Neurocomputing 521 (2023) 65–78

Contents lists available at ScienceDirect

Neurocomputing

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

Context-driven pyramid registration network for estimating large topology-preserved deformation ☆

Peng Wang [a], Yunqi Yan [b], Lijun Qian [b], Shiteng Suo [b], Jianrong Xu [b], Yi Guo [a,*], Yuanyuan Wang [a]

[a] School of Information Science and Engineering, Fudan University, Shanghai 200433, PR China
[b] Department of Radiology, Shanghai Jiao Tong University School of Medicine Affiliated Renji Hospital, Shanghai 200127, PR China

ARTICLE INFO

Article history:
Received 24 November 2021
Revised 10 August 2022
Accepted 27 November 2022
Available online 5 December 2022

Keywords:
Deformable image registration
Unsupervised registration
Convolutional neural networks
Brain MRI
Liver CT

ABSTRACT

Deep learning-based deformable image registration methods have become attractive alternatives to traditional methods because of their great performance and fast run time. However, it is still challenging for these methods to estimate large topology-preserved deformation, and the contextual information that is important for large deformation is also under-mined. To address these issues, we propose a novel unsupervised context-driven pyramid registration network for estimating large topology-preserved deformation named CPRNet. Specifically, based on the multi-resolution feature pyramids, we first design multi-receptive-field guidance modules, aiming at exploiting the multi-scale spatial correlation between features of two pyramids. Then we devise multi-view context fusion modules to dynamically fuse deep contextual information containing high-level semantic information from different views of feature maps. Further, we develop a residual estimation strategy to estimate the deformation in a coarse-to-fine manner. Moreover, a deformation field regularization module is proposed to address the challenge of balancing the registration performance and topology preservation. The experiments both on liver computed tomography (CT) images and brain magnetic resonance (MR) images demonstrate that our proposed method provides effective and accurate registration on various datasets with a fast run time. Compared with existing learning-based registration methods, our proposed method exceeds the performance in most trials while maintaining desirable topology preservation capability and can potentially fit various image registration tasks.

1. Introduction

Deformable image registration (DIR) is the process of establishing a nonlinear correspondence between a pair of images and is usually used to eliminate the displacement between two related medical images. The registration task generally can be divided into intra-patient registration and inter-patient registration. Intra-patient registration is more correlated with clinical diagnostic tasks, such as image-guided surgery and radiotherapy as well as motion analysis, while inter-patient registration is more correlated with medical image analysis, such as atlas-based segmentation, landmark localization, and shape analysis. Compared with intra-patient registration, there is much large complex deformation in inter-patient registration.

Traditional registration methods, such as Elastix [22] and Ants [2], have been rigorously developed and studied, but require computationally intensive optimization for each image pair, making them impractical for real-time clinical scenarios.

Recently, learning-based registration methods achieve fast registration by building on deep learning developments. However, many of these are supervised approaches relying on accurate ground truth with a vast number of instances [41,34,5]. Some recent works present unsupervised learning-based models but largely may not work very well in registering images with large complex deformation [3,7,24,42]. Therefore, the registration of images with large displacement and complex deformation while keeping the topology is still a challenge. Constructing multi-resolution pyramids is considered as a natural and effective way to estimate the large complex deformation. Traditional multi-resolution-based methods rely on the iterative optimization to find optimal solutions in each resolution. However, the estimation of deformation by neural network is completed at one time, and thus the

☆ This research was supported by the National Natural Science Foundation of China (Grant 61871135, 81627804 and 81830058) and the Science and Technology Commission of Shanghai Municipality (Grant 20DZ1100104).

* Corresponding author.
E-mail addresses: guoyi@fundan.edu.cn (Y. Guo), yywang@fundan.edu.cn (Y. Wang).

https://doi.org/10.1016/j.neucom.2022.11.088
0925-2312/

sufficient contextual information, which indicates the semantic and spatial correlation of tissues and organs between paired images, should be an important factor in estimating optimal deformation, especially for large complex deformation (see Fig. 1).

Based on the above analysis, we propose a novel context-driven pyramid registration network, which targets the challenge of registering images with large displacements and complex deformation while enhancing topology preservation. The main contributions of our work are summarized below:

- We propose a novel unsupervised registration network by incorporating multi-scale and multi-view contextual information into a pyramid framework to estimate large complex deformation in a coarse-to-fine manner.
- We devise a multi-receptive-field guidance module, which aims at exploiting multi-scale spatial correlation between features of two images, working as a guidance for facilitating the estimation of complex deformation.
- We design a multi-view context fusion module, which fuses deep contextual information from different views of feature maps. This module can help the proposed network prevent it from falling into the local optimum by providing high-level semantic information.
- We introduce a deformation field regularization module to address the challenge of balancing the registration performance and topology preservation, which can accelerate the network convergence and more efficiently enhance the topology preservation.

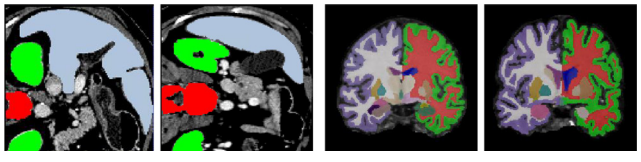
To evaluate the performance of the proposed method, we conduct extensive experiments on two medical image datasets with different modalities and different objects: liver CT images and brain MR images. Qualitative and quantitative evaluation of the experimental results demonstrates our proposed CPRNet can address the challenge of large displacements and complex deformation in DIR while encouraging topology preservation.

This paper is organized as follows. A brief review of related works on learning-based registration methods is introduced in Section 2. In Section 3, we describe our proposed CPRNet in detail. Experiments and results are illustrated in Sections 4 and 5, respectively. In Section 6, the discussion is provided, and conclusions are presented in Section 7.

2. Related work

2.1. Learning-based image registration

Recently, many learning-based registration methods have been proposed. Compared with traditional methods [22,2], they are comparable in performance with less run time. We categorize those learning-based registration methods into supervised methods and unsupervised methods. Supervised learning methods usually require the ground truth of the deformation field for model training. Some learning-based registration methods propose to train the model using deformation fields generated by classical registration methods [33,4,5], but those methods are inevitably affected by the quality of the deformation field, making it difficult for them to make greater progress. Another reasonable way to train supervised registration models is to use random synthetic transformations. These supervised registration methods that rely on random synthetic transformations have been successfully applied to different registration applications, especially including brain MR images [40,41,34] and chest CT images [23,11,12]. Segmentations or landmarks can also be used as ground truths to train supervised registration models [31,46,16,45]. Although these supervised learning methods achieve good registration performance, the generation of ground truth is a difficult issue, especially in large-scale datasets. In addition, the generated ground truth may deviate from the real displacement between two medical images, resulting in poor generalization.



(a) Paired Liver CT Images (b) Paired Brain MR Images

Fig. 1. Examples of 2D paired slices with large complex deformation taken from 3D liver CT images and brain MR images. Different color represents different organs and tissues.

Different from supervised methods, unsupervised methods take the fixed image as the ground truth and train the model by the spatial similarity between the fixed image and registered image, which saves the trouble of acquiring the ground truth. The goal of unsupervised learning-based registration is to estimate an optimal deformation field by minimizing the following energy loss function:

$$\phi = \text{argmin}\mathcal{L}_{\text{sim}}(F, M \circ \phi) + \lambda \mathcal{L}_{\text{reg}}(\phi) \tag{1}$$

where ϕ represents the deformation field, $\mathcal{L}_{\text{sim}}(\cdot)$ denotes the dissimilarity between the fixed image F and warped moving image (registered image) $M \circ \phi$, $\mathcal{L}_{\text{reg}}$ represents the regularization constraint and λ is the weight of regularization constraint. To solve the optimization problem in Eq. 1, a large number of unsupervised registration methods have been proposed. DIRNet [38] is one early method of applying the unsupervised learning strategy to 2D image registration. [3] introduce an unsupervised learning-based network named VoxelMorph to explore the 3D image registration. It adopts the U-Net [32] structure and achieves good performance in brain MR images. The follow-up works include DIF-VoxelMorph [7] and FAIM [24], which are based on VoxelMorph's U-Net structure.

However, the above-mentioned methods are enforced to make a straightforward estimation of the deformation field, and thus may not work very well when estimating complicated deformation fields with large displacements. Therefore, [37] propose stacking multiple separately trained CNNs for direct affine and deformable image registration. [43] propose an end-to-end recursive cascaded network to refine the deformation field progressively and effectively register images with large displacements. The main problem of these cascaded networks is that they need massive parameters for training, resulting in heavy computational burdens.

2.2. Multi-resolution image registration

In traditional registration methods, multi-resolution or multi-level registration strategies are commonly used to solve the challenge of large displacements and complex deformation [26,22]. This idea is also widely used in learning-based registration methods. [10] incorporate a multi-level registration strategy into the training of a U-Net and let it grow progressively during training. [15] introduce a multi-level framework that can compensate and handle large deformation by computing deformation fields on different scales and functionally composing them. In [17], a CNN-based multi-resolution pyramid combined with feature warping is used to progressively estimate the deformation fields. [29] also propose a registration network based on multi-resolution Laplacian pyramid for large deformation in brain MR image registration,

and they train the network stage-by-stage to guarantee better performance.

