This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3095409, IEEE Journal of Biomedical and Health Informatics

Few-shot Learning for Deformable Medical Image Registration with Perception-Correspondence Decoupling and Reverse Teaching

Yuting He†, Tiantian Li†, Rongjun Ge, Jian Yang, Youyong Kong, Jian Zhu, Huazhong Shu, Guanyu Yang*, Shuo Li*

***Abstract*—Deformable medical image registration estimates corresponding deformation to align the regions of interest (ROIs) of two images to a same spatial coordinate system. However, recent unsupervised registration models only have correspondence ability without perception, making misalignment on blurred anatomies and distortion on task-unconcerned backgrounds. Label-constrained (LC) registration models embed the perception ability via labels, but the lack of texture constraints in labels and the expensive labeling costs causes distortion internal ROIs and overfitted perception. We propose the first few-shot deformable medical image registration framework, *Perception-Correspondence Registration (PC-Reg)*, which embeds perception ability to registration models only with few labels, thus greatly improving registration accuracy and reducing distortion. 1) We propose the *Perception-Correspondence Decoupling* which decouples the perception and correspondence actions of registration to two CNNs. Therefore, independent optimizations and feature representations are available avoiding interference of the correspondence due to the lack of texture constraints. 2) For few-shot learning, we propose *Reverse Teaching* which aligns labeled and unlabeled images to each other to provide supervision information to the structure and style knowledge in unlabeled images, thus generating additional training data. Therefore, these data will reversely teach our perception CNN more style and structure knowledge, improving its generalization ability.**

Our experiments on three datasets with only five labels demonstrate that our PC-Reg has competitive registration accuracy and effective distortion-reducing ability. Compared with LC-VoxelMorph($\lambda = 1$), we achieve the 12.5%, 6.3% and 1.0% Reg-DSC improvements on three datasets, revealing our framework with great potential in clinical application.

***Index Terms*—Deformable medical image registration, Perception-Correspondence Decoupling, Few-shot learning, Reverse Teaching, Distortion-reducing.**

(†*Equal contribution: Y. He; T. Li.* **Corresponding authors: G. Yang; S. Li.*)

Y. He†, T. Li†, R. Ge, Y. Kong, H. Shu and G. Yang* are with the LIST, Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing 210096, China. (e-mail: yang.list@seu.edu.cn)

G.Yang * is also with the School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China

J. Yang is with the Beijing Engineering Research Center of Mixed Reality and Advanced Display, School of Optics and Electronics, Beijing Institute of Technology, Beijing, China.

J. Zhu is with the Department of Radiation Oncology, Shandong Cancer Hospital, Shandong University, Jinan 250117, China.

S. Li* is with the Department of Medical Imaging, Western University, London, ON N6A 3K7, Canada. (e-mail: sli287@uwo.ca)

This research was supported by the National Key Research and Development Program of China (2017YFC0109202), National Natural Science Foundation under grants (31800825, 31571001, 61828101), Excellence Project Funds of Southeast University and Scientific Research Foundation of Graduate School of Southeast University (YBPY2139). We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

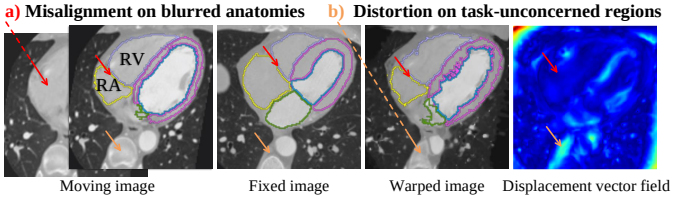

Fig. 1: The loss of perception ability limits the fine alignment for different semantic regions, resulting in: a) the blurred anatomical boundaries will make the misalignment [1] on the anatomical structures; b) the large complication and non-corresponding structures on task-unconcerned regions make distortion and weaken the registration.

I. Introduction

DEFORMABLE medical image registration [2], [3] on single modal medical images aligns anatomical structures to the same spatial coordinate system. Therefore, radiologists will visually analyze the regions of interest (ROIs) in a unified anatomical space, making this one of the key image processing steps in clinical diagnosis and research [2]. Recently, deep learning (DL)-based deformable medical image registration [3]–[6] learns a common representation of the data distribution via convolutional neural networks (CNN) to estimates their corresponding deformation (correspondence action) in images in one inference, achieving rapid registration and expanding clinical application prospects.

However, the unsupervised DL-based deformable registration models [1], [3], [6] (Fig. 2(a)) only learn the correspondence actions for a displacement vector field (DVF) to align image pairs and have no perception ability for semantic regions (perception action), resulting in: **1)** *Mis-alignment on blurred anatomies.* The unsupervised registration CNN lacks the discriminative feature representation for different semantic regions, and is failed to perceive the soft tissues which present low contrast and blurred boundaries such as the right atrial cavity (RA) and right ventricular cavity (RV) (Fig. 1 (a)), resulting in the misalignment. **2)** *Distortion on task-unconcerned regions.* The task-unconcerned regions (Fig. 1 (b)) have significant and non-corresponding structures between moving and fixed images. The registration CNN without the perception ability seeks the alignment of all anatomies, making the ROIs have to compromise with them, thus limiting the registration accuracy on ROIs and leading to large background distortion.

Label-constrained registration models [4], [5], [7], [8] (Fig. 2(b)) take a single CNN to learn the mixed perception

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3095409, IEEE Journal of Biomedical and Health Informatics

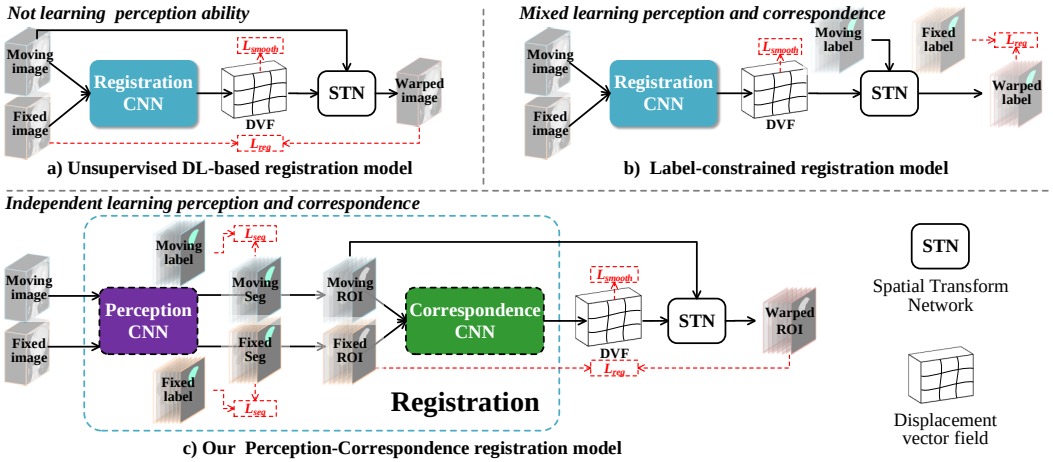

Fig. 2: The Advantages of our PC-Reg in learning process compared with the existing DL-based registration models. a) The unsupervised DL-based registration model only learns correspondence action without perception ability. b) The label-constrained registration model mixes the perception and correspondence in a single CNN being limited by the lack of texture in labels. c) Our PC-Reg decouples the registration to perception and correspondence CNNs to perceive the ROIs and align the corresponding regions independently preserving texture.

and correspondence ability for ROIs via segmentation labels, however: **1)** The lack of texture constraints in labels distort ROIs. The segmentation labels lose the texture information of the ROIs constraining the correspondence action to only focus on the alignment of the ROIs' edges and ignore the texture of ROIs. Therefore, this deformation process will distort the texture of the ROIs and make the deformed images lose authenticity. **2)** The perception performance is limited by the label amount. The high-cost labeling on medical images makes few labels available (*few-shot situation*), so that the perception action will be overfitted to the few labeled images limiting its generalization [9]. Therefore, the label-constrained models will lose the perception ability for general anatomies, and during the test, the registration model will make mis-perception and wrong correspondence weakening the registration.

