Medical Image Analysis 78 (2022) 102379

Contents lists available at ScienceDirect

Medical Image Analysis

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

Dual-stream pyramid registration network☆

Miao Kang[a,b], Xiaojun Hu[a], Weilin Huang[a,*], Matthew R. Scott[a], Mauricio Reyes[c]

[a] Malong LLC, Wilmington, USA
[b] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China
[c] ARTORG Center for Biomedical Engineering Research, Univ. of Bern, Switzerland

ARTICLE INFO

Article history:
Received 22 September 2020
Revised 25 January 2022
Accepted 27 January 2022
Available online 18 February 2022

Keywords:
Medical image registration
Encoder-decoder network
Deformable registration
3D segmentation
Brain MRI

ABSTRACT

We propose a Dual-stream Pyramid Registration Network (referred as Dual-PRNet) for unsupervised 3D brain image registration. Unlike recent CNN-based registration approaches, such as VoxelMorph, which computes a registration field from a pair of 3D volumes using a single-stream network, we design a two-stream architecture able to estimate multi-level registration fields sequentially from a pair of feature pyramids. Our main contributions are: (i) we design a two-stream 3D encoder-decoder network that computes two convolutional feature pyramids separately from two input volumes; (ii) we propose sequential pyramid registration where a sequence of pyramid registration (PR) modules is designed to predict multi-level registration fields directly from the decoding feature pyramids. The registration fields are refined gradually in a coarse-to-fine manner via sequential warping, which equips the model with a strong capability for handling large deformations; (iii) the PR modules can be further enhanced by computing local 3D correlations between the feature pyramids, resulting in the improved Dual-PRNet^{++} able to aggregate rich detailed anatomical structure of the brain; (iv) our Dual-PRNet^{++} can be integrated into a 3D segmentation framework for joint registration and segmentation, by precisely warping voxel-level annotations. Our methods are evaluated on two standard benchmarks for brain MRI registration, where Dual-PRNet^{++} outperforms the state-of-the-art approaches by a large margin, i.e., improving recent VoxelMorph from 0.511 to 0.748 (Dice score) on the Mindboggle101 dataset. In addition, we further demonstrate that our methods can greatly facilitate the segmentation task in a joint learning framework, by leveraging limited annotations.

1. Introduction

Deformable image registration has been widely used in image diagnostics, disease monitoring, and surgical navigation, with the goal of learning the anatomical correspondence between a moving image and a fixed image. A registration process mainly consists of three steps: establishing a deformation model, designing a function for similarity measurement, and a learning step for parameter optimization. Traditional deformable registration methods, such as Demons (Vercauteren et al., 2009), Large Diffeomorphic Distance Metric Mapping (LDDMM) (Glaunès et al., 2008) and symmetric normalization (SyN) (Avants et al., 2008), often cast the deformable registration as a complex optimization problem that involves intensive computation by densely measuring voxel-level similarities. Recent deep learning technologies have advanced this task considerably by developing learning-based approaches, which allow them to leverage the strong feature learning capability of deep networks (Miao et al., 2016; Yang et al., 2017; Hering et al., 2019a; 2019b; Nielsen et al., 2019; Kuckertz et al., 2020), resulting in fast training and accurate inference, e.g., by taking orders of magnitude less time.

However, learning-based approaches for medical image registration often require strong supervised information, such as ground-truth registration fields or anatomical landmarks. While obtaining a large-scale medical dataset with such strong annotations is extremely expensive, which inevitably limits the clinical application of supervised approaches. Recently, unsupervised learning-based registration methods have been developed, by learning a registration function that maximizes the similarity between a moving volume and a fixed volume. For example, Balakrishnan et al. proposed VoxelMorph able to learn a parameterized registration function using a convolutional neural network (CNN) (Balakrishnan et al., 2018). Furthermore, they introduced an auxiliary loss able to integrate segmentation masks into the loss function, as described in Balakrishnan et al. (2019). However, Lewis et al. demonstrated

☆ This work was initially presented at MICCIA 2019.

* Corresponding author.
E-mail addresses: whuang@malongtech.com, whuang@robots.ox.ac.uk (W. Huang).

https://doi.org/10.1016/j.media.2022.102379
1361-8415/

that the performance of existing CNN-based approaches can be limited in real-world clinical applications, where two medical images or volumes may have significant spatial displacements or large slice spaces (Lewis et al., 2018).