Different from these similar works, we facilitate the estimation of large complex deformation by utilizing multi-scale and multi-view contextual information. Our proposed two novel modules, the multi-receptive-field guidance module and multi-view context fusion module, are embedded into the multi-resolution feature pyramid and thus enable the network to be trained in an end-to-end manner without deep supervision. We further introduce a residual learning strategy into the multi-resolution framework to relieve the estimation burden. In addition, we have conducted extensive experiments both on liver CT images and brain MR images to demonstrate that our proposed CPRNet can handle large displacements and complex deformation.

2.3. Topology preservation

Learning-based registration methods are comparable to traditional registration methods in performance, even outperforming them. However, those learning-based methods may grant a high degree of freedom to the deformation field, thus sacrificing the preservation of topology and leading to unrealistic transformation results, such as folding. There have been some efforts to regularize the deformation field. [37] penalize the bending energy of the deformation field directly, assuming that the local deformation field should be smooth. [3] encourage a smooth deformation field using a diffusion regularizer on its spatial gradients. Using the scaling and squaring algorithm to integrate the estimated deformation field is an effective way to reduce folding in some works [8,29], but increases the computational cost as well. Cycle consistency [20] and inverse consistency [42] can also be helpful for topology preservation. Inspired by traditional methods, [16] propose to restrict the determinant of the Jacobian to physiologically meaningful values by combining a volume change penalty with a curvature regularizer in the loss function. Another reasonable way to ensure topology preservation is to require the Jacobian determinant of the deformation field to be positive everywhere, either as a hard constraint or by a penalty method [24,28]. However, those penalties may not work well with large displacement or complex deformation. The reason is that the preservation of topology and the improvement of registration accuracy may be contradictory in the training process, which makes it difficult for the network to converge.

3. Method

Fig. 2 is the overall framework of our proposed CPRNet. It adopts a top-to-down estimation of the deformation field from low to high resolution and consists of four main parts: two multi-resolution feature pyramids, the deformation field estimation module (DFEM), the refined module, and the deformation field regularization module.

3.1. Multi-resolution feature pyramid

Let F and M denote the fixed image and the moving image, respectively. As Fig. 2 shows, given F and M, we generate two multi-resolution 4-level feature pyramids F_i and $M_i (i = 1, 2, 3, 4)$ through pyramid layers with shared weights. The design of multi-resolution feature pyramids can enlarge the receptive fields of CPRNet, which facilitates the capture of locations where large displacements occur. Each pyramid layer is composed of a convolution layer (kernel size = 3, stride = 2) followed by a leaky rectified linear unit layer. The numbers of channels of pyramid layers are 4, 8, 16 and 16 from low resolution to high resolution.

3.2. Deformation field estimation module

The DFEM estimates a deformation field at each resolution and generates the context needed in the next resolution. As illustrated in Fig. 3, the DFEM is composed of four parts, including the feature warping module, the multi-receptive-field guidance module (MRF-GM), the multi-view context fusion module (MV-CFM), and the residual deformation field estimation module (RDFEM). Fig. 3 also demonstrates how to estimate the deformation field at each resolution. For each resolution, the DFEM takes as inputs the fixed image feature F_i, the moving image feature M_i, the deformation field estimated from the last level ϕ_{i-1} and context from the last level c_{i-1} generated by the MV-CFM and outputs the deformation field ϕ_i and context c_i. Specifically, M_i is first warped by $\times 2$ upsampled ϕ_{i-1}, and the warped M_i is denoted as $\widehat{M}_i$. Then, the MRF-GM computes guidance maps d_i between F_i and $\widehat{M}_i$. Next, the RDFEM predicts a residual deformation field $\hat{\phi}_i$ under the guidance of d_i and c_{i-1}. Finally, the deformation field ϕ_i is obtained by adding the $\times 2$ upsampled ϕ_{i-1} and $\hat{\phi}_i$.

3.2.1. Feature warping

Motivated by the image warping used in the recursive cascaded network [43] and feature warping in optical flow estimation [35,18] for addressing large-displacement flow, we propose to reduce the misaligned area between M_i and F_i by warping the moving image feature M_i with the $\times 2$ upsampled deformation field ϕ_{i-1} from the last level. Different from [17], which also use the feature warping to refine the deformation field, we only estimate a residual deformation field $\hat{\phi}_i$ at each resolution, and the details are in Section 3.2.4. The deformation field estimation process becomes easier if M_i and F_i are captured close to each other through feature warping, since the correspondence only needs to be searched in a smaller area. The feature warping process can be written as:

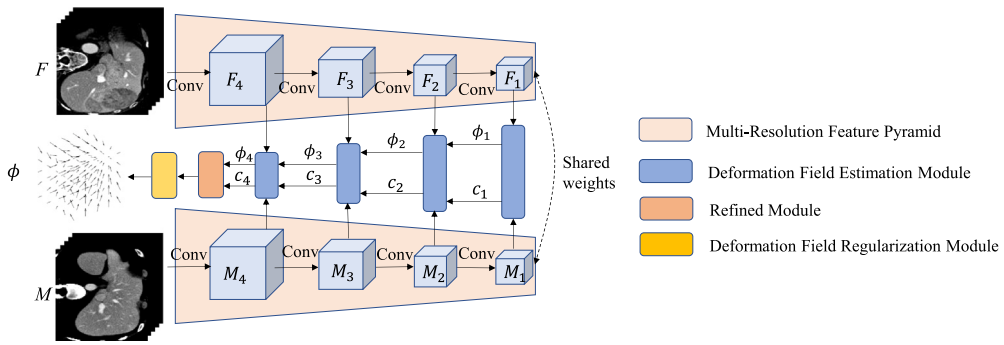


Fig. 2. The framework of CPRNet following a top-to-down estimation of the deformation field.

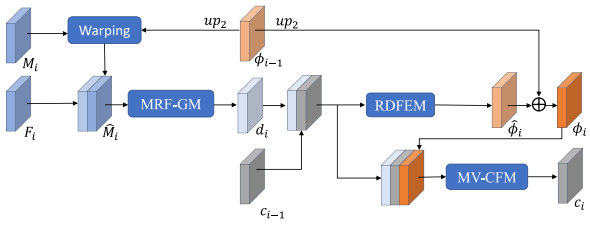


Fig. 3. Details of the deformation field estimation module. MRF-GM: multi-receptive-field guidance module; MV-CFM: multi-view context fusion module; RDFEM: residual deformation field estimation module.

$$\widehat{M}_i = M_i \circ (up_2(\phi_{i-1})) \tag{2}$$

where $\circ$ represents the warping operation and up_2 represents the upsampling process. We use bilinear interpolation to implement the warping operation and compute the gradients to the input features for backpropagation according to the spatial transformer network [19]. The upsampling of ϕ_{i-1} is conducted by using a transpose convolution layer (kernel size = 4,stride = 2).

3.2.2. Multi-receptive-field guidance module

Given the F_i and M_i, a simple way to learn the displacement between F_i and M_i is stacking multiple convolutional layers in series, like estimating optical flow in Flownet-S [9]. Another way is to construct the correlation between features of the fixed image and moving image (in a so-called correlation layer) before estimating the displacement, which has shown great success in the 2D optical flow estimation, like Flownet-C [9] and PWC-net [35]. In [14], the correlation layer, which is transferred to the medical image registration, has been shown effective in estimating large displacements and complex deformation, but costs excessive memory storage.

Inspired by their work, we propose a multi-receptive-field guidance module to automatically capture the spatial correlation between F_i and $\widehat{M}_i$ from the perspective of multi-receptive-field for relieving the heavy computation burdens, where we try to use the receptive field to mimic the search range of the correlation layer. In addition, the setting of multiple receptive fields can also make the calculated guidance maps integrate multi-scale contextual information, which is important for estimating complicated deformation. As shown in Fig. 4, there are four different branches in the MRF-GM, where $1 \times 1 \times 1$ convolutions are used to fuse two input features $\left(F_i, \widehat{M}_i\right)$ and $3 \times 3 \times 3$ convolutions are responsible for enlarging the receptive field. As suggested in Inception [36], pooling operations are essential for deep learning methods, and thus, we add a parallel pooling branch in the MRF-GM, which should have an additional beneficial effect. For those convolution layers in the MRF-GM at different resolutions, the numbers of channels are 16, 32, 64, and 128 from low resolution to high resolution. In addition, every convolution layer in the MRF-GM is followed by a leaky rectified linear unit layer which is ignored in Fig. 4 for simplicity. Features from four branches are concatenated and fed into the RDFEM as a cue to estimate deformation fields.

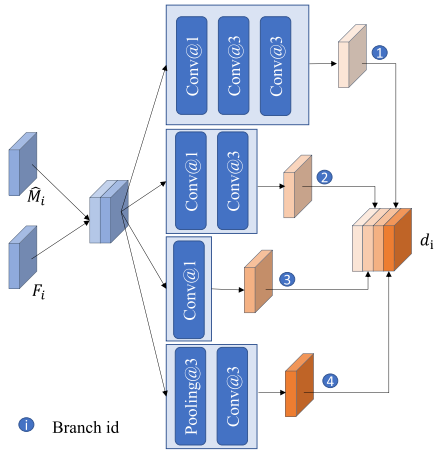


Fig. 4. The structure of the multi-receptive-field guidance module.

