, Changsha, China. His research interests include computer vision and medical image analysis.

Shuo Li is the associate professor in department of medical imaging, Western University, Canada. He also is the founding scientific director of digital imaging group of London. His research interests include Artificial Intelligence, Big Data, Machine Learning, Medical Image Analysis.