We are committed to few-shot deformable medical image registration which aligns the images by DL-models only trained with few labels: **1)** To reduce the distortion internal ROIs, we decouple the perception and correspondence actions into two CNNs (perception CNN and correspondence CNN) for independent optimizations and feature representations of each action (Fig. 2(c)). Therefore, when the labels promote the accuracy of perception action, the correspondence action still will be constrained by the texture to align the whole ROIs, so that the distortion internal ROIs will be reduced. **2)** To learn the perception ability of general anatomies only with few labels, we propose the Reverse Teaching which utilizes the correspondence CNN to reversely teach the perception CNN more anatomical knowledge from unlabeled images. Therefore, as the perception CNN is improved, the correspondence CNN will also make more accurate alignment on ROIs achieving the excellent registration performance only with few labels. The details are as follow:

Perception-Correspondence Decoupling for texture preservation: We propose the *Perception-Correspondence Decoupling* to preserve the texture internal ROIs when improving the perception via labels. It decouples [10] the perception and correspondence actions to two CNNs which will learn the independent optimization and feature representation for them. Therefore, when embedding the anatomical knowledge from the labels into registration [4], [5], [7], the independent optimization of perception CNN will not interfere with the correspondence CNN, thus weakening the distortion in ROIs caused by the lack of texture in labels. The perception CNN also will perceive the ROIs and remove the background so that the correspondence CNN only needs to learn to align ROIs without the interference of labels, thus making fine registration with the preservation of texture internal ROIs. What's more, the more accurate the perception results, the correspondence will obtain more accurate anatomical structures, achieving better registration performance indirectly. Therefore, it gives us an idea to improve the registration performance via improving the perception ability.

Reverse Teaching for few-shot learning: We propose the *Reverse Teaching* to improve the generalization of perception CNN only with few labels to improve the perception of ROIs, thus finally improving the registration performance. It utilizes the correspondence CNN to generate additional training data using diverse images and few labels, thus reversely teaching the perception CNN more anatomical structure and style knowledge: **1)**Warped image and warped label pairs with structure knowledge. The labeled images are aligned to unlabeled images by the correspondence CNN together with their labels for warped images and labels, so that the warped image and label pairs will have the structure information of unlabeled images. Therefore, this additional structure information will be used to teach the perception CNN more anatomical structures knowledge. **2)** Warped image and fixed label pairs with style knowledge. The unlabeled images are aligned to the labeled images by the correspondence CNN for warped images, so that these warped images with the style information from the unlabeled images will be labeled by the fixed labels for warped image and fixed label pairs. Therefore, this additional style information will be used to teach the perception CNN more anatomical style knowledge.

We propose the *Perception-Correspondence Registration*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3095409, IEEE Journal of Biomedical and Health Informatics

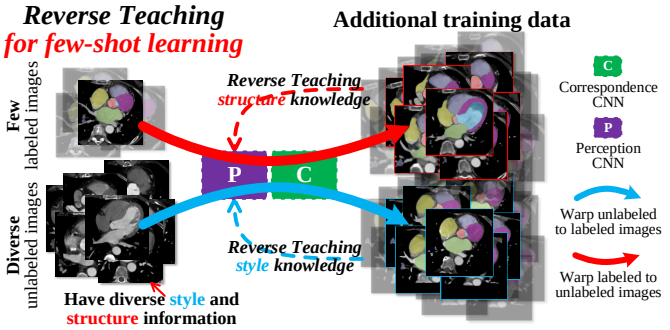

Fig. 3: For few-shot learning, our Reverse Teaching generates additional training data via aligning few labeled images and diverse unlabeled images with each other thus teaching the perception CNN rich structure and style knowledge.

(PC-Reg) framework for few-shot deformable medical image registration for the first time. Our contributions are as follow:

- To the best of our knowledge, we propose a novel few-shot learning framework, *Perception-Correspondence Registration (PC-Reg)*, for few-shot deformable medical image registration for the first time. It only needs few labels, greatly improving the registration accuracy on ROIs and reducing the distortion of deformation.
- We propose the *Perception-Correspondence Decoupling* which preserves the texture internal ROIs when improving the perception via labels. It decouples the perception and correspondence actions to two CNNs for the independent optimizations and feature representations. Therefore, the lack of the texture in labels will not interfere with the correspondence CNN, so that when the perception CNN learns to perceive more accurate ROIs, the correspondence CNN still will preserve the texture internal ROIs.
- We propose the *Reverse Teaching* which improves the generalization of perception action with only few labels. It utilizes the correspondence CNN to generate additional training data using diverse images and few labels, thus reversely teaching the perception CNN more anatomical knowledge. Therefore, the perception CNN with more powerful generalization ability will provide accurate ROIs for correspondence CNN, finally improving the registration performance.

The rest of the paper is organized as follows. We review the related works about deformable medical image registration, perception and correspondence in computer vision, and few-shot learning in medical image analysis in Sec. II. Then, we specifically introduce our proposed PC-Reg in Sec. III, including our Perception-Correspondence Decoupling (Sec. III-A), our Reverse Teaching (Sec. III-B), and the detailed network structures (Sec. III-C). Then the datasets, comparison methods, implementation, and evaluation metrics are described in Sec. IV. The results are shown and analyzed in Sec. V. Finally, our work is well concluded in Sec. VI.

II. Related work

1) Deformable medical image registration: Deformable medical image registration [2], [3] aligns anatomical structures in medical images to the same spatial coordinate system. The traditional methods [2], such as BSpline [11] and Symmetric Normalization (SyN) [12], are optimized by intensive iterative computation via the intensity similarity in image pair, resulting in poor time efficiency [4]. Taking advantage of GPUs and powerful representation ability, the DL-based methods [3]–[5] have been applied to the deformable medical image registration by predicting a DVF directly, achieving great time efficiency, and competitive accuracy.

Our work focus on DL-based registration which has three types: **1)** Supervised registrations [13], [14]. The real DVFs, which are obtained by simulation and deformed image pairs or from classical registration methods, are used to train a network for the estimation ability of deformation. However, the accuracy of registration is limited by the bias of ground truth generated from the classical registration methods. **2)** Unsupervised registrations [1], [15]–[19]. Inspired by the spatial-transformer [20], [21], image interpolation is integrated into the network to make the registration an end-to-end trainable process [1], [5]. Therefore, the intensity similarity such as mean square distance [1], normalized cross-correlation [1], etc. is used to optimize the network directly to learn the deformation without any labels. While this unsupervised strategy lacks ROIs perception ability and seeks the alignment of all anatomies, so that it is easily disturbed by outliers and task-unconcerned regions limiting the performance on ROIs. **3)** Weakly supervised registrations [5], [7], [8]. These methods take auxiliary information like segmentation maps and landmarks to guide the network to perceive the ROIs, so that weaken the damaging deformation of backgrounds. However, the auxiliary information has to be fine labeled which is difficult and time-consuming for 3D medical images, making the overfitting only with few labels. Besides, the DVF will be distorted due to the lack of real texture constraints in labels [4], [6], decreasing the clinical availability and value.

2) Perception and correspondence in computer vision: Learning perception and correspondence are fundamental problems in computer vision. The object detection task [22] is the perception of the object's position and size in images, the instance segmentation [23] and semantic segmentation [24]–[26] tasks are the pixel level perception for instance objects and semantic objects, whereas the image recognition task [27] is also an image-level perception for objects in the images. For correspondence problems, the optical flow estimation [28] and unsupervised deformable registration [1], [16], [17], [29] are the pixel-level correspondence tasks for the pixels in image pairs, and the object tracking [30] and matching [31] tasks are the patch-level correspondence tasks for the objects in patches.

Our few-shot deformable medical image registration task is a problem that combines the learning of perception and correspondence. The few segmentation labels guide the model to learn to perceive the ROIs, and the model also learns the correspondence of these ROIs for registration at the same time. Due to the limitation of the labels (Sec. I), we decouple the perception and correspondence actions to two models which are mutually complementary, and draw on the experience of the above researches to guide the design of our framework.