3.2.3. Multi-view context fusion module

As discussed in Section 1, deep contextual information which encodes high-level semantic information is important for estimating deformation. For example, the well-worked U-net structure in DIR [3,24] incorporates deep contextual information by fusing the features of the deep layer and shallow layer via skip-connection. It motivates us to extract deep contextual information at each resolution for the estimation of deformation at the next resolution. Generally, deep contextual information needs to be obtained by successive multiple convolutional layers. With the advantage of multi-resolution pyramid network, we can effortlessly extract the required deep contextual information. For our network, three kinds of feature maps can be considered as deep contextual information for the next resolution, including the context from the last resolution, the guidance maps, and the deformation field estimated at the current resolution. Therefore, we design a module to fuse contextual information from the above-mentioned three views. As shown in Fig. 3, the input of the context fusion module consists of feature maps from three views: 1) the context c_{i-1}, 2) MRF guidance maps d_i, and 3) the deformation field ϕ_i:

$$c_i = H_c([d_i, \phi_i, c_{i-1}]) \tag{3}$$

where H_c represents the MV-CFM and $[d_i, \phi_i, c_{i-1}]$ refers to the concatenation of feature maps from d_i, ϕ_i and c_{i-1}. Particularly, c_1 is obtained by the Eq. 4:

$$c_1 = H_c(d_1) \tag{4}$$

The proposed MV-CFM is composed of a transpose convolution layer (kernel size = 4,stride = 2) followed by a leaky rectified linear unit layer. To reduce the computational cost of the network, we limit the maximum extracted contextual information of each resolution by fixing the channel number of convolutional layers, and thus the network automatically fuses the helpful context for prediction. The numbers of channels of context layers are 128, 64, 32, and 16 from low resolution to high resolution.

3.2.4. Residual deformation field estimation module

The RDFEM is composed of a convolutional layer (kernel size = 3, stride = 1). As shown in Fig. 3, it takes d_i and c_{i-1} as its input and outputs the residual deformation field $\hat{\phi}_i$:

$$\hat{\phi}_i = H_r([d_i, c_{i-1}]) \tag{5}$$

where H_r represents the RDFEM and $[d_i, c_{i-1}]$ refers to the concatenation of feature maps from d_i and c_{i-1}. Particularly, $\hat{\phi}_1 = H_r(d_1)$ since there is no context at the lowest resolution. ϕ_i is obtained by Eq. 6:

$$\phi_i = \begin{cases} \hat{\phi}_i, & \text{if} i = 1 \\ up_2(\phi_{i-1}) + \hat{\phi}_i, & \text{otherwise} \end{cases} \tag{6}$$

3.3. Refined module

Since the resolution of the deformation field ϕ_4 generated in the last level is only half of the original image resolution, we use the

refined module to make the final deformation field ϕ consistent with the original image in resolution. The refined module is composed of a convolution layer (kernel size = 3, stride = 1). The refined module is fed with the concatenation of the deformation field ϕ_4 and the context c_4, and outputs the final deformation field ϕ:

$$\phi = R([up_2(\phi_4), c_4]) \tag{7}$$

where R represents the refined module and $[up_2(\phi_4), c_4]$ refers to the concatenation of feature maps from $up_2(\phi_4)$ and c_4.

3.4. Deformation field regularization module

Image registration is an ill-posed problem, and there could be a trade-off between the registration performance and topology preservation constraints. Giving large weight to the regularization constraint in Eq. 1, the deformation field could be excessively regularized, penalizing the registration accuracy. To address this issue, we propose a deformation field regularization module, which applies local Gaussian smoothing filtering to the region where the estimated deformation field is folded. We define C_1 as the set of voxels in the deformation field ϕ whose Jacobian determinant value is non-positive.

$$C_1 = \{p | \forall p \in \Omega, \det(\phi(p)) \leqslant 0\} \tag{8}$$

where Ω denotes the cuboid (or grid) on which input images are defined and $\det(\cdot)$ denotes the Jacobian determinant of deformation field ϕ at voxel p. C_2 is defined as follows:

$$C_2 = \{p | \exists q \in \delta_r(p), q \in C_1\} \tag{9}$$

where δ_r is the neighborhood of the voxel p with radius r. The operation of the deformation field regularization module at voxel p is formulated as follows:

$$\phi_r(p) = \begin{cases} \phi(p) * g, & \text{if} p \in C_2 \\ \phi(p), & \text{otherwise} \end{cases} \tag{10}$$

$$g = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^3} e^{-\frac{|x|^2+|y|^2+|z|^2}{2\sigma^2}} \tag{11}$$

where $*$ represents a convolution operation and g is a 3-D Gaussian kernel. As shown in Fig. 5, the module performs local Gaussian smoothing filtering on the area where the folding occurs, which is equivalent to the operation of unfolding.

3.5. Training loss

The training loss is composed of a similarity loss and two regularization losses. Their definitions are as follows:

Similarity loss: The similarity loss measures the spatial, structural, or intensity similarity between the fixed image and the moving image warped by the deformation field ϕ_r. Here, we utilize the normalized correlation coefficient (NCC) as our similarity metric. The similarity loss is defined as:

$$\mathcal{L}_{\text{sim}}(F, M \circ \phi_r) = \quad 1 - NCC(F, M \circ \phi_r) \tag{12}$$

Smooth loss: To encourage the smoothness of the deformation field, we regularize it with the smooth loss. The smooth loss is defined as:

$$\mathcal{L}_{\text{smo}} = \Sigma_{p\in\Omega} \|\nabla \phi_r(p)\|_2^2 \tag{13}$$

Anti-folding loss: Because areas with negative Jacobian determinants are considered to be folded, the anti-folding loss aims specifically at penalizing areas in deformation fields that have many negative Jacobian determinants.

$$\mathcal{L}_{\text{ant}} = \frac{1}{\Omega} \sum_{p\in\Omega} \sigma(-\det(\phi_r)(p)) \tag{14}$$

where $\sigma(\cdot)$ works as the ReLU activation function that is linear for all positive values and zero for all negative values and $\det(\cdot)$ denotes the Jacobian determinant of the deformation field ϕ_r.

Hence, the overall training loss of our proposed method can be written as:

$$\mathcal{L}(F, M, \phi_r) = \mathcal{L}_{\text{sim}} + \lambda_1 \mathcal{L}_{\text{smo}} + \lambda_2 \mathcal{L}_{\text{ant}} \tag{15}$$

where λ_1 and λ_2 are weights to balance the contributions of the smooth loss and anti-folding loss, respectively.

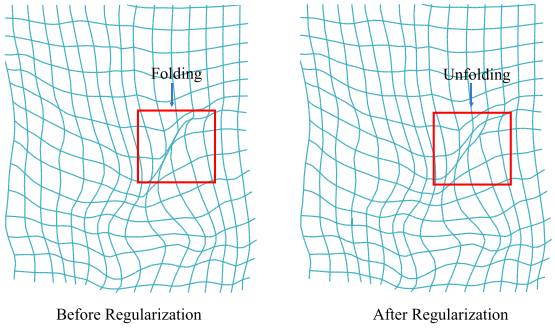


Fig. 5. Demonstration of deformation field regularization.

4. Experiments

We evaluated the effectiveness and generalization performance of our proposed method on two kinds of medical image datasets with different modalities and different objects: liver CT images and brain MR images. We performed subject-to-subject registration both in liver CT and brain MR image registration experiments. The subject-to-subject registration means all images can be fixed images and belongs to the intra-patient registration. In contrast to the atlas-based registration, it more reflects the generalization ability of the registration model.

4.1. Datasets

In liver CT experiments, we use a large-scale, multi-site dataset of 1025 liver scans from two publicly available datasets as the training datasets: MSD [43] and BFD [44]. Three publicly datasets with liver segmentations serve as testing datasets, namely, Sliver (20 scans) [44], LITS (131 scans) [25], and BGCV (100 scans) [39].

In Brain MRI experiments, we collected 414 brain MRI scans from OASIS-1 dataset [27]. The initial OASIS-1 dataset consists of a cross-sectional collection of 416 subjects covering the adult life span aged 18 to 96 including individuals with early-stage Alzheimer's Disease (AD). Two scans were excluded due to preprocessing failure. We generated 35 anatomical structures labels for each subject using SAMSEG [30]. The dataset was divided into 300, 14, and 100 scans for training, validation, and testing respectively.

4.2. Implementation

For liver CT image registration experiments, all the images in the above datasets are resampled into $128 \times 128 \times 128$ voxels after cropping out the unnecessary parts around the liver. λ_1 and λ_2 in Eq. 15 are set to 1.0 and 1×10^{-4}, respectively. For the brain MR image registration experiments, standard brain MR image deformable registration pre-processing procedures were performed, including skulls removal using FreeSurfer [13], resampling into $128 \times 128 \times 128$ voxels, and affine spatial normalization. λ_1 and λ_2 in Eq. 15 are set to 0.1 and 1×10^{-4}, respectively. The radius

r in Eq. 9 is set to 5 in all experiments. The proposed CPRNet was implemented using TensorFlow [1] and executed on one card of 32G NVIDIA Tesla v100 GPU. We trained the proposed CPRNet using a batch size of 4 for 10^5 iterations with the Adam optimizer [21] both in liver CT and brain MR image registration experiments. The learning rate is initially 1×10^{-4} and halved after 6×10^4 steps and again after 8×10^4 steps. The source code is available at https://github.com/MedicalReg/CPRNet-main.