Recent approaches on optical flow estimation attempted to handle large displacements by gradually refining the estimated flows (Ranjan and Black, 2017; Hui et al., 2018). For example, Ranjan et al. estimated multi-resolution optical flows with a Spatial Pyramid Network (SPN) to warp a moving volume at each pyramid level. Hui et al. introduced feature warping to replace image warping in the process of pyramidal feature refinement, resulting in a lightweight yet effective network. More recently, Eppenhof et al. attempted to train neural networks progressively to handle the problem of large displacements, by expanding the networks gradually with additional layers that are trained on higher resolution data (Eppenhof et al., 2019). This inspired us to design a sequential warping mechanism able to warp two volumes gradually in a coarse-to-fine manner. In addition to learning meaningful feature representation, medical image registration also requires strong pixel-wise correspondences between moving and fixed volumes, which naturally involves learning local correlations between intermediate features of the moving and fixed volumes. Therefore, current optical flow estimation methods, such as Dosovitskiy et al. (2015); Sun et al. (2018); Hui et al. (2018), utilize a correlation layer to enable the network to identify the actual correspondences from convolutional features (Dosovitskiy et al., 2015). This also inspired us to develop a new 3D correlation layer capable of learning such correlations to further enhance feature representation.

In addition, our registration network is able to capture the semantic correspondence between moving and fixed volumes. This allows it to accurately warp the anatomical annotations of moving volumes to the fixed volumes, providing rough supervised information for training a segmentation network on the target volumes where the annotations are not available (Estienne et al., 2019; Zhao et al., 2019a; Hu et al., 2019b). Recent work in Xu and Niethammer (2019) shows that such warped labels can be used as auxiliary data to improve the performance of segmentation when the training data with annotations is very limited.

Furthermore, Estienne et al. proposed an U-ReSNet (Estienne et al., 2019) which is a lightweight framework for joint registration and segmentation, with excellent results achieved. This further confirmed the benefits of joint learning. In addition, Wang et al. presented a label transfer network (LT-Net) able to propagate a segmentation map from the atlas to unlabelled images, by learning the reversible voxel-wise correspondences (Wang et al., 2020). In this work, we demonstrate that our Dual-PRNet^{++} can be integrated into a 3D segmentation framework for joint segmentation and registration, which facilitates the segmentation task using limited manual annotations.

Contributions. This paper extends our preliminary version of Dual-PRNet presented at the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2019 conference (Hu et al., 2019a), with two main extensions: we introduce Dual-PRNet^{++} by improving sequential pyramid registration (PR) with the enhanced PR^{++} modules which boost the performance; and we apply our Dual-PRNet^{++} for joint 3D segmentation and registration. The overall contributions can be summarized as: (i) a two-stream 3D encoder-decoder network is designed to compute two convolutional feature pyramids separately from two input volumes, generating stronger deep features for deformation estimation; (ii) we propose sequential pyramid registration where a sequence of registration fields is estimated by a set of designed pyramid registration (PR) modules. The estimated registration fields perform sequential warping over the decoding layers, which refine the feature pyramids gradually in a coarse-to-fine manner. This equips the model with a strong capability for handling large deformations; (iii) the PR module can be further enhanced by computing local 3D correlations (between two feature pyramids) followed by multiple residual convolutions, which aggregate richer local details of anatomical structure for better estimating the deformation fields, resulting in the improved Dual-PRNet^{++}. In addition, the 3D correlations with more convolutional layers in the PR^{++} module are able to enlarge receptive fields which further enhance the ability to handle large deformations; (iv) our registration networks can be integrated into a 3D segmentation network, resulting in a unified 3D framework for joint segmentation and registration. Finally, our methods are evaluated on brain MRI registration, where the Dual-PRNet^{++} outperforms the state-of-the-art approaches by a large margin. In addition, on 3D segmentation with limited annotations, we demonstrate that our methods can greatly facilitate the segmentation task via joint framework, by accurately warping voxel-level annotations.

2. Related work

In this section, we briefly review recent approaches on learning-based medical image registration, particularly on using deep learning methods. More comprehensive studies on this topic can be referred to Boveiri et al. (2020); Fu et al. (2020); Haskins et al. (2020).