3) Few-shot learning in medical image analysis: Few-shot

learning learns models from the data with little supervision information [9], relieving the burden of collecting large-scale labeled data. In medical image analysis, few-shot learning is urgent and has achieved success in many scenarios such as histopathological image classification [32], brain imaging modality recognition [33], segmentation [25], etc. They aim to reduce the label amount of medical images whose labeling works are professional and costly, reducing the research cost.

However, there are no efforts for few-shot deformable medical image registration to reduce the labeling cost and improve the perception accuracy in label-constrained registration. Our Reverse Teaching utilizes the alignment ability of correspondence CNN to generate additional supervision information using diverse images and few labels, thus reversely teaching the perception CNN more anatomical style and structure knowledge and improving its generalization ability.

III. Methodology

As shown in Fig. 2 and 4, our PC-Reg only needs few labels achieving few-shot deformable medical image registration. It works in two aspects: **1)** The **Perception-Correspondence Decoupling** (Sec. III-A) decouples the perception and correspondence actions of the registration into two CNNs, so that their independent optimization will isolate the interference of the label on the correspondence action, thus simultaneously weakening the distortion in ROIs caused by the lack of texture in labels and perceiving the ROIs with clear boundaries; **2)** The **Reverse Teaching** (Sec. III-B) aligns the labeled and unlabeled images to map the supervision information in the labels to the unlabeled images, thus generating the additional training data to reversely teach the perception CNN more anatomical style and structure knowledge making the more accurate perception of ROIs. The details of the perception and correspondence CNNs will be illustrated in Sec. III-C.

A. Perception-Correspondence Decoupling for real texture preservation

As shown in Fig. 2, our PC-Reg decouples the registration $R(\cdot)$ to perception CNN $P(\cdot)$ and correspondence CNN $C(\cdot)$ to make independent optimizations and feature representations. Therefore the registration framework will learn to perceive the ROIs with clear boundaries and further make accurate correspondence without the distortion of real texture.

1) Perception for ROIs with clear boundaries: The decoupled perception CNN provides the ROIs which have clear boundaries to correspondence CNN for accurate correspondence. As shown in Fig. 4, the moving x_m and the fixed x_f images are put into the perception CNN $P(\cdot)$ for the moving y'_m and fixed y'_f segmentations in N semantic channels without textures. These segmentations excluding the background channel are multiplied with their original images x for moving r_m and fixed ROIs r_f in $N-1$ semantic channels with real textures, thus eliminating the task-unconcerned backgrounds and perceiving the ROIs which have clear boundaries. The cross-entropy loss function $\mathcal{L}_{ce}$ are calculated on fixed and moving segmentations with their labels ($\mathcal{L}_{fce} = \mathcal{L}_{ce}(y'_f, y_f)$, $\mathcal{L}_{mce} = \mathcal{L}_{ce}(y'_m, y_m)$) to optimize the perception CNN:

$$\mathcal{L}_{ce}(y', y) = -\frac{1}{NK}\sum_{n=0}^{N}\sum_{k=0}^{K} y_{k,n}\log y'_{k,n}, \tag{1}$$

here, N is the number of the semantic channels from the perception CNN, the K is the number of the voxels in each semantic channel, the y is the pixel-wise label and the y' is the segmentation.

2) Correspondence for texture-preserved registration: The decoupled correspondence CNN aligns the perceived anatomies accurately with texture-preserved ability. The perceived moving and fixed ROIs are concatenated and put into the correspondence CNN $C(\cdot)$ for a DVF ϕ which will have fine deformation on ROIs without the distortion on backgrounds. The DVF deforms the moving ROIs, moving labels and moving images via the spatial transform network (STN) [20] $T(\cdot)$ for warped ROIs r_w, warped labels y_w and warped images x_w. The warped ROIs with real textures will be used to calculate the local normalized cross-correlation loss [1] $\mathcal{L}_{LNCC}$ with the fixed ROIs for fine deformation with texture-preserving due to the constraint of the texture internal ROIs, and a smooth regularization loss [1] $\mathcal{L}_r$ is calculated to penalize local spatial variations in the DVF:

$$\mathcal{L}_{corr} = \sum_{n=0}^{N-1}\mathcal{L}_{LNCC}(r_{w_n}, r_{f_n}) + \lambda_0\mathcal{L}_r(\phi) \tag{2}$$

Here, the λ_0 is the weight to balance two loss functions. The inference of our decoupled registration process is $x_w = T(x_m, R(x_f, x_m)) \rightarrow T(x_m, C(P(x_f)_{\{1,\dots,N\}} * x_f, P(x_m)_{\{1,\dots,N\}} * x_m)))$, where the $\rightarrow$ means decoupling.

We takes two-stage training strategy to avoid the potential interference of poor perception in the beginning. In the first stage, the perception CNN is trained independently and the correspondence CNN is only train on the accurate ROIs from the few labeled images without inaccurate input, achieving an initial perception and correspondence ability. Therefore, in the second stage which train with unlabeled images, the inaccurate segmentation are reduced from perception CNN, avoiding the very poor segmentation owing to the random initialization, and stabilizing the training process.

B. Reverse Teaching for few-shot learning

Our Reverse Teaching aligns the few labeled images and the diverse unlabeled images, thus mapping the supervision information in the labels to the unlabeled images (Fig. 3) for the additional training data *(warped image and label pairs, warped image and fixed label pairs)*. Therefore, as shown in Fig. 4, these data are put into the perception CNN for segmented warped images, thus reversely teaching the perception CNN rich structure and style knowledge in the unlabeled images for more accurate perception.

1) Warp labeled images for structure knowledge: Our Reverse Teaching warps the few labeled images to the diverse unlabeled images teaching the perception CNN rich structure knowledge. The unlabeled images have rich structure knowledge, the labeled images are warped to the unlabeled images

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3095409, IEEE Journal of Biomedical and Health Informatics

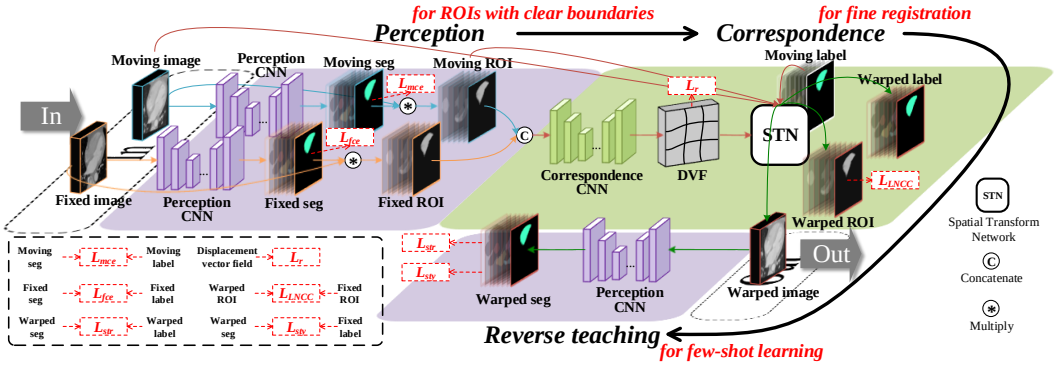

Fig. 4: The details of our proposed PC-Reg. We decouple the perception and correspondence actions in the registration process to two CNNs, thus embedding anatomical knowledge of ROIs for accurate perception and weakening the distortion caused by the lack of texture in labels. The correspondence CNN also performs the Reverse Teaching which teaches the perception CNN rich structure and style knowledge from unlabeled images, perceiving more accurate ROIs.

by our PC-Reg. Therefore, the warped labeled images will have the anatomical structure information of the unlabeled images, thus generating diverse warped image and label pairs to teach our perception CNN diverse structure knowledge. As shown in Fig. 4, the warped images x_w from correspondence CNN are putted into the perception CNN for warped segmentations y'_w. The warped segmentations are calculated the cross-entropy loss (Equ. 1) with the warped labels y_w for structure loss $\mathcal{L}_{str} = \mathcal{L}_{ce}(y'_w, y_w)$, so that perception CNN will generalize to anatomical structures.