4.3. State-of-the-art methods

We compare our proposed CPRNet with four recent state-of-the-art learning-based methods: VoxelMorph [3], VTN [44], FAIM [24] and CycleMorph [20]. VoxelMorph often serves as the baseline method in unsupervised DIR. VTN achieves state-of-the-art registration performance on liver CT datasets, which has been proposed recently. FAIM encourages topology preservation through anti-folding penalty terms. Cyclemorph introduces the idea of cycle consistency to enhance topology preservation, and we choose the global CycleMorph version for comparison. We also compare our proposed method with a cascaded network proposed by [43] since cascaded networks achieve great registration performance in registering largely-displaced images. Due to limitations of memory and computing resources, we only compare our proposed method with VTN(ADD) which cascades an affine subnetwork and two deformable subnetworks. Letters in brackets represent the type and number of subnetworks in the cascaded network. For a fair comparison, we reproduce the experiments of other methods under the same training setting as ours, except for the hyper-parameters which are finetuned respectively in each method to achieve the best performances.

4.4. Evaluation

To assess the registration performance of different methods, the Dice coefficient, landmark distance, and folding ratio were used.

4.4.1. Dice coefficient

The Dice coefficient measures the spatial overlap of anatomical segmentation maps between the fixed image and the registered image. When two images are perfectly aligned, the value of the Dice coefficient is equal to 1.0. In liver CT image registration, we calculate the Dice coefficient with liver segmentation ground truth. In brain subject-to-subject MR image registration, we calculate an average Dice coefficient with 35 anatomical structures.

4.4.2. Landmark distance

The landmark distance (Lm dist) is the average Euclidean distance between warped anatomical landmarks and the ground truth in voxel units. We only use this metric on the Sliver dataset since only this dataset contains six anatomical landmarks as the ground truth. A lower Lm dist means a better registration performance.

4.4.3. Folding ratio

The folding ratio can be obtained by calculating the ratio of voxels whose Jacobian determinants are non-positive in the deformation field. As image folding is an abnormal phenomenon in medical image registration that violates the anatomical correspondence, the lower the folding ratio is, the better the registration performance will be.

5. Results

5.1. Liver CT image registration

Table 1 shows quantitative evaluation results with the Dice coefficient, Lm dist, and folding ratio for comparing different methods. To further quantify the difference between the proposed method and comparative methods, Wilcoxon rank-sum test between the comparative methods and the proposed CPRNet is performed, in which the p-value level is marked in Table 1. In general, $p < 0.05$ denotes statistically improvements, and $p < 0.01$ denotes obvious statistically improvements. By comparison, we find that the proposed CPRNet significantly improves the Dice coefficient on three evaluation datasets from $0.879, 0.834, 0.824$ to $0.931, 0.884, 0.851$ and reduces the Lm dist by more than 3.16 compared with the baseline method VoxelMorph. Also, the proposed CPRNet achieves a significant registration performance improvement in terms of Dice coefficient and folding ratio compared with VTN, FAIM, and CycleMorph. It's worth noticing that the cascaded network VTN(ADD) effectively improves the registration accuracy in terms of the Dice coefficient and Lm dist compared with other methods, but its folding ratio is as high as approximately 1.3%. Instead, CPRNet achieves competitive performance compared with VTN(ADD) in terms of the Dice coefficient and Lm dist, while the folding ratio is reduced to lower than 0.2%.

Fig. 6 shows visualization results of different methods on an example with a large displacement. In Fig. 6(a), it can be seen that the registered images of VoxelMorph, VTN, FAIM, and CycleMorph are slightly aligned to the fixed image. The registered image of VTN (ADD) is well aligned to the fixed image, while some details in the moving image become unnatural. For comparison, the proposed CPRNet achieves better visual performance both in details and anatomical correspondence. The flow field in Fig. 6(b) reflects displacements of deformation fields along three different directions. It is observed that the flow fields of VoxelMorph, VTN, FAIM and CycleMorph are different from other better-performing methods' flow fields, which indicates that those methods may not find an optimal deformation field. Fig. 6(c) shows the anatomical correspondence, which reflects the Dice coefficient and Lm dist, and it is observed that the CPRNet reaches the better performance in anatomical correspondence. We visualize regions with non-positive Jacobian determinants with the red color in deformation fields, as shown in Fig. 6(d). We observe that the proposed CPRNet effectively reduces the red area considered to be folded compared with the other methods.

5.2. Brain MR image registration

Table 2 shows the results of the Dice coefficient and folding ratio for brain MR image registration with various comparative methods. The p-value level from the Wilcoxon rank-sum test between the comparative methods and the proposed CPRNet is also shown in Table 2. It is observed that the proposed CPRNet outperforms these recent existing learning-based methods both on Dice coefficient and folding ratio. The proposed CPRNet achieves 2.27% Dice coefficient improvements in comparison with the baseline method VoxelMorph. Both FAIM and CycleMorph achieve significant improvements in topology preservation, with a much lower folding ratio than VoxelMorph and VTN. As a comparison, the proposed CPRNet further reduces the folding ratio while improving the Dice coefficient compared to FAIM and CycleMorph. Fig. 7 depicts the box plots of the Dice coefficients obtained for 35 labeled anatomical structures on the OASIS dataset. It can be observed that our proposed CPRNet achieves better registration

Table 1
Comparison among different methods in liver CT image registration experiments. Standard deviations across instances are in parentheses. Wilcoxon rank-sum test with Bonferroni correction between the comparative methods and the proposed method CPRNet is performed on each dataset. One asterisk (∗) denotes $p < 0.05$ and two asterisks (∗∗) denote $p < 0.01$.

Method	Sliver			LiTS		BGCV	
	Dice	Lm Dist	Folding(%)	Dice	Folding(%)	Dice	Folding(%)
VoxelMorph	0.879(0.037) ∗∗	16.90(6.30) ∗∗	0.82(3.16e−3) ∗∗	0.834(0.059) ∗∗	0.65(2.10e−3) ∗∗	0.824(0.059) ∗∗	0.47(2.15e−3) ∗∗
VTN	0.883(0.035) ∗∗	15.91(6.14) ∗∗	0.40(2.11e−3) ∗∗	0.840(0.057) ∗∗	0.29(1.74e−3) ∗∗	0.818(0.054) ∗∗	0.14(8.02e−4) ∗
FAIM	0.873(0.039) ∗∗	17.10(6.47) ∗∗	0.39(7.85e−4) ∗∗	0.825(0.068) ∗∗	0.33(7.60e−4) ∗∗	0.821(0.061) ∗∗	0.26(6.45e−3) ∗∗
CycleMorph	0.881(0.036) ∗∗	16.20(6.21) ∗∗	0.82(2.34e−3) ∗∗	0.836(0.058) ∗∗	0.63(1.79e−3) ∗∗	0.814(0.056) ∗∗	0.32(1.30e−3) ∗∗
VTN(ADD)	0.935(0.020)	12.28(4.72)	1.32(5.67e−3) ∗∗	0.893(0.044) ∗	1.21(5.59e−3) ∗∗	0.856(0.052)	1.29(5.17e−3) ∗∗
CPRNet	0.931(0.023)	12.75(4.82)	1.95e−2(1.60e−4)	0.884(0.048)	1.30e−2(1.10e−4)	0.851(0.052)	8.01e−3(8.07e−5)