Deep learning technologies have recently been applied to medical image registration. For example, Hu et al. explored the strong capability of CNN to learn deformable image registration, with promising results achieved (Hu et al., 2017). In Miao et al. (2016), CNN regressors were employed to directly estimate transformation parameters, while De Vos et al. attempted to develop a patch-based end-to-end unsupervised deformable image registration network (DIRNet) de Vos et al., 2017), where a spatial transformer network (STN) (Jaderberg et al., 2015) was applied for estimating a deformation field. However, the deformation field estimated by STN is unconstrained, which may cause severe distortions. To overcome this limitation, VoxelMorph (Balakrishnan et al., 2018; 2019) was proposed. It estimates a deformation field by using an encoder-decoder CNN with a regularization penalty on the deformation field. Furthermore, Kuang and Schmah developed an unsupervised method, named as FAIM, which extends VoxelMorph by introducing an explicit penalty loss computing negative Jacobian determinants (Kuang and Schmah, 2018).

However, these methods may fail to estimate large displacements in complex deformation fields, and recent efforts have been devoted to handle this issue by developing stacked multiple networks (de Vos et al., 2019; Zhao et al., 2019b; Kim et al., 2021).

For example, Zhao et al. designed recursive cascaded networks where multiple VoxelMorph are cascaded recursively, which were employed to warp the images gradually (Zhao et al., 2019b). Kim et al. proposed CycleMorph, which consists of two registration networks, taking inputs by switching their orders with a cycle consistency. It can be extended to multi-scale implementation performing on large volumes. This allows the model to better capture transformation relationships at different levels, but at the cost of a high complexity and computational burden as it requires multiple models. In de Vos et al. (2019), multiple ConvNets were stacked into a larger architecture to perform image warping in a coarse-to-fine manner. More recently, several attempts have been made by cascading an affine alignment subnetwork and a deformable subnetwork to improve the performance (Zhu et al., 2020; de Vos et al., 2019; Huang et al., 2021; Zhao et al., 2020). However, sequential combination of multiple networks will result in an accumulation of interpolation artifacts, which may affect the quality of the estimated deformation field.

Therefore, recent approaches attempted to estimate deformation fields at multiple resolutions (Sokooti et al., 2017; Hering et al., 2019a; Mok and Chung, 2020; Liu et al., 2019; Jiang et al., 2020; Lei et al., 2020; Risheng et al., 2021). For example, a RegNet was introduced in Sokooti et al. (2017), which can be trained by using a large set of artificially generated displacement vector fields (DVF), and then the feature maps computed at multiple scales are concatenated to equip the network with fusion information. In Hering et al. (2019a), mlVIRNET was introduced by creating an image pyramid (not feature pyramid), where a single-stream network is applied multiple times for computing the deformation fields at different image resolutions.

Furthermore, Mok et al. proposed a L-level Laplacian pyramid framework (named as LapIRN) to mimic the conventional multi-resolution strategy, which warps the images from the previous level (Mok and Chung, 2020). Eppenhof et al. attempted to expand the networks progressively with additional layers that are trained on higher resolution data (Eppenhof et al., 2019), and a final deformation field can be estimated by averaging multi-resolution deformation fields computed from the pyramidal structure of a U-Net (Çiçek et al., 2016).

These approaches commonly stack the moving volume and fixed volume together as the input of a single-stream CNN, which largely discards transformation relationships between the two volumes. Two-stream encoders have recently been developed, which are able to encode the two volumes separately for better aggregating multi-level correlations in the feature spaces. For instance, Krebs et al. proposed an efficient latent variable model, which maps similar deformations close to each other in an encoding space (Krebs et al., 2019), while Hering et al. developed a 2.5D two-stream convolutional transformer architecture, which is a memory-efficient weakly supervised learning model for multi-modal image registration (Hering et al., 2019b). In Liu et al. (2019), a dual-stream network was developed to predict multi-resolution deformation fields from different convolutional layers independently, which are then enlarged and averaged to generate a final deformation field. Similarly, Liu et al. utilized a dual-stream encoder to obtain two feature pyramids, and then computed a single transformation field with a contrastive loss and a single-stream decoder (Liu et al., 2020). Besides, the two-steam design was further applied in Kuckertz et al. (2020) where two generators with U-Net architecture and two discriminators using patchGAN (Isola et al., 2017) were developed.

In this paper, we design a dual-stream network to compute two meaningful feature pyramids separately, and directly estimate sequential deformation fields in the feature space, in a single pass. Refinements on both registration fields and convolutional features are performed in a layer-wise, sequential, and coarse-to-fine manner, providing an efficient approach to align the two volumes gradually and more accurately in the feature space. This results in an end-to-end trainable model for unsupervised 3D image registration.