2) Warp unlabeled images for style knowledge: Our Reverse Teaching warps the diverse unlabeled images to the few labeled images teaching the perception CNN rich style knowledge. The unlabeled images have rich style knowledge, and are warped to the labeled images by our CP-Reg. Therefore, the warped unlabeled images with the style information of the unlabeled images will be labeled by the fixed labels, thus generating diverse warped image and fixed label pairs to teach our perception CNN diverse style knowledge. As shown in Fig. 4, the warped images x_w from correspondence CNN are putted into the perception CNN for warped segmentations y'_w. The warped segmentations are calculated the cross-entropy loss (Equ. 1) with the fixed labels y_f for style loss $\mathcal{L}_{sty} = \mathcal{L}_{ce}(y'_w, y_f)$, so that perception CNN will generalize to more image styles.

In general, when it comes different combinations of labeled and unlabeled images, the perception CNN is trained with four different losses $\mathcal{L}_{perc}$: **1)** fixed y_f and moving y_m labels are available; **2)** fixed labels y_f are available; **3)** moving labels y_m are available; and **4)** no label is available:

$$\mathcal{L}_{perc} = \begin{cases} \mathcal{L}_{fce} + \mathcal{L}_{mce} + \mathcal{L}_{str} + \lambda_1 \mathcal{L}_{sty} & \text{if } y_f \text{ and } y_m \\ \mathcal{L}_{mce} + \mathcal{L}_{str} & \text{if } y_m \\ \mathcal{L}_{fce} + \lambda_1 \mathcal{L}_{sty} & \text{if } y_f \\ 0 & \text{if no label} \end{cases} \tag{3}$$

where the λ_1 is the weight to weaken the interference of misaligned structures warped image and fixed label pairs.

We take two-stage training strategy to perform our Reverse Teaching in the second stage, thus weakening the influence of mis-alignment at the beginning of training. At the beginning of the training process, our PC-Reg was only trained by the few labeled images, so that the perception and correspondence ability of the CNNs are weak which will make the distortion and too many misaligned structures in the additional training data. If this data is used for our Reverse Teaching, the training process will be interfered. Therefore, our Reverse Teaching participates in the training process in the later stage to generalize the perception to more images, thus further improving the registration accuracy.

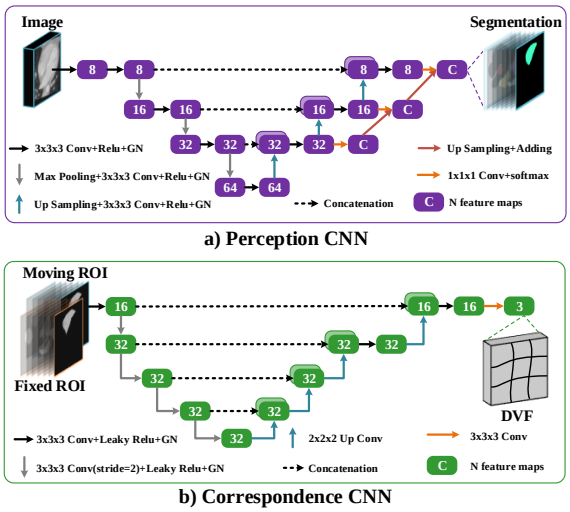

Fig. 5: The detailed network structures of our perception CNN and correspondence CNN.

C. Details of the networks

1) Perception CNN: Our perception CNN segments the mask of the ROIs on fixed and moving images. As shown in Fig. 5(a), it follows U-Net [34] structure consisting of the encoder and decoder, and using the deep supervision [35] to enhances the accuracy. The encoder down-samples the input images saving GPU memory and expanding the receptive field, the decoder restores the low-resolution feature maps for the output segmentations with original resolution. It has four resolution stages with two $3\times3\times3$ convolution layers followed

by a group normalization [36] (GN) and a rectified linear units (ReLU). Maxpooling is used to down-sample and the trilinear interpolation is used to up-sample. The concatenation in the same resolution adds detailed information to the decoding. Each resolution stage in decoder outputs the segmentation via $1 \times 1 \times 1$ convolutions for deep supervision.

2) Correspondence CNN: Our correspondence CNN takes moving and fixed ROIs and outputs the DVF for fine deformation on ROIs. As shown in Fig. 5(b), it follows the 3D U-Net [34] with five resolution stages. In the encoder, each resolution stage has one $3 \times 3 \times 3$ convolution layer, and maxpooling is used to down-sample the feature maps between each stage. In the decoder, the first two stages have one $3 \times 3 \times 3$ convolution and the last two stages has two $3 \times 3 \times 3$ convolution for finer estimation in details. Each convolution is followed by a GN [36] and a ReLU. Finally, an $1 \times 1 \times 1$ convolution is used to estimate a DVF with three channels for the deformation of each voxel in the x, y, z-direction.

IV. Experiments Configurations

1) Datasets: We evaluate our excellent registration performance of our PC-Reg on three different organs including cardiac, cervical vertebra and brain:

Cardiac CT cross-object registration: This dataset is from the MM-WHS 2017 Challenge [37] including 20 labeled CT images with and 40 unlabeled CT images. These labels segment seven clinical ROIs on the heart including the ascending aorta (AO), left atrial cavity (LA), left ventricular cavity (LV), myocardium of the left ventricle (Myo), pulmonary artery (PA), right atrial cavity (RA) and right ventricular cavity (RV). The cardiac regions are cropped and contained in a rectangular box for affine alignment firstly. Then they were resampled to 128×128×96 for fitting the input size of the network. Five labeled images are selected randomly together with the 40 unlabeled images as a training set making our few-shot situation. The remaining 15 labeled images produce 210 image pairs as the test set.

Cervical vertebra CT cross-object registration: The CT images of 43 patients are involved in this dataset. The segmentation of seven cervical vertebras (C1, C2, C3, C4, C5, C6, C7) is labeled by two radiologists with cross-check. Then these images are resampled to 128×128×128 for fitting the input size of the network. 27 images are randomly selected as the training set and the remaining 16 images are used for the testing set. In the training set, we select 5 images with labels and 22 images without labels randomly making our few-shot situation. The remaining 16 labeled images are pairwise coupled making up the test set.

Brain MR cross-object registration: This dataset is from LBPA40 [38] which has 40 brain MR images with 56 labeled regions and we selects 4 regions (B1, B2, B3, B4) as our ROIs. These images are cropped and resampled to 128×144×112 to fit the input size of our network. 30 images are randomly selected as the training set and the remaining 10 images are used for the testing set. In the training set, we select 5 images with labels and 25 images without labels randomly making a few-shot situation. The remaining 10 labeled images produce 90 image pairs as the test set.

Before being fed into the network, each image pair has an initialization via affine transformation[1] following [1] for global alignment, so that our PC-Reg will concentrate on the deformation registration.

2) Comparison settings: To demonstrate the superiority of our proposed method, we compared our proposed PC-Reg with five widely-used deformable registration methods, including two traditional methods (BSpline [11], SyN [12]) and four DL-based methods (unsupervised (Unsup-) VoxelMorph [1], label-constrained (LC-) VoxelMorph [4], CycleMorph [19] and DeepRS [8]). The BSpline implemented by Elastix[1] with its default parameters [11], and the SyN [12] is implemented in the publicly available Advanced Normalization Tools (ANTs) [39] with a mutual information similarity metric. The Unsup-VoxelMorph [1] uses LNCC loss function together with the smooth loss. The LC-VoxelMorph [4] uses Dice loss together with the smooth loss in our few-shot situation, and we test different weights ($\lambda = 1$ and $\lambda = 0.1$) for smooth loss to evaluate the distortion of DVF due to the lack of real texture constraints in the labels. All methods are based on an initialization via affine transformation[1].