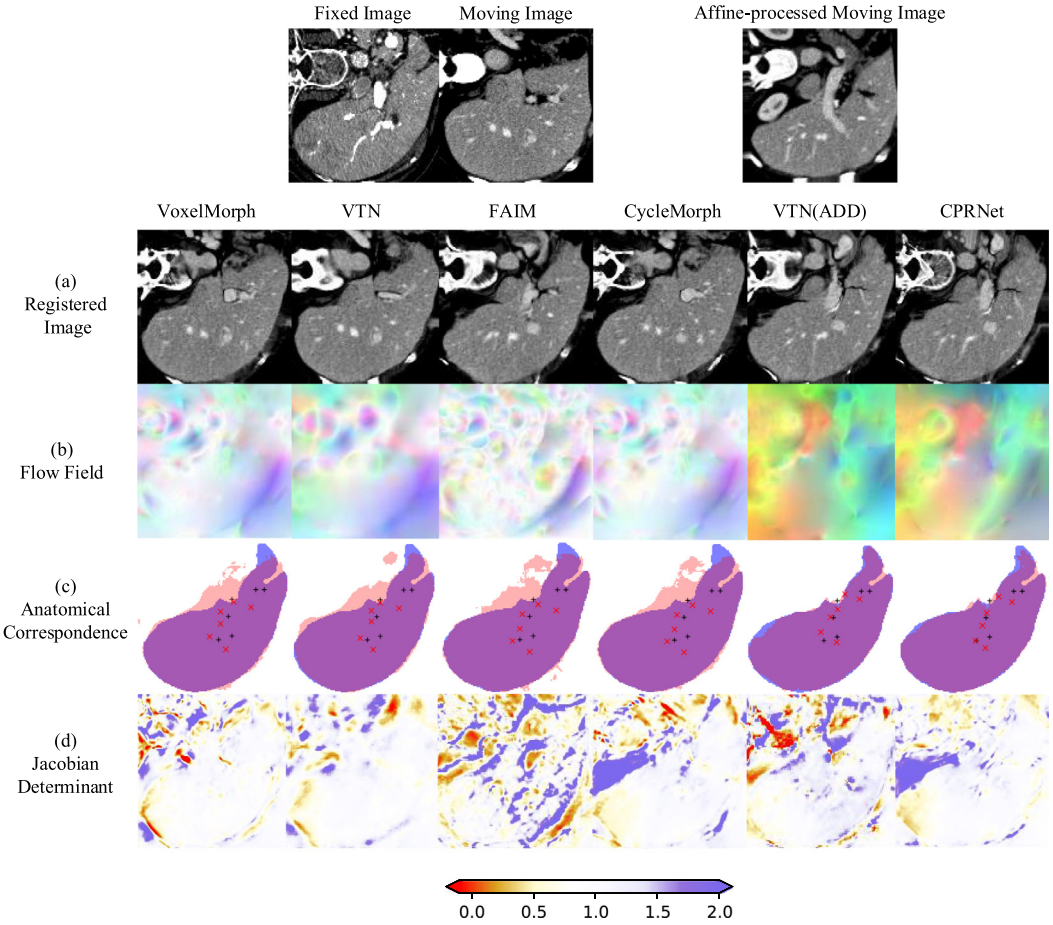


Fig. 6. A visual example showing the registration performance of different methods. (a) Registered image; (b) Flow fields are drawn by mapping the absolute value of three components (x, y, z) of deformation fields into color channels (R, G, B) respectively; (c) Black plusses represent landmarks of the fixed image, and red crosses represent landmarks of the registered image. Segmentations are sliced. Blue areas represent the segmentation of the fixed image, and red areas represent the segmentation of the registered images. Coincident areas are in purple; (d) The colormap of the Jacobian determinant with singularities (folding) indicated in the bright red. The affine-processed moving image is obtained through the affine subnetwork of VTN (ADD).

performance on most anatomical structures compared with other methods.

The complicated anatomical structure of the brain usually leads to complex deformation in brain MRI registration. Fig. 8 shows visualization results of different methods on an example with a potentially complex deformation. According to the tissue in the orange boxes, it can be observed in Fig. 8(a) that the registered image obtained by the proposed CPRNet is structurally more similar to the fixed image than other methods, which can be specifically verified with the segmentation boundaries of several selected brain structures. The colormaps of Jacobian determinants in Fig. 8(c) show that the proposed CPRNet mitigates some image foldings present in the comparative methods.

5.3. Ablation studies

In order to explore the reasons for the superior registration performance of the proposed CPRNet and evaluate the effectiveness of our innovations, we compare our method by its key variants in this section.

5.3.1. Multi-resolution feature pyramid

First, we conduct an experiment on the liver CT image registration to compare the performance of CPRNet with different numbers of multi-resolution feature pyramid layers. Table 3 reports the results of the Dice coefficient and Lm dist. From $n = 2$ to $n = 4$, the Dice coefficient is increased from 0.871, 0.819, 0.813 to

Table 2
Comparison among different methods in brain MR image registration experiments. Standard deviations across instances are in parentheses. Wilcoxon rank-sum test with Bonferroni correction between the comparative methods and the proposed method CPRNet is performed. One asterisk (∗) denotes $p < 0.05$ and two asterisks (∗∗) denote $p < 0.01$.

Method	OASIS	
	Dice	Folding(%)
VoxelMorph	0.749(0.027) ∗	0.57(6.00e−4) ∗∗
VTN	0.721(0.025) ∗∗	0.68(7.69e−4) ∗∗
FAIM	0.727(0.035) ∗∗	1.28e−2(2.14e−5) ∗∗
CycleMorph	0.757(0.022) ∗	2.73e−2(7.10e−5) ∗∗
VTN(ADD)	0.754(0.020) ∗	4.34e−2(1.17e−4) ∗∗
CPRNet	0.766(0.023)	2.23e−3(1.41e−5)

$0.931, 0.884, 0.851$ and Lm dist declines from 17.32 to 12.75, which indicates that the registration performance greatly improves as the number of multi-resolution feature pyramid layers increases. However, the performance of 5-level CPRNet is inferior to that of 4-level CPRNet. Therefore, we adopt the best-performing 4-level CPRNet as the backbone architecture in the following ablation experiments.

5.3.2. Multi-receptive-field guidance module

To verify the contribution of the multi-receptive-field correlation module in the MRF-GM, we conduct an experiment on the liver CT image registration to compare the performance improvement of different branch groups in the MRF-GM. For the liver CT image registration experiments, as shown in Table 4, the CPRNet with more receptive fields increases the Dice coefficient by approximately 1–4 points and reduces the Lm dist by more than 0.62 compared with those removing any branch in the MRF-GM, and the results of Wilcoxon rank-sum test in Table 4 shows that there are statistical differences between the ablation methods and the proposed CPRNet. It proves that parallel branches with different receptive fields are all conducive to better registration performance.

We also conduct an ablation study to compare the CPRNet with and without MRF-GM. The CPRNet without MRF-GM is implemented by replacing the MRF-GM with one branch of multiple convolution layers as previous works do. Specifically, we replace the proposed MRF-CM with three convolutional layers (kernel size = 3). The results are shown in Table 5. It can be observed that the proposed CPRNet without MRF-GM achieves 0.916, 0.872, and 0.833 in Dice coefficient on three different datasets, and the Lm dist of it is 13.60. Compared with the ablation method, the results of our proposed method are better.

5.3.3. Multi-view context fusion module

To evaluate the contribution of the proposed MV-CFM, we conduct an experiment on liver CT image registration for the registration performance comparison between the proposed CPRNet with and without the MV-CFM. As shown in Table 6, CPRNet + MV-CFM achieves average enhancements of $5.5\%, 5.2\%$, and 2.5% in the Dice coefficient on three evaluation datasets, compared with the CPRNet without the MV-CFM. Additionally, the Lm dist decreases by 2.85. The results indicate that the CPRNet benefits from MV-CFM since it can fuse multi-view contextual information. Fig. 9 shows axial CT slices example from the resulting registered image for CPRNet with and without the MV-CFM. The red box locates where the kidney is, and it is observed that the kidney has a large displacement along the z-axis. In Fig. 9(c), it can be seen that the CPRNet without MV-CFM fails to align the kidney in the moving image with the kidney in the fixed image, which indicates that it falls into a local optimum. As a comparison, the kidney and liver are well aligned in Fig. 9(d). According to the sampling grid corresponding to the red box in Fig. 9(c) and (d), the color of the grid measures the displacement along the z-axis. It can be seen from grids that the CPRNet with the MV-CFM finds that the kidney has a displacement along the z-axis. This indicates that the contextual information provided by MV-CFM can prevent the proposed CPRNet from the local optimum.

To further verify the contribution of every view in the MV-CFM, we conduct an experiment with different view groups, and the results are illustrated in Table 7. The removal of any view in the MV-CFM results in a decrease of 1%–2% in the Dice coefficient and an increase of 0.68–1.55 in the Lm dist compared with the CPRNet with the complete MV-CFM. The results of the Wilcoxon rank-sum test in Table 7 further verify that the CPRNet with the complete MV-CFM achieves statistically improvements over the ablation methods. All of the above results prove that our proposed MV-CFM can improve the registration performance by integrating multi-view contextual information.

5.3.4. Residual learning strategy

To explicitly demonstrate the contribution of residual learning strategy (RLS), we compare our proposed CPRNet with and without RLS. The CPRNet without RLS is implemented by removing the feature warping operation and addition operation of the deformation field from the workflow of our proposed method. The results are shown in Table 8. It can be seen that the RLS makes the proposed CPRNet achieve an increase of more than 2.5% in the Dice coefficient and a reduction of 3.56 in the Lm dist, which indicates that it makes a great contribution to facilitating the registration performance.

5.3.5. Deformation field regularization module

We evaluate registration performance with and without deformation field regularization module on the brain MR image registration experiments. Fig. 10 shows the training loss curve of the proposed CPRNet with and without deformation field regularization. We can observe the network with deformation field regularization converges after about 20 k iterations while the network without deformation field regularization starts to converge after about 50 k iterations and is prone to violent oscillations. Moreover, the network with deformation field regularization converges to a lower value according to the zoom-in view. Table 9 presents the effect of the proposed deformation field regularization module on Dice coefficient and folding ratio with varying anti-folding weights. When the λ_2 increases from 1e-5 to 1e-3, the network without regularization penalizes the Dice coefficient from 0.761 to 0.750 and reduces the folding ratio from 3.30e-2 to 3.97e-3. By comparison, the network with regularization keeps the folding ratio below 1e-2 while maintaining the Dice coefficient above 0.759.

Fig. 11 shows the registration results of CPRNet with and without deformation field regularization module when the weight λ_2 is equal to $1e-4$. It can be observed that the CPRNet with deformation field regularization module provides better anatomical correspondence and fewer folding artifacts.

6. Discussion

In this study, we present a novel context-driven pyramid registration network for unsupervised DIR, which is dedicated to overcoming the challenge of registering medical images with large complex deformation while enhancing topology preservation.