3. Dual-stream pyramid registration network

In this section, we describe the details of the proposed Dual-PRNet and Dual-PRNet^{++}, including three main components: (i) a dual-stream encoder-decoder network for computing feature pyramids, (ii) sequential pyramid registration, and (iii) the improved pyramid registration (PR) modules: PR^{++} modules.

3.1. Preliminaries

The goal of 3D medical image registration is to estimate a deformation field Φ which can warp a moving volume $M \subset R^{H\times W\times D}$ to a fixed volume $F \subset R^{H\times W\times D}$, so that the warped volume $W = M \circ \Phi \subset R^{H\times W\times D}$ can be accurately aligned to the fixed one F. We use $M \circ \Phi$ to denote the application of a deformation field Φ to the moving volume with a warping operation, with image registration being formulated as an optimization problem:

$$\hat{\Phi} = \arg\min_{\Phi} \mathcal{L}(F, M, \Phi) \tag{1}$$

$$\mathcal{L}(F, M, \Phi) = \mathcal{L}_{sim}(F, M \circ \Phi) + \lambda \mathcal{L}_{smooth}(\Phi) \tag{2}$$

where $\mathcal{L}_{sim}$ is a function measuring the image similarity between the warped image ($M \circ \Phi$) and the fixed image (F), and $\mathcal{L}_{smooth}$ is a regularization constraint on the deformation field (Φ), which enforces spatial smoothness. Both $\mathcal{L}_{sim}$ and $\mathcal{L}_{smooth}$ can be defined in various forms, and recent efforts have been devoted to developing a powerful approach to computing the deformation field Φ. For example, VoxelMorph (Balakrishnan et al., 2018; 2019) uses a CNN to compute a deformation field, $\Phi = f_\theta(F, M)$, where θ are learnable parameters of the CNN. In VoxelMorph, the deformation warping operation is implemented by using a spatial transformer network (Jaderberg et al., 2015), $M \circ \Phi = f_{stn}(M, \Phi)$, and a single-stream encoder-decoder architecture with skip connections (similar to U-Net (Ronneberger et al., 2015)) is used. Two volumes, M and F, are stacked as the input of VoxelMorph. More details of VoxelMorph are described in Balakrishnan et al. (2018, 2019).

3.2. Dual-stream architecture

Our Dual-PRNet^{++} is built on the encoder-decoder architecture of VoxelMorph, but improves it by introducing a dual-stream design, as shown in Fig. 1. Specifically, the backbone of Dual-PRNet^{++} consists of a two-stream encoder-decoder with shared parameters. We apply an encoder with the same architecture of U-Net (Ronneberger et al., 2015), which contains five convolutional blocks. Except for the first convolutional block, each block has a 3D down-sampling convolutional layer with a stride of 2, followed by a ReLU operation. Thus the encoder reduces the spatial resolution of the input volumes by a factor of 16 in total, as shown in Fig. 2. In the decoding stage, we apply skip connections to the corresponding convolutional maps in the encoding and decoding process. The lower-resolution convolutional maps (from decoding layers) are up-sampled and concatenated with the higher-resolution ones (from encoding layers), following by a 3×3×3 convolution layer and ReLU operation, as shown in Fig. 1. Finally, we obtain two feature pyramids with multi-resolution convolutional features computed from the moving volume and the fixed volume separately.

The proposed dual-stream design allows us to compute feature pyramids from two input volumes separately, and then predict deformable fields from the learned, stronger and more discriminative convolutional features, which is the key to improve the performance. This is different from existing single-stream networks, such as Balakrishnan et al. (2018, 2019) and Kuang and Schmah (2018), which compute the convolutional features from two stacked volumes, and estimate the deformation fields using single-stream convolutional filters. Furthermore, our dual-stream architecture can compute two paired feature pyramids where layer-wise deformation fields can be estimated sequentially at multiple scales. This allows the model to generate a sequence of deformation fields by designing a new sequential pyramid registration method.