3) Implementation details: Our PC-Reg is implemented by Tensorflow[2] on a single NVIDIA TitanX GPU with 12 GB memory. We set $\lambda_0 = 1$ for our PC-Reg to achieve the accurate registration and the smooth DVF at the same time following the setting of unsupervised VoxelMorph [1] in our experiments. The λ_1 is 0.5, so that the error supervision in misaligned structures will be weakened when the labels are used as pseudo-labels for the aligned unlabeled images in our Reverse Teaching. All DL-based models are optimized by Adam whose learning rate is 1×10^{-4}. The training batch size is 1 to save the memory and iterated 400 epochs (200 for PC-Reg and 200 for PC-Reg with Reverse Teaching). To improve the generalization ability, the random mirror in x, y and z-axis, and rotation in $[-20°, 20°]$ are conducted for data augmentation. Source code is released at https://github.com/YutingHe-list/PC-Reg-RT.

4) Evaluation metrics: The Dice similarity coefficient (DSC) [%] and average surface distances (ASD) [26] between warped label $y_w = T(y_m, \phi)$ and fixed label y_f are calculated to assess the registration results. The higher of DSC and the lower of ASD mean the better accuracy of registration. To evaluate our texture-preserving ability, Jacobian matrix $J_\phi(p) = \nabla\phi(p) \in R^{3\times3}$ is calculated to capture the local properties of ϕ around voxel p [4]. For those voxels of $J_\phi(p) \leq 0$ are recorded as singularities that indicate folds. We calculated the fraction of $J_\phi(p) \leq 0$ [%] for each DL-based method to quantitatively measure the texture-preserving ability of DVF, the small this metric, the better the texture-preserving ability of deformation [17]. The GPU times and CPU times of these models are counted to analyze their time efficiency. The DSC [%] between the segmentation y' and their labels y are calculated to analyze the perception accuracy of our perception CNN. The standard deviation (std) of these metrics are also provided to evaluate the stability of these models.

[1]https://www.elastix.org/

[2]https://github.com/tensorflow/tensorflow

V. Results and Analysis

Our framework eliminates the interference of backgrounds, perceives the ROIs with clear boundaries, preserves the real texture and improves the generalization ability in the few-shot situation, thus achieving competitive registration on ROIs.

A. Quantitative evaluation demonstrates our advantages

As illustrated in Tab. I, our PC-Reg with Reverse Teaching (PC-Reg-RT) achieves the competitive registration accuracy on three registration tasks with small $|J_\phi| \leq 0$ bringing effective texture-preserving ability. Besides, taking the advantage of one inference, our time efficiency is comparable to other DL-based models ($< 1s$) which is much faster than traditional models.

1) Our competitive registration accuracy: As shown in Tab. I, our PC-Reg achieves the competitive registration accuracy on cardiac CT, cervical vertebra CT and brain MR cross-object registration tasks only with five labels. Our Perception-Correspondence Decoupling (PC-Reg) takes the perception CNN to eliminate the task-unconcerned backgrounds and perceives ROIs which have clear boundaries independently to correspondence CNN, thus achieving 79.0% Reg-DSC and 1.93 ASD on cardiac CT and 81.4% Reg-DSC and 0.66 ASD on cervical vertebra CT. Compared with the LC-VoxelMorph($\lambda = 1$), it achieves 5.8% (a) and 1.0% (b) Reg-DSC improvement, and 0.5 (a), 0.06 (b), and 0.04(c) ASD reducing owing to our decoupling which will provide independent feature representations for two actions and isolate the interference of labels. When our Reverse Teaching (PC-Reg-RT) performs, it achieves 6.7% (a), 5.3% (b), 1.0% (c) Reg-DSC improvement and 0.61 (a), 0.25 (b) and 0.6 (c) ASD reducing on cardiac (a), cervical vertebra (b) and brain (c) registration, because the rich structure and style information in unlabeled data effectively improves the perception CNN (6.3% (a), 20.6% (b) and 2.9% (c) Seg-DSC improvements).

2) Our effective texture-preserving ability: Our PC-Reg eliminates the distortion on task-unconcerned backgrounds and avoids the distortion in ROIs caused by the lack of real texture in labels, making the registration only need to focus on the ROIs and achieving texture-preserving ability, especially in the cervical vertebra CT whose backgrounds are significant and varied. Both in cardiac CT, cervical vertebra CT and brain MR, our PC-Reg and PC-Reg-RT have a stable small fraction ($< 1\%$) $|J_\phi| \leq 0$ and bringing texture-preserving deformation. Unsup-VoxelMorph is impacted by these backgrounds with many non-corresponding structures thus having large fractions (4.48% (a), 10.96% (b), 1.37% (c)). Especially, in cervical vertebra CT, the SyN and Unsup-VoxelMorph seek to achieve global alignment which will pay attention to the backgrounds, bringing severe distortion and worse results (39.4% and 50.1% Reg-DSC (b)) which are even worse than Affine only (64.8% Reg-DSC (b)). The ASD and $|J_\phi| \leq 0$ of SyN cannot be counted due to such exaggerated deformation. When directly adding the labels, the LC-VoxelMorph($\lambda = 1$) achieves a more smooth DFV (0.38% (a)) on cardiac CT due to the attention on ROIs, but its Reg-DSC is reduced to 73.2% (a). When the weight of the smooth loss is reduced to 0.1, the Reg-DSC of LC-VoxelMorph($\lambda = 0.1$) is improved to 77.0% (a) and 82.3% (b) owing to the stronger constraints on the boundaries of ROIs. But its $|J_\phi| \leq 0$ increases to 3.43% (a) and 1.85% (b) because the lack of real texture constraints in the labels causes the distortion of DVF. DeepRS is seriously interfered by the labels and non-corresponding structures in background, thus achieving terrible distortion with high $|J_\phi| \leq 0$ on three tasks. Our PC-Reg-RT avoids the damaging deformation in backgrounds and the distorted deformation in ROIs, achieving small $|J_\phi| \leq 0$ (0.37% (a) and 0.11% (b)) and high Reg-DSC (85.7% (a) and 86.7% (b)). Therefore, compared with Unsup-VoxelMorph and LC-VoxelMorph($\lambda = 0.1$), our PC-Reg-RT has the better texture-preserving ability with high accuracy.

3) Our great time efficiency: Taking the advantage of one inference, our time efficiency is comparable to other DL-based models which is much faster than traditional models. The traditional B-Spline and SyN have to take more than 10s on each task which cannot be applied to some scenarios with real-time requirements.

4) Details of structures: Our PC-Reg-RT achieves competitive registration performance on the cardiac CT, cervical vertebra CT and brain MR (Fig. 6). In the cervical vertebra CT, the results of Unsup-VoxelMorph and SyN are deformed exaggeratedly and distortedly due to the large damaging interference of the backgrounds, so that they even reduced the alignment of the ROIs and get large performance fluctuations.

B. Qualitative evaluation shows our visual superiority

The DVFs and the registration results in Fig. 7 demonstrate that our PC-Reg has fine registration on boundaries and few fold in ROIs, and eliminates the distortion of backgrounds on three datasets owing to our decoupling and Reverse Teaching.

1) Fine registration on boundaries: As shown in cardiac case 2 and cervical vertebra case 1 in Fig. 7, our PC-Reg-RT achieves fine registration on boundaries owing to the decoupled perception action that provides the ROIs with clear boundaries for the correspondence action. Due to the lack of anatomical priori knowledge of ROIs for perception, the Unsup-VoxelMorph, SyN and BSpline perform rough registration even worse than affine only, especially in the cervical vertebra CT (b) whose structures are small and have large variations. When adding the labels, the LC-VoxelMorph($\lambda = 0.1$), LC-VoxelMorph($\lambda = 1$) and DeepRS get ROIs perception ability and achieve the alignment of ROIs, but they lose the fineness of boundaries due to mixed perception and correspondence actions complicating the feature representation.