First, to overcome the challenge of registering images with large displacement, we address this issue by devising a multi-resolution pyramid architecture and the MV-CFM built on it. Previous methods mainly make a straightforward estimation of deformation, which are generally not suitable for scenarios with large displacement

Fig. 7. Boxplots of Dice coefficients for 35 labeled anatomical structures on the OASIS dataset.

and complex deformation. To solve the problem, we first consider constructing multi-resolution feature pyramids to estimate the deformation field in a coarse-to-fine manner, which has been proved effective in our experiments. The multi-resolution feature pyramids give a much wider search range to perceive the large displacement. As illustrated in Table 3, we can observe that the registration performance of 2-level CPRNet is much worse than those of others. The reason behind this is that the 2-level CPRNet has difficulty capturing locations where large displacements occur. In addition, we also find that the 5-level CPRNet fails to improve the registration performance. The reason is that the 4-level CPRNet has been enough to deal with all large displacements since its smallest resolution is as low as $8 \times 8 \times 8$. With the advantages of the multi-resolution framework, we propose the MV-CFM incorporate multi-view contextual

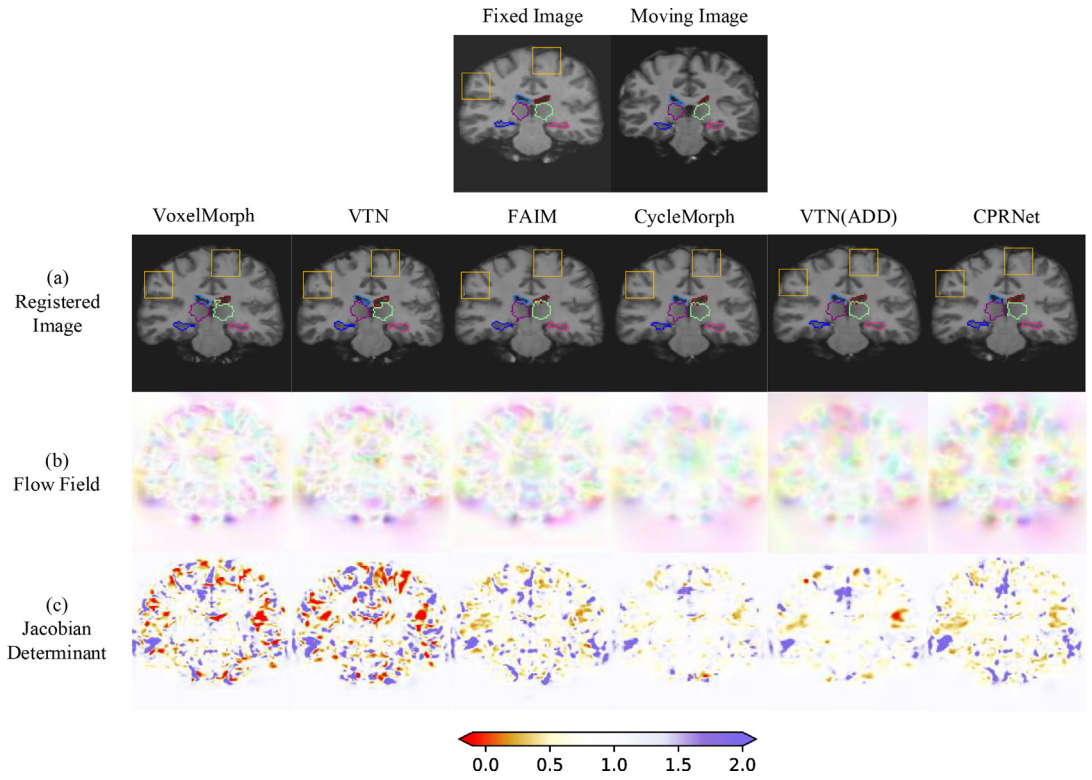


Fig. 8. A visual example showing the registration performance of different methods on the brain MR image registration. (a) Registered image with overlaid boundaries of ventricles, thalamus and hippocamus. Orange boxes highlight our accurate performance; (b) Flow fields are drawn by mapping the absolute value of three components (x, y, z) of deformation fields into color channels (R, G, B) respectively; (c) The colormap of the Jacobian determinant with singularities (folding) indicated in the bright red.

Table 3
Comparison of results with different number of layers in the multi-resolution feature pyramid on the liver CT datasets. Standard deviations across instances are in parentheses.

Method	Sliver		LiTS	BGCV
	Dice	Lm Dist	Dice	Dice
$n = 2$	0.871(0.040)	17.32(6.57)	0.819(0.065)	0.813(0.061)
$n = 3$	0.905(0.033)	16.31(6.37)	0.858(0.056)	0.825(0.058)
$n = 4$	0.931(0.023)	12.75(4.82)	0.884(0.048)	0.851(0.052)
$n = 5$	0.930(0.023)	12.70(4.78)	0.881(0.049)	0.834(0.060)

Table 4
Comparison of the improvement brought by each branch of the MRF-GM. Standard deviations across instances are in parentheses. Wilcoxon rank-sum test with Bonferroni correction between the ablation methods and the proposed method CPRNet is performed on each dataset. One asterisk $(*)$ denotes $p < 0.05$ and two asterisks $(**)$ denote $p < 0.01$.

Branch id				Sliver		LiTS	BGCV
1	2	3	4	Dice	Lm Dist	Dice	Dice
X	✓	✓	✓	0.897(0.034) **	15.16(5.84) **	0.838(0.054) **	0.826(0.055) **
✓	**X**	✓	✓	0.914(0.030) **	13.73(5.23) *	0.873(0.052) *	0.832(0.057) **
✓	✓	**X**	✓	0.921(0.026) *	13.93(5.16) *	0.874(0.051) *	0.837(0.057) *
✓	✓	✓	**X**	0.925(0.026) *	13.37(4.87) *	0.880(0.049)	0.841(0.053) *
✓	✓	✓	✓	0.931(0.236)	12.75(4.82)	0.884(0.048)	0.851(0.052)

Table 5
Comparison of the impact of MRF-GM on our proposed network. Standard deviations across instances are in parentheses.

Method	Sliver		LiTS	BGCV
	Dice	Lm Dist	Dice	Dice
w/o MRF-GM	0.916(0.027)	13.60(5.12)	0.872(0.051)	0.833(0.056)
w/ MRF-GM	0.931(0.023)	12.75(4.82)	0.884(0.048)	0.851(0.052)

information for facilitating the estimation of large deformation. As illustrated in Table 6, the proposed MV-CFM is a crucial module for the pyramid framework that significantly improves the registration accuracy when registering largely-displaced images. One reason for the performance improvement brought by MV-CFM is that the MV-CFM provides sufficient semantic information which is

Table 6
Comparison of the impact of MV-CFM on our proposed network. Standard deviations across instances are in parentheses.

Method	Sliver		LiTS	BGCV
	Dice	Lm Dist	Dice	Dice
w/o MV-CFM	0.876(0.036)	15.64(6.06)	0.832(0.057)	0.826(0.056)
w/ MV-CFM	0.931(0.023)	12.75(4.82)	0.884(0.048)	0.851(0.052)

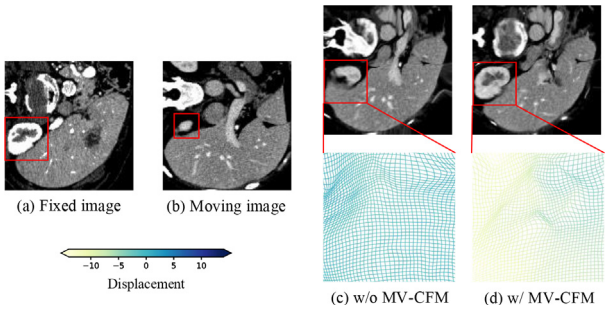


Fig. 9. A visualization of the impact of the MV-CFM on registration. The deformation field of the area in the red box is drawn in the form of a grid, where the depth of the color represents the displacement in the *z*-axis. The bottom color bar gives the mapping between the displacement and the color.

beneficial to the global alignment, as shown in Fig. 9. Another reason is that it ensures information interactions between layers of the multi-resolution pyramid and makes each DFEM receive additional supervision from the loss function through the MV-CFM. When compared with other learning-based methods, we have also certified the promising advantages of the proposed method in the experiments. As shown in Fig. 6, most learning-based methods struggle to register images with a large displacement. Although the cascaded network VTN(ADD) [43] can work very well in registering large-displaced images, the folding in the deformation field may accumulate during the subnetwork cascading.

Next, to enhance the ability in handling complex deformation, two important contributions to our proposed method should be discussed. One is the residual learning strategy. The proposed residual learning strategy allows the proposed CPRNet to only estimate a residual deformation field which is extremely beneficial for learning a complicated deformation field. Fig. 12 gives an insight into how the proposed CPRNet progressively learns to estimate the deformation field through our residual learning strategy. As the resolution increases, the color of the flow fields becomes more complex, which indicates that the deformation field is refined progressively. Another contribution is that we integrate the MRF-GM into our proposed CPRNet to help estimate a residual deformation field reasonably. Guidance maps generated by the MRF-GM encode multi-scale correlations between features of the fixed image and moving image. The results in Table 4 prove that multi-receptive-field guidance maps can facilitate the estimation of the complex deformation field. An explanation is that multi-receptive-field guidance maps can help the proposed CPRNet well judge various degrees of deformation. We have validated the proposed method on brain MRI scans, and the results show that the network can yield a better registration performance than recent learning-based methods on complex deformation scenarios, as shown in Table 2 and Fig. 8.