Notice that we modify the backbone applied in the original Dual-PRNet (Hu et al., 2019b) by increasing the convolutional blocks from four to five, but reducing the number of channels from 32 at each layer to [8, 16, 16, 32, 32] for the five layers, and also removing the refine units in the original design to keep a lightweight and effective model. This results in a large reduction of the model parameters from 410K to 175K (which might also alleviate the potential overfitting), but maintains the similar performance. For ex-

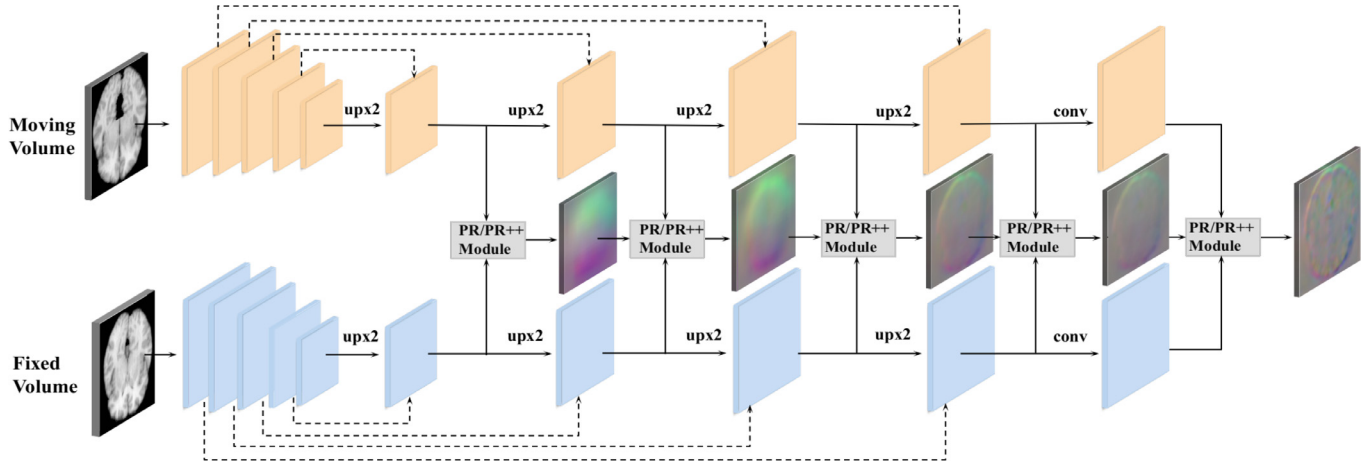


Fig. 1. Framework of the proposed Dual-PRNet^{++}, which is a dual-stream encoder-decoder network, integrated with new sequential pyramid registration including a sequence of pyramid registration (PR) modules or PR^{++} modules. The dual-stream model computes two convolutional feature pyramids separately from two input volumes, while the PR / PR^{++} modules estimate a sequence of deformation fields which can warp the pyramid features gradually in a coarse-to-fine manner. Finally, the final deformation field is generated by sequentially warping the estimated fields, as shown in Fig. 4.

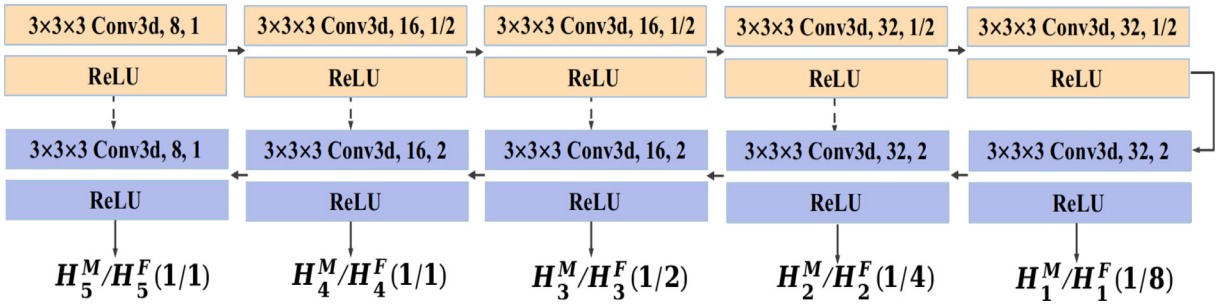


Fig. 2. Backbone of the proposed Dual-PRNet^{++}. It consists of an encoder (yellow) and a decoder (blue), each of which has five convolutional blocks. The convolutional layers are indicated by the filter size, the number of output channels, and spatial resolution (w.r.t. the feature maps of previous layer). H_l^M and H_l^F denote the feature maps computed from the moving volume and the fixed volume separately, with various spatial resolutions (w.r.t. the input volumes) at different convolutional blocks.

ample, it improves the Dice score from 0.631 to 0.653 on the Mindboggle101, but has a reduction with 0.767→0.743 on the LPBA40. In this work, we will use the new backbone for Dual-PRNet^{++} in all our experiments.