2) Eliminate the distortion of backgrounds: Our PC-Reg-RT perceives the ROIs via perception CNN thus eliminating the distortion outside the ROIs (background). As shown in case 2 of the cervical vertebra in Fig. 7, our PC-Reg and PC-Reg-RT focus on the ROIs, so that the significant and non-corresponding structures of the task-unconcerned anatomical structures in the background will not interfere with the correspondence process. Therefore, the independent deformation on ROIs eliminates the distortion of backgrounds. The SyN, BSpline and Unsup-VoxelMorph have no perception ability and are interfered by the non-corresponding structures of backgrounds bringing exaggerated deformation. Al-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3095409, IEEE Journal of Biomedical and Health Informatics

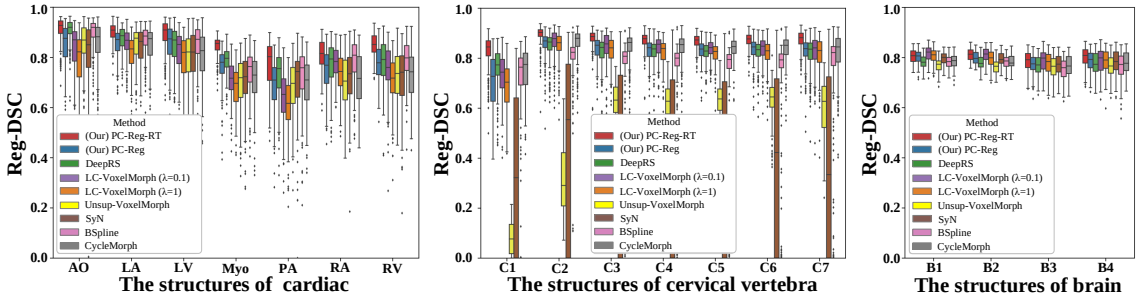

Fig. 6: The box-plots present the Reg-DSC of the cardiac, cervical vertebra, and brain structures for different registration methods. Our PC-Reg-RT achieves the competitive registration accuracy of each structure.

TABLE I: The quantitative evaluation demonstrates the advantages of our framework on our three tasks. Our PC-Reg-RT achieves state-of-the-art performance on registration (Reg)-DSC and ASD. Our time efficiency is comparable to other DL-based models ($< 1s$) which is much faster than traditional models. Our small fraction ($< 1\%$) $|J_\phi| \leq 0$ illustrates our effective texture-preserving ability. The improvement of segmentation (Seg)-DSC after equipping our Reverse Teaching (-RT) shows its great generalization improvement in few-shot situation.

Method	Reg-DSC (%)	ASD	$\lvert J_\phi\rvert \leq 0$ (%)	CPU time (s)	GPU time (s)	Seg-DSC (%)
a) Cardiac CT cross-object registration						
Affine only	64.0±12.5	3.37±0.86	-	5.98±0.55	-	-
BSpline [11]	80.8±10.4	1.69±0.63	**0.34±0.51**	40.19±1.59	-	-
SyN [12]	75.5±12.7	2.31±0.90	0.50±0.16	23.70±4.33	-	-
Unsup-VoxelMorph [1]	75.8±11.8	2.18±0.74	4.48±1.61	-	0.22±0.16	-
LC-VoxelMorph($\lambda = 1$) [4]	73.2±11.6	2.43±0.68	0.38±0.23	-	0.23±0.41	-
LC-VoxelMorph($\lambda = 0.1$) [4]	77.0±11.6	2.04±0.58	3.43±0.79	-	0.23±0.31	-
CycleMorph [19]	76.5±9.4	2.12±1.01	0.64±0.18	-	0.22±0.25	-
DeepRS [8]	81.5±7.2	1.71±0.77	7.03±1.18	-	0.65±0.10	87.4±6.4
(Our) PC-Reg	79.0±9.9	1.93±0.56	0.40±0.18	-	0.54±0.51	83.1±12.9
(Our) PC-Reg-RT	**85.7±7.3**	**1.32±0.38**	0.37±0.28	-	0.54±0.19	**89.4±6.1**
b) Cervical vertebra CT cross-object registration						
Affine only	64.8±10.2	1.37±0.34	-	6.35±0.70	-	-
BSpline [11]	74.2±18.5	1.15±1.58	0.45±0.98	38.62±1.72	-	-
SyN [12]	39.4±34.7	-	-	21.30±9.70	-	-
Unsup-VoxelMorph [1]	50.1±22.1	3.09±0.58	10.96±0.41	-	0.29±0.17	-
LC-VoxelMorph($\lambda = 1$) [4]	80.4±8.4	0.72±0.25	0.25±0.09	-	0.29±0.16	-
LC-VoxelMorph($\lambda = 0.1$) [4]	82.3±7.6	0.65±0.22	1.85±0.42	-	0.29±0.17	-
CycleMorph [19]	82.5±6.8	0.62±0.35	0.13±0.06	-	0.34±0.38	-
DeepRS [8]	81.7±5.7	0.65±0.31	2.06±0.39	-	0.86±0.13	**86.3±8.6**
(Our) PC-Reg	81.4±8.1	0.66±0.23	0.16±0.08	-	0.74±0.60	63.8±20.4
(Our) PC-Reg-RT	**86.7±5.0**	**0.41±0.15**	**0.11±0.06**	-	0.71±0.23	84.4±12.6
c) Brain MR cross-object registration						
Affine only	75.5±3.7	1.25±0.21	-	7.14±0.51	-	-
BSpline [11]	77.0±3.9	1.15±0.22	**0**	40.32±0.62	-	-
SyN [12]	78.5±3.8	1.07±0.21	**0**	19.67±1.46	-	-
Unsup-VoxelMorph [1]	76.5±3.7	1.09±0.19	1.37±0.19	-	0.30±0.30	-
LC-VoxelMorph($\lambda = 1$) [4]	79.0±4.1	1.07±0.22	0.14±0.03	-	0.30±0.29	-
LC-VoxelMorph($\lambda = 0.1$) [4]	79.9±3.9	1.02±0.20	1.37±0.15	-	0.30±0.28	-
CycleMorph [19]	77.7±3.7	1.06±0.20	**0**	-	0.32±0.42	-
DeepRS [8]	77.6±3.6	1.05±0.19	1.34±0.24	-	1.04±0.12	81.7±3.4
(Our) PC-Reg	79.0±3.4	1.03±0.18	0.02±0.01	-	0.71±0.39	79.4±3.5
(Our) PC-Reg-RT	**80.0±3.4**	**0.97±0.18**	0.04±0.02	-	0.71±0.40	**82.3±3.3**

though the labels embed the perception ability into the LC-VoxelMorph($\lambda = 1$), LC-VoxelMorph($\lambda = 0.1$) and DeepRS making the registration pay more attention to ROIs, they also make the distortion on the backgrounds due to their mixed feature representation for perception and correspondence.

3) Few folds in ROIs: As shown in case 1 of cardiac in Fig. 7, our PC-Reg takes the ROIs of fixed and moving images to train the correspondence CNN for DVF, so that the real texture constraints in the ROIs will avoid the distortion of DVF and make the great texture-preserving ability of deformation. While the LC-VoxelMorph($\lambda = 1$) has serious distortion in ROIs decreasing the clinical availability, because it directly takes the labels which have no real texture information. When its λ of smooth loss decreases to 0.1, it shows more serious distortion due to the stronger constraint of labels. The unsupervised methods (Unsup-VoxelMorph, SyN and BSpline) utilizes the textures and structures information as the constraints to avoid the exaggerated distortion in the ROIs, but the large deformation on backgrounds makes the global distortion.