Last, to enhance the topology preservation, we propose the deformation field regularization module. The background of proposing the regularization module is that adding an anti-folding constraint to the network alone will hinder the convergence of the network and affect the registration accuracy. In order to alleviate the interference of the anti-folding constraint, similar to unfolding, we apply an external force to the predicted deformation field. The experimental results in Table 9 and Fig. 10 prove the proposed regularization module can handle the trade-off between registration performance and topology preservation. The reason behind this is that the regularization model leverages the anti-folding prior knowledge to lower the upper bound of the objective. In this way, the objective function is forced to find a lower point in the loss landscape during the optimization under the same training setting, thus improving the trade-off between registration accuracy and topology preservation. Moreover, we also observe that the regularization module allows the network to converge faster in the early training stage. This is mainly due to the balanced gradient contributions between the anti-folding loss and similarity loss. Without the regularization module, the anti-folding loss would become larger than the counterpart due to the significant number of folding areas, overwhelming the optimization of the similarity loss. In contrast, the pre-regularization of the deformation field reduces the dominancy of the anti-folding loss, enabling the network to concentrate on both losses and to have an unbiased gradient descent direction than before. In addition, the regularization module can effectively compensate for the defect in the anti-folding loss function that the negative determinant with a small value is penalized insufficiently. Compared with other learning-based methods that also focus on topology preservation, such as FAIM [24] and CycleMorph [20], the results in Table 1 and Table 2 show that our proposed method achieves a consistently better performance in reducing folding.

In order to more comprehensively demonstrate our proposed method, here we discuss the model size and run time of our proposed method. Model size measures the number of parameters to be trained. It is associated with the computation burden. Run time is the average time taken during training for each pair of

Table 7
Comparison of the performance improvement brought by each view of the MV-CFM. Standard deviations across instances are in parentheses. Wilcoxon rank-sum test with Bonferroni correction between the ablation methods and the proposed method CPRNet is performed on each dataset. One asterisk (∗) denotes $p < 0.05$ and two asterisks (∗∗) denote $p < 0.01$.

			Sliver		LiTS	BGCV
Context	Correlation maps	Deformation field	Dice	Lm Dist	Dice	Dice
X	✓	✓	0.922(0.028) ∗	13.43(5.79) ∗	0.878(0.059) ∗	0.839(0.054) ∗
✓	**X**	✓	0.921(0.029) ∗	13.70(5.26) ∗	0.878(0.051) ∗	0.842(0.053) ∗
✓	✓	**X**	0.912(0.030) ∗∗	14.30(5.38) ∗∗	0.869(0.052) ∗∗	0.829(0.056) ∗∗
✓	✓	✓	0.931(0.023)	12.75(4.82)	0.884(0.048)	0.851(0.052)

Table 8
Comparison of the impact of residual learning strategy. Standard deviations across instances are in parentheses.

Method	Sliver		LiTS	BGCV
	Dice	Lm Dist	Dice	Dice
w/o RLS	0.895(0.033)	16.31(6.37)	0.849(0.056)	0.826(0.053)
w/ RLS	0.931(0.023)	12.75(4.82)	0.884(0.048)	0.851(0.052)

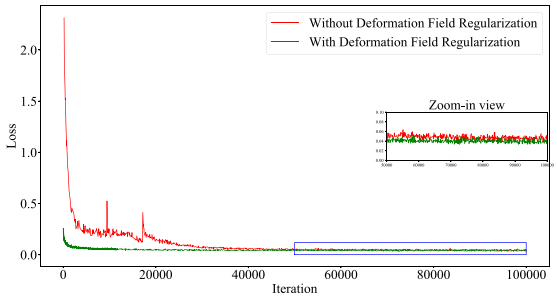


Fig. 10. Training loss curve of the proposed CPRNet with and without deformation field regularization when the anti-folding constraint $\lambda_2 = 1e-4$.

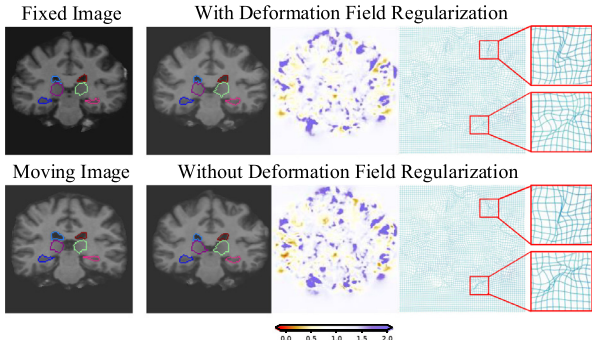


Fig. 11. Registration results of the CPRNet with and without deformation field regularization module.

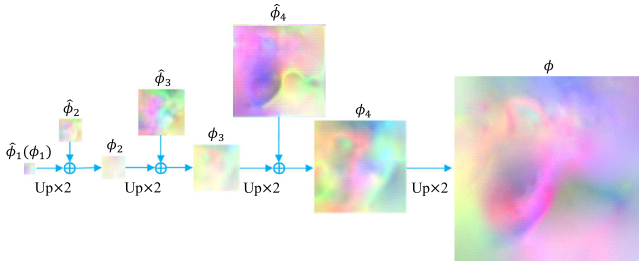


Fig. 12. The residual deformation fields $\hat{\phi}_i$ and composite deformation fields ϕ_i estimated at each resolution are visualized in the form of flow fields. Flow fields are drawn by mapping the absolute value of three components (x, y, z) of deformation fields into color channels (R, G, B) respectively. White area indicates zero displacement.

images to be registered. Measurements for the run time were evaluated on one card of 32G NVIDIA Tesla v100 GPU with a batch size of 1. In terms of model size, the proposed CPRNet contains 25,079,384 parameters. The cascaded network VTN(ADD) with similar performance to our proposed method, includes 212,012,085 parameters. Considering the registration accuracy and topology preservation, the proposed CPRNet is efficient in general. In addition, the proposed CPRNet's average run time on GPU is 0.11s which also meets the requirements of applying it in real-time registration applications.

This study still has some limitations that should be acknowledged. Although we have improved the topology preservation through the deformation field regularization module, the folding phenomenon is still found in some complex deformation. As can be seen in Table 9, the folding ratio has not reached zero. We argue that the possible reason is that the Jacobian determinant penalty mainly focuses on very local topological errors and cannot detect very large smooth folds [6]. In other words, folding may exist at some pixels but is not detected. Therefore, a loss function that penalizes global folds in learning-based registration remains explored. In addition, the proposed regularization module has been verified in this paper to accelerate network convergence and reduce the folding phenomenon, and thus more regularization modules that can better perform "unfolding" operations can be further explored. We also notice the landmark distance is high in some cases, which indicates the alignment inside the liver is not perfect. Although this metric is indeed not easy to improve due to the large variants among different images, our proposed method still has room for improvement in the alignment of some anatomical details. In future studies, we will explore additional constraints on landmark distance or local similarity metrics (e.g. local NCC) to help the model register the anatomical details more accurately and make it generally applicable in other medical image registration tasks, such as lung CT registration.

7. Conclusion

In this paper, we propose a novel unsupervised context-driven pyramid registration network called CPRNet for 3D medical image registration, which can address the challenge of registering medical images with large displacements and complex deformation while enhancing topology preservation. The proposed CPRNet is built based on two multi-resolution feature pyramids. We facilitate the registration performance through two novel modules, including the multi-receptive-field guidance module and multi-view context fusion module. We further develop a residual estimation strategy for progressively estimating the deformation. In addition, we design a deformation field regularization module to accelerate the network convergence and enhance the topology preservation.

Table 9
Dice coefficient and folding ratio with different anti-folding weights λ_2. Standard deviations across instances are in parentheses.

λ_2	Without Regularization		With Regularization	
	Dice	Folding(%)	Dice	Folding(%)
1e−5	0.761(0.026)	3.30e−2(6.52e−5)	0.765(0.023)	4.61e−3(2.00e−6)
1e−4	0.753(0.022)	4.06e−3(3.88e−5)	0.766(0.023)	2.23e−3(1.41e−5)
1e−3	0.750(0.024)	3.97e−3(1.75e−5)	0.759(0.024)	1.11e−3(8.05e−6)

Extensive experiments both in the liver CT image registration and the brain MR image registration demonstrate that our method achieves state-of-the-art performances. Compared with existing learning-based registration methods, our proposed method is well balanced among registration accuracy, topology preservation, model size and run time, which has potential to be applied to various registration tasks.

CRediT authorship contribution statement

Peng Wang: Methodology, Validation, Writing – original draft. **Yunqi Yan:** Formal analysis. **Lijun Qian:** Data curation. **Shiteng Suo:** Investigation, Resources. **Jianrong Xu:** Conceptualization, Investigation. **Yi Guo:** Conceptualization, Writing – review & editing, Project administration. **Yuanyuan Wang:** Supervision, Project administration, Funding acquisition.

Data availability

The authors do not have permission to share data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, et al., Tensorflow: large-scale machine learning on heterogeneous systems. software available from tensorflow. org. 2015. https://www.tensorflow.org.

[2] B. Avants, N. Tustison, G. Song, Advanced normalization tools (ants), Insight J. 2 (2009) 1–35.

[3] G. Balakrishnan, A. Zhao, M. Sabuncu, J. Guttag, A. Dalca, Voxelmorph: a learning framework for deformable medical image registration, IEEE Trans. Med. Imag. 38 (2019) 1788–1800.

[4] X. Cao, J. Yang, J. Zhang, D. Nie, M. Kim, Q. Wang, D. Shen, Deformable image registration based on similarity-steered cnn regression, International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2017) 300–308.