3.3. Sequential pyramid registration

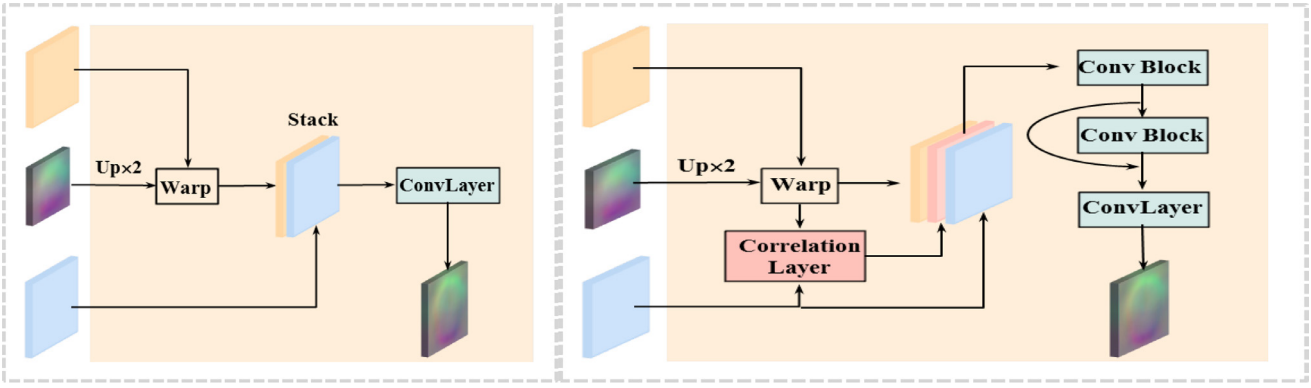
VoxelMorph computes a single deformation field from the convolutional features at the last up-sampling layer in the decoding process, which might make it difficult to handle multi-scale deformations precisely, which are often in case for different anatomical structures of the brain. In this work, we propose a new pyramid registration by designing a set of pyramid registration (PR) modules, which are implemented sequentially at each decoding layer. This allows the model to predict multi-scale deformation fields with increasing resolutions, generating a sequence of pyramid deformation fields, as shown in Fig. 1.

PR module. Each PR module estimates a deformation field at each decoding layer. As input, the PR module uses a pair of convolutional features, together with a deformation field computed from the previous layer (except for the first decoding layer where the deformation field is not available). As output, the PR module yields an estimated deformation field at a given resolution level, which is used in the next pyramid level. The PR module includes a sequence of operations with feature warping, stacking, and convolution (as shown in Fig. 3(a)), which are implemented repeatedly over the decoding layers.

Sequential operations. Specifically, the first deformation field (Φ_1) is computed at the first decoding layer. We first stack the two convolutional features computed at the first decoding layer, and then apply a 3D convolution with size of 3×3×3 to estimate a deformation field. The deformation field (Φ_1) is 3D maps with the same shape of the corresponding convolutional feature maps. It is able to extract coarse-level context information, such as high-level anatomical structure of the brain, which is then encoded into the convolutional features computed at the next decoding layer via feature warping: (i) the current deformation field is up-sampled by using bilinear interpolation with a factor of 2, denoted as $u(\Phi_1)$, and (ii) then it is applied to warp the convolutional maps of the moving volume in the next layer, by using a grid sample operation, as shown in Fig. 3(a). Then the warped convolutional maps are stacked again with the corresponding convolutional features generated from the fixed volume, followed by a convolution operation to estimate a new deformation field. This process is implemented repeatedly at each decoding layer, and can be formulated as,

$$\Phi_l = C_l^{3\times3\times3}\left(H_l^M \circ u(\ \Phi_{l-1}),\ H_l^F\right) \tag{3}$$

where $l = 1, 2, ..., N$, indicates the number of decoding layers. $C_l^{3\times3\times3}$ denotes a 3D convolution at the lth decoding layer. The operator $\circ$ is the warping operation that maps the coordinates of H_l^M to H_l^F using $u(\ \Phi_{l-1})$, where H_l^M and H_l^F are the convolutional feature pyramids computed from the moving volume and the fixed volume at the lth decoding layer.



(a) PR module (b) PR^{++} module

Fig. 3. The proposed (a) Pyramid Registration (PR) module, and (b) its extension: PR^{++} module, which improves the PR module by computing correlation features with further enhancement by residual convolutions.