C. High Perception accuracy with Few Supervision

Only with five labels, our Reverse Teaching generates additional training data from diverse unlabeled images, thus teaching the perception CNN rich structure and style knowledge and improving its generalization for the perception of more accurate ROIs (89.4% on cardiac CT, 84.4% on cervical vertebra CT, and 82.3% on brain MR). It improves perception CNN achieving 6.3%, 20.6% and 2.9% Seg-DSC

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3095409, IEEE Journal of Biomedical and Health Informatics

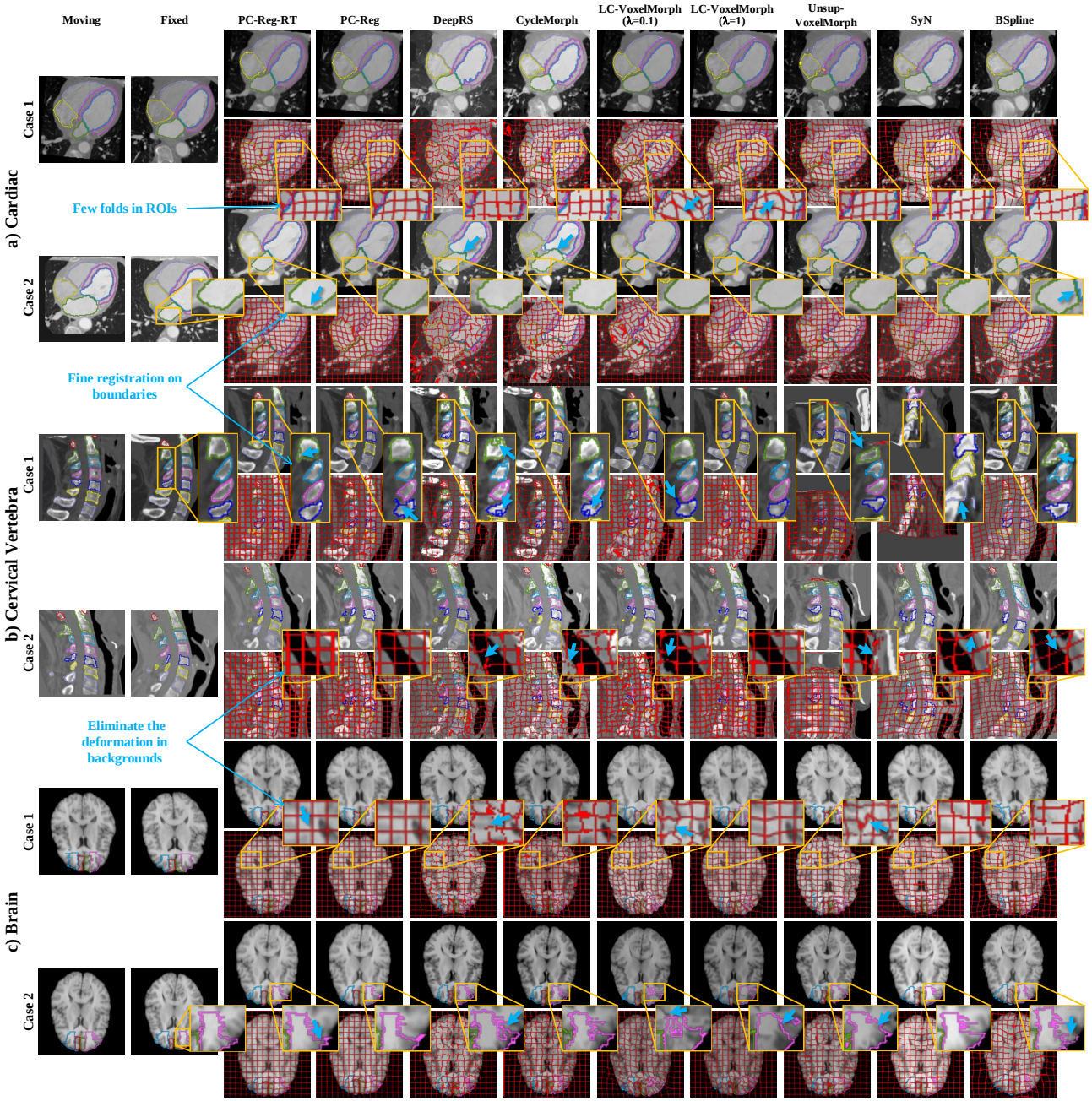

Fig. 7: The qualitative evaluation shows the visual superiority of our PC-Reg on our three registration tasks. Our Perception-Correspondence Decoupling eliminates the damaging deformation of backgrounds, perceives the ROIs with clear boundaries, and reduce the distortion of DVF, thus achieving fine registration on ROIs and great texture-preserving ability.

improvement on our three tasks (Tab. I). As shown in Fig. 8, the segmentation results are significantly improved visually, reflected in fine boundaries and few mis-segmentations. The enlarged regions in cardiac CT show the complete segmented regions and smooth boundaries. Our PC-Reg-RT makes clear distinctions between different cones in cervical vertebra CT, effectively suppressing the mis-segmentation between cones caused by overfitting (PC-Reg).

D. Ablation study

1) Our effectiveness in fewer-shot situation: As illustrated in Fig. 9, when the label amount decreases, our PC-Reg-RT still will achieve fine registration performance on Cardiac CT. **a)** Under very fewer labels, our Reverse Teaching brings significant improvement and our PC-Reg-RT still has excellent performance. The PC-Reg only has 64.1% Reg-DSC and 15.4% Seg-DSC with one label, while our Reverse Teaching improves the generalization ability and makes our PC-Reg-RT achieve 15.0% Reg-DSC and 65.8% Seg-DSC improvements. Such significant improvements are owing to the embedding of rich structure and style knowledge from unlabeled images. **b)** With the increasing of the label amount, our PC-Reg-RT will be further enhanced. When the label amount is increased to five, our PC-Reg-RT achieves 6.6% Reg-DSC and 8.2%

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2021.3095409, IEEE Journal of Biomedical and Health Informatics

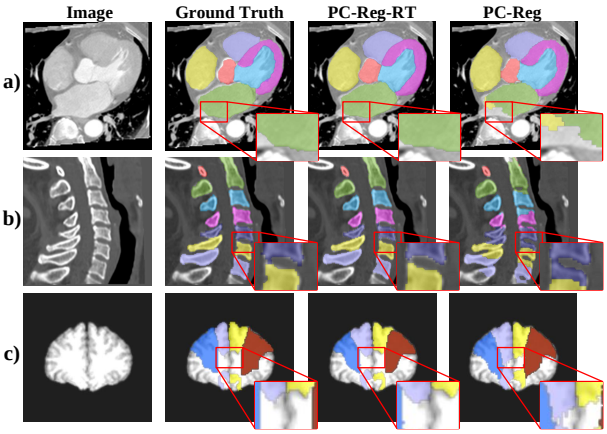

Fig. 8: Our Reverse Teaching significantly improves the generalization of perception CNN and perceives more accurate ROIs only with five labels. a) cardiac CT, b) cervical vertebra CT, c) brain MR.

TABLE II: The ablation study analyses the contributions of the additional structure and style information provided by our correspondence CNN in our Reverse Teaching on Cardiac CT.

$\mathcal{L}_{str}$	$\mathcal{L}_{sty}$	Reg-DSC(%)	Seg-DSC(%)
		79.0±9.9	83.1±12.9
✓		82.1±9.2	84.1±14.0
	✓	85.4±7.3	89.1±6.1
✓	✓	85.7±7.3	89.4±6.1

Seg-DSC improvements owing to the good performance at the beginning of training (PC-Reg).

2) Contributions of structure and style knowledge from unlabeled images: As shown in Tab. II, the ablation study of style loss $\mathcal{L}_{sty}$ and structure loss $\mathcal{L}_{str}$ shows the contribution of rich structure and style knowledge from unlabeled images which improves our accuracy for few-shot learning. The structure information brings 3.1% Reg-DSC and 1.0% Seg-DSC improvements to our PC-Reg-RT because the structure information in diverse unlabeled images is mapped to our few labeled images providing the perception CNN diverse structure knowledge. The rich style information in unlabeled images is deformed to the labels guiding the perception CNN to learn rich style knowledge, thus getting 6.4% Reg-DSC and 6.0% Seg-DSC improvements. When using all additional information, our PC-Reg-RT achieves astonishing 6.7% Reg-DSC and 6.3% Seg-DSC improvements.

VI. Discussion and Conclusion

In this paper, we propose the *Perception-Correspondence Decoupling Registration (PC-Reg)* for few-shot deformable medical image registration, greatly improving the registration accuracy on ROIs with texture-preserving. Our Perception-Correspondence Decoupling decouples the perception and correspondence actions of registration into two CNNs, so that their independent optimizations and feature representations isolate the interference of the labels, and the anatomical knowledge of ROIs is embedded into the registration process without the damaging of texture caused by the lack of texture in labels. For few-shot learning, our Reverse Teaching utilizes the alignment ability of the corresponding CNN to generate additional

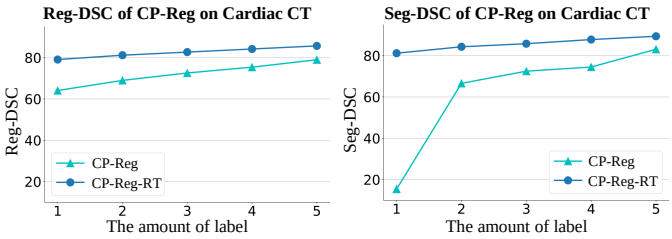

Fig. 9: Our Reverse Teaching keeps the registration performance in fewer-shot situation. The line chart shows the Reg-DSC and Seg-DSC of our PC-Reg(-RT) in different label amount on Cardiac CT.