[5] 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 (2018) 1900–1911.

[6] S. Chun, J. Fessler, A simple regularizer for b-spline nonrigid image registration that encourages local invertibility, IEEE J. Sel. Top. Signal Process. 3 (2009) 159–169.

[7] A. Dalca, G. Balakrishnan, J. Guttag, M. Sabuncu, Unsupervised learning for fast probabilistic diffeomorphic registration, International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2018) 729–738.

[8] A. Dalca, G. Balakrishnan, J. Guttag, M. Sabuncu, Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces, Med. Image Anal. 57 (2019) 226–236.

[9] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, T. Brox, Flownet: Learning optical flow with convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2758–2766.

[10] K. Eppenhof, M. Lafarge, M. Veta, J. Pluim, Progressively trained convolutional neural networks for deformable image registration, IEEE Trans. Med. Imaging 39 (2019) 1594–1604.

[11] K.A. Eppenhof, M.W. Lafarge, P. Moeskops, M. Veta, J.P. Pluim, Deformable image registration using convolutional neural networks, in: Medical Imaging 2018: Image Processing, International Society for Optics and Photonics, 2018, p. 105740S.

[12] K.A. Eppenhof, J.P. Pluim, Pulmonary ct registration through supervised learning with convolutional neural networks, IEEE Trans. Med. Imaging 38 (2018) 1097–1105.

[13] B. Fischl, Freesurfer, NeuroImage 62 (2012) 774–781.

[14] M. Heinrich, Closing the gap between deep and conventional image registration using probabilistic dense displacement networks, International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2019) 50–58.

[15] A. Hering, B. van Ginneken, S. Heldmann, mlvirnet: Multilevel variational image registration network, International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2019) 257–265.

[16] A. Hering, S. Häger, J. Moltz, N. Lessmann, S. Heldmann, B. van Ginneken, Cnn-based lung ct registration with multiple anatomical constraints, Med. Image Anal. 102139 (2021).

[17] X. Hu, M. Kang, W. Huang, M. Scott, R. Wiest, M. Reyes, Dual-stream pyramid registration network, International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2019) 382–390.

[18] T. Hui, X. Tang, C. Loy, Liteflownet: A lightweight convolutional neural network for optical flow estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8981–8989.

[19] M. Jaderberg, K. Simonyan, A. Zisserman, K. Koray, Spatial transformer networks, Advances in Neural Information Processing Systems (2015) 2017–2025.

[20] B. Kim, D.H. Kim, S.H. Park, J. Kim, J.G. Lee, J.C. Ye, Cyclemorph: Cycle consistent unsupervised deformable image registration, Med. Image Anal. 102036 (2021).

[21] D. Kingma, J. Ba, Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[22] S. Klein, M. Staring, K. Murphy, M. Viergever, J. Pluim, Elastix: a toolbox for intensity-based medical image registration, IEEE Trans. Med. Imaging 29 (2009) 196–205.

[23] J. Krebs, T. Mansi, H. Delingette, L. Zhang, F.C. Ghesu, S. Miao, A.K. Maier, N. Ayache, R. Liao, A. Kamen, Robust non-rigid registration through agent-based action learning, International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2017) 344–352.

[24] D. Kuang, T. Schmah, Faim–a convnet method for unsupervised 3d medical image registration, International Workshop on Machine Learning in Medical Imaging, Springer (2019) 646–654.

[25] LiTS, 2018. Liver tumor segmentation challenge. Website. Available at https://competitions.codalab.org/competitions/15595.

[26] J.A. Maintz, M.A. Viergever, A survey of medical image registration, Med. Image Anal. 2 (1998) 1–36.

[27] D. Marcus, T. Wang, J. Parker, J. Csernansky, J. Morris, R. Buckner, Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults, J. Cogn. Neurosci. 19 (2007) 1498–1507.

[28] T. Mok, A. Chung, Fast symmetric diffeomorphic image registration with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 4644–4653.

[29] T. Mok, A. Chung, Large deformation diffeomorphic image registration with laplacian pyramid networks, 2020. arXiv preprint arXiv:2006.16148.

[30] O. Puonti, J.E. Iglesias, K. Van Leemput, Fast and sequence-adaptive whole-brain segmentation using parametric bayesian modeling, NeuroImage 143 (2016) 235–249.

[31] M.M. Rohé, M. Datar, T. Heimann, M. Sermesant, X. Pennec, Svf-net: Learning deformable image registration using shape matching, International Conference on Medical Image Computing and Computer-assisted Intervention, Springer (2017) 266–274.

[32] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, International Conference on Medical Image Computing and Computer-assisted Intervention, Springer (2015) 234–241.

[33] T. Sentker, F. Madesta, R. Werner, gdl-fire_4D) Deep learning-based fast 4d ct image registration, International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2018) 765–773.

[34] H. Sokooti, B. De Vos, F. Berendsen, B. Lelieveldt, I. Išgum, M. Staring, Nonrigid image registration using multi-scale 3d convolutional neural networks, International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2017) 232–239.

[35] D. Sun, X. Yang, M. Liu, J. Kautz, Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8934–8943.

[36] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[37] B. de Vos, F. Berendsen, M. 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.

[38] 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, 2017, pp. 204–212.

[39] Z. Xu, C. Lee, M. Heinrich, M. Modat, D. Rueckert, S. Ourselin, R. Abramson, B. Landman, Evaluation of six registration methods for the human abdomen on clinically acquired ct, IEEE Trans. Biomed. Eng. 63 (2016) 1563–1572.

[40] X. Yang, R. Kwitt, M. Styner, M. Niethammer, Fast predictive multimodal image registration, in: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), IEEE, 2017, pp. 858–862.

[41] X. Yang, R. Kwitt, M. Styner, M. Niethammer, Quicksilver: Fast predictive image registration–a deep learning approach, NeuroImage 158 (2017) 378–396.

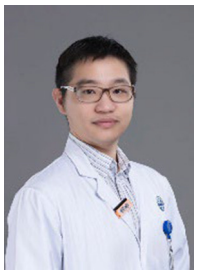
[42] J. Zhang, Inverse-consistent deep networks for unsupervised deformable image registration, 2018. arXiv preprint arXiv:1809.03443.

[43] S. Zhao, Y. Dong, E. Chang, Y. Xu, Recursive cascaded networks for unsupervised medical image registration, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 10600–10610.

[44] S. Zhao, T. Lau, J. Luo, I. Eric, C. Chang, Y. Xu, Unsupervised 3d end-to-end medical image registration with volume tweening network, IEEE J. Biomed. Health Inf. 24 (2019) 1394–1404.

[45] Z. Zheng, W. Cao, Z. He, Y. Luo, Progressive anatomically constrained deep neural network for 3d deformable medical image registration, Neurocomputing 465 (2021) 417–427.
[46] B. Zou, Z. He, R. Zhao, C. Zhu, W. Liao, S. Li, Non-rigid retinal image registration using an unsupervised structure-driven regression network, Neurocomputing 404 (2020) 14–25.

Peng Wang received his M.S. degree in Electronic Engineering from Fudan University, Shanghai, China, in 2020. He is currently pursuing a B.S. degree in Biomedical Engineering at Fudan University, Shanghai, China. His research mainly focuses on deep learning-based image registration.



Yunqi Yan received his B.S. and M.S. degrees in Clinical Medicine in 2008 and 2010, respectively, from the School of Medicine, Shanghai Jiao Tong University. Since 2010, he has worked as a radiologist at Renji Hospital, School of Medicine, Shanghai Jiao Tong University. His research interests are abdominal imaging, gastrointestinal tumors, and inflammatory bowel disease.

Lijun Qian is a radiologist from Renji Hospital, Shanghai Jiao Tong University, School of Medicine, Shanghai, China, since 2007. His primary research interests include hepatobiliary imaging in liver transplantation, local-regional therapy assessment of hepatocellular carcinoma, and pancreatic imaging.

Shiteng Suo received his B.S. and M.S. degrees in Biomedical Engineering in 2010 and 2012, respectively, from Shanghai Jiao Tong University. He is currently a radiologic technologist at Renji Hospital, School of Medicine, Shanghai Jiao Tong University. His research interests are MR imaging, image processing, and machine learning.

Jianrong Xu received his B.S. and M.D. degrees in Clinical Medicine in 1983 and 1990, respectively, from Shanghai Medical University. He is currently a radiologist at Renji Hospital, School of Medicine, Shanghai Jiao Tong University. His research interests are abdominal imaging and computer-aided diagnosis.

Yi Guo received her Ph.D. degree in Biomedical Engineering from Fudan University, Shanghai, China, in 2013. She is currently a professor at the Department of Electronic Engineering, Fudan University. Her current research interests include medical signal and image processing.

Yuanyuan Wang received his B.S., M.S., and Ph.D. degrees in Electronic Engineering from Fudan University, Shanghai, China, in 1990, 1992, and 1994, respectively. From 1994 to 1996, he was a Post-Doctoral Research Fellow with the School of Electronic Engineering and Computer Science, University of Wales, Bangor, U.K. In 1996, he joined the Department of Electronic Engineering, Fudan University, as an Associate Professor, and was then promoted to a Full Professor in 1998, where he is currently the Director of the Biomedical Engineering Center. He has authored or co-authored six books and 500 research articles. His research interests include medical ultrasound techniques and medical image processing.