3.4. PR^{++} modules

Sequential pyramid registration with a set of PR modules was originally introduced in our preliminary version (Hu et al., 2019a). In this extension, we introduce PR^{++} modules to enhance sequential pyramid registration. It improves the PR module by computing correlation features which are further enhanced by residual convolutions, as shown in Fig. 3(b). With respect to the PR module, the PR^{++} module includes two additional operations: 3D correlation and residual convolution, which are the key to enhance the learned features and in turn to boost the performance. Specifically, we design a 3D correlation layer to compute correlation features between the warped features (from the moving volume) and the features from the fixed volume (see Fig. 3(b)). Then the correlation features, together with the two stacked features, are further processed by two convolution blocks with a residual connection to further enhance the representation.

3D correlation layer. In PR^{++} module, a 3D correlation layer is designed to compute the local correlations between the two input volumes in the convolutional feature space. This allows us to aggregate the correlated features which are not directly explored in the original PR module, but can emphasise local details in deep representation.

Specifically, let P_i^W and P_j^F denote the central voxel of the 3D blocks (with size of $(2k+1)^3$) sampled from the feature maps of the warped moving volume and the fixed volume. The correlation relationship between the two sampled 3D blocks can be computed as:

$$C(w_i, f_j) = \frac{1}{(2k+1)^3} \sum_{n_w, n_f \in [-k,k]^3} p_{i+n_w}^W \times p_{j+n_f}^F \quad (4)$$

where $n \in [-k,k]^3$ means n iterates over a 3D neighborhood $[-k,k] \times [-k,k] \times [-k,k]$ of P_i^W or P_j^F. In our experiments, k was empirically set to 1. Given a local 3D block on the feature maps of the (warped) moving volume, it is time-consuming to compute the dense correlations over all the 3D blocks sampled from the feature maps of the fixed volume. Therefore, given a 3D block with P_i^W, we only compute the local correlations by sampling a set of P_j^F within a 3D neighborhood of $d \times d \times d$, which can be implemented as 3D convolutions. We use a stride $s_w = 1$ to densely sample P_i^W from the warped feature maps, and set the correlation neighborhood with $d = 3$ on the corresponding fixed feature maps, where P_j^F is sampled with a stride of $s_f = 2$. Each sampled block has the same size of $[-k,k] \times [-k,k] \times [-k,k]$, and we compute direct correlations between two sampled blocks using Eq. (4). This generates 3D correlation maps (P^C) with shape of $[2 \times FL(d/s_f) + 1]^3 \times (H/s_w) \times (W/s_w) \times (D/s_w)$, where $[2 \times FL(d/s_f) + 1]^3 = 27$ is the number of channels. FL indicates a *Floor* computation. The generated correlation maps have the same 3D shape as the feature maps of the moving and fixed volumes, which ensure that the three maps can be stacked together for further processing.

Convolutional enhancement. Our dual-stream architecture computes two separate feature pyramids from two input volumes. However, the key to the registration task is to learn the strong anatomical correspondence between the two volumes in the feature space, which inspired us to design a new mechanism to further aggregate the computed pyramid features. The key function of the proposed PR^{++} module is to provide a powerful approach for learning richer local details from the two features, which ensure more accurate estimations of the deformation fields at multiple levels. To enrich the learned features, the computed correlation maps are stacked with the two pyramid features at each decoding layers: the warped features from the moving volume and the pyramid features from the fixed volume, as shown in Fig. 3(b). The correlation maps have 27 channels over all decoding layers, while the number of channels of the two pyramid features varies over different layers: [8, 16, 16, 32, 32] for the five decoding layers in our experiments.

To this end, we apply two 3D convolution blocks for further processing the stacked features, as shown in Fig. 3(b). Each convolution block consists of two 3×3×3 convolutional layers, followed by a ReLU operation. The first convolution block reduces the channels of the stacked features considerably from [43, 59, 59, 91, 91] to [8, 16, 16, 32, 32] at the five decoding layers, which are consistent with the numbers of channels applied in the PR module for computational efficiency. In addition, a residual connection is applied to the second convolutional block, in an effort to preserve more context information, and at the same time, to extract discriminative features between the two volumes. Finally, a convolution layer is used to estimate the deformation field. The new PR^{++} module is applied to our dual-stream registration framework, resulting in the enhanced version Dual-PRNet^{++}, which boost the performance of the original Dual-PRNet, as demonstrated in our experiments.