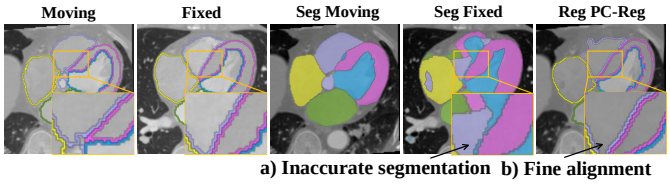

Fig. 10: The correspondence CNN has self-correction ability owing to the training process with inaccurate perception. Therefore, even the inaccurate segmentation of fixed image comes (a), our PC-Reg still achieves good registration results (b).

training data from diverse unlabeled images, thus teaching the perception CNN rich structure and style knowledge for better generalization and the perception of more accurate ROIs. Our experiments on cardiac CT, cervical vertebra CT and brain MR images only with five labels demonstrate competitive registration accuracy, effective texture-preserving ability and great time efficiency of our PC-Reg. Compared with LC-VoxelMorph($\lambda = 1$), we achieve the 12.5%, 6.3% and 1.0% Reg-DSC improvements, revealing our framework with great potential for the deformable registration in clinical practice.

Our PC-Reg has significantly improved the deformable medical image registration only with few labels, but the decoupling makes more modules resulting in the potential accumulation of errors between these modules. Fortunately, timely checking and correcting in each module will effectively avoid the errors' accumulation. The post-processing [40] between the perception and correspondence will optimize the perceived ROIs and the generated additional training data, thus avoiding the interference of error results. The metric learning [8] also can generate a weight map to weaken the mis-aligned regions in the training, avoiding the inefficient learning.

During the training, our correspondence CNN will learn self-correction ability to make fine registration even with inaccurate ROIs from our perception CNN. The registration results are constrained for smooth deformation and same-texture alignment via smooth loss and texture constraint in LNCC loss. When inaccurate segmentation results comes, the correspondence CNN is constrained to output smooth and texture-alignment registration results, thus this inaccurate training process will teach the correspondence CNN the ability to self-correction. As shown in Fig. 10, when inaccurate fixed ROIs are used as input (a), the registration still will make accurate registration results (b).

Future work: The multi-modal registration is still a challenging task in medical image deformable registration, so the

future study of our PC-Reg on multi-modal registration task has great significance.

REFERENCES

[1] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and 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.

[2] J. B. A. Maintz and M. A. Viergever, "A survey of medical image registration," *Medical Image Analysis*, vol. 2, no. 1, pp. 1–36, 1998.

[3] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: a survey," *Machine Vision and Applications*, vol. 31, no. 1, p. 8, 2020.

[4] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: A learning framework for deformable medical image registration," *IEEE Transactions on Medical Imaging*, vol. 38, pp. 1788–1800, 2019.

[5] Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, E. Bonmati, G. Wang, S. Bandula, C. M. Moore, M. Emberton *et al.*, "Weakly-supervised convolutional neural networks for multimodal image registration," *Medical image analysis*, vol. 49, pp. 1–13, 2018.

[6] K. A. J. Eppenhof, M. W. Lafarge, M. Veta, and J. P. W. Pluim, "Progressively trained convolutional neural networks for deformable image registration." *IEEE Transactions on Medical Imaging*, pp. 1–1, 2019.

[7] A. Hering, S. Kuckertz, S. Heldmann, and M. P. Heinrich, "Enhancing label-driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking," in *Bildverarbeitung für die Medizin 2019*. Springer, 2019, pp. 309–314.

[8] Y. He, T. Li, G. Yang, Y. Kong, Y. Chen, H. Shu, J.-L. Coatrieux, J.-L. Dillenseger, and S. Li, "Deep complementary joint model for complex scene registration and few-shot segmentation on medical images," in *16th European Conference on Computer Vision*, vol. 1, 2020, pp. 770–786.

[9] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys (CSUR)*, 2019.

[10] A. Caldas, A. Micaelli, M. Grossard, M. Makarov, P. Rodriguez-Ayerbe, and D. Dumur, "On task-decoupling by robust eigenstructure assignment for dexterous manipulation," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 5654–5661.

[11] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast mr images," *IEEE transactions on medical imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[12] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Medical Image Analysis*, vol. 12, no. 1, pp. 26–41, 2008.

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

[14] J. Wang and M. Zhang, "Deepflash: An efficient network for learning-based medical image registration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4444–4452.

[15] H. Li and Y. Fan, "Non-rigid image registration using fully convolutional networks with deep self-supervision," *arXiv preprint arXiv:1709.00799*, 2017.

[16] Z. Shen, X. Han, Z. Xu, and M. Niethammer, "Networks for joint affine and non-parametric image registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4224–4233.

[17] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning for fast probabilistic diffeomorphic registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 729–738.

[18] T. C. Mok and A. Chung, "Fast symmetric diffeomorphic image registration with convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4644–4653.

[19] B. Kim, D. H. Kim, S. H. Park, J. Kim, J.-G. Lee, and J. C. Ye, "Cyclemorph: Cycle consistent unsupervised deformable image registration," *Medical Image Analysis*, vol. 71, p. 102036, 2021.

[20] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[21] R. Ge, G. Yang, Y. Chen, L. Luo, C. Feng, H. Zhang, and S. Li, "Pv-lvnet: Direct left ventricle multitype indices estimation from 2d echocardiograms of paired apical views with deep neural networks," *Medical image analysis*, vol. 58, 2019.

[22] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, no. 2, pp. 261–318, 2020.

[23] A. M. Hafiz and G. M. Bhat, "A survey on instance segmentation: state of the art," *International Journal of Multimedia Information Retrieval*, pp. 1–19, 2020.

[24] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: A review," *Artificial Intelligence Review*, pp. 1–42, 2020.

[25] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transformations for one-shot medical image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 8543–8553.

[26] Y. He, G. Yang, J. Yang, Y. Chen, Y. Kong, J. Wu, L. Tang, X. Zhu, J.-L. Dillenseger, P. Shao *et al.*, "Dense biased networks with deep priori anatomy and hard region adaptation: Semi-supervised learning for fine renal artery segmentation," *Medical Image Analysis*, p. 101722, 2020.

[27] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60 – 88, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841517301135

[28] P. Liu, M. Lyu, I. King, and J. Xu, "Selflow: Self-supervised learning of optical flow," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4571–4580.

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

[30] G. Ciaparrone, F. L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video multi-object tracking: A survey," *Neurocomputing*, vol. 381, pp. 61–88, 2020.

[31] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *International Journal of Computer Vision*, pp. 1–57, 2020.

[32] A. Medela, A. Picon, C. L. Saratxaga, O. Belar, V. Cabezón, R. Cicchi, R. Bilbao, and B. Glover, "Few shot learning in histopathological images: reducing the need of labeled data on biological datasets," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 1860–1864.

[33] S. Puch, I. Sánchez, and M. Rowe, "Few-shot learning with deep triplet networks for brain imaging modality recognition," in *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*. Springer, 2019, pp. 181–189.

[34] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[35] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial intelligence and statistics*, 2015, pp. 562–570.

[36] Y. Wu and K. He, "Group normalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[37] X. Zhuang and J. Shen, "Multi-scale patch and multi-modality atlases for whole heart segmentation of mri," *Medical Image Analysis*, vol. 31, pp. 77 – 87, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841516000219

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

[39] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ants similarity metric performance in brain image registration." *NeuroImage*, vol. 54, no. 3, pp. 2033–2044, 2011.

[40] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.