3.5. Final deformation field

The proposed Dual-PRNet^{++} generates five sequential deformation fields with increasing resolutions, indicated as [Φ_1, Φ_2, Φ_3, Φ_4, Φ_5]. To compute the final deformation field, an estimated deformation field is up-sampled by a factor of 2, and then is warped by the following deformation field being estimated. Such up-sampling and warping operations are implemented repeatedly

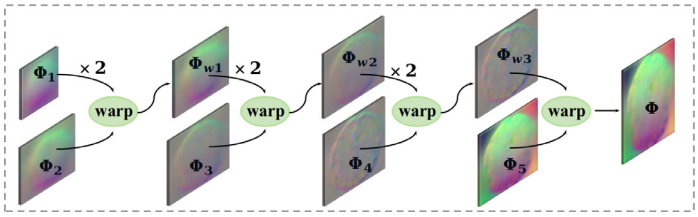


Fig. 4. The final deformation field is computed by sequentially warping the current field with the previous one (×2 up-sampling).

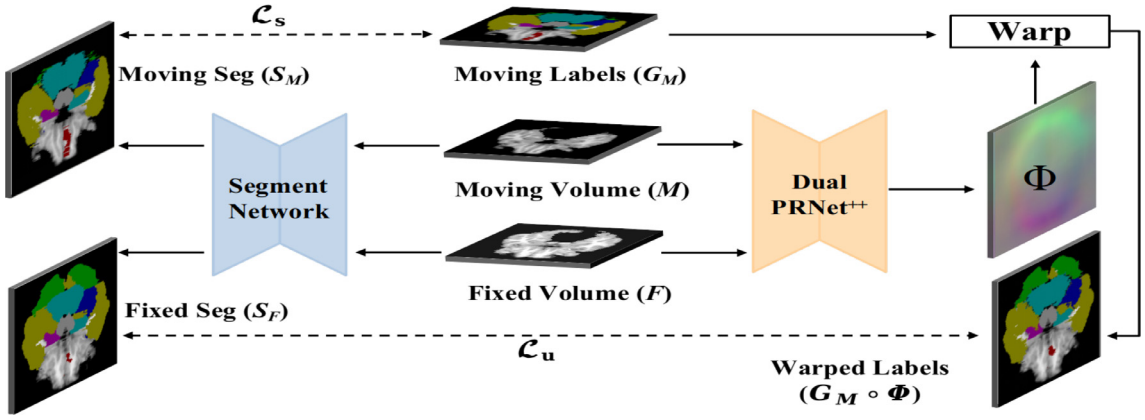


Fig. 5. Joint segmentation and registration framework, where the proposed Dual-PRNet^{++} is used as a joint registration network, and is trained jointly with a 3D segmentation network. Dual-PRNet^{++} is applied to warp the moving labels, which are then used as supervision of the corresponding fixed volume.

and sequentially to generate the final deformation field (as shown in Fig. 4), which encodes rich multi-level context information with multi-scale deformations. This allows the model to propagate strong context information over hierarchical decoding layers, where the estimated deformation fields are refined gradually in a coarse-to-fine manner, and thus aggregate both high-level context information and low-level detailed features. The high-level context information equips our model with the ability to work with large-scale deformations, while the fine-scale features allows it to model detailed anatomical structure information. We integrate PR modules or PR^{++} modules into our dual-stream architecture, resulting in an end-to-end trainable model. By simply following VoxelMorph (Balakrishnan et al., 2018; 2019), a negative local cross correlation (NLCC) is applied as the loss function, coupled with a smooth regularization, e.g., a diffusion regularizer which computes approximate spatial gradients using differences between neighboring voxels, as detailed in Balakrishnan et al. (2018).

4. Joint segmentation and registration

Recent segmentation methods using deep learning technologies often require massive manually annotated data, which is labor-intensive and expensive, particularly for 3D medical images. It is appealing to develop unsupervised or weakly supervised methods for accurate segmentation on 3D medical images. The proposed Dual-PRNet or Dual-PRNet^{++} is able to transfer a moving volume to a fixed volume, which inspired us to adopt such ability to roughly map the available segmentation labels from a source domain to a target domain where the annotations are not provided. This enables us to train a segmentation network on the target MRI domain by using the transferred anatomical labels, without any manual annotation.

In this work, we integrate the registration network into a segmentation network to form a unified framework, as shown in Fig. 5. The framework is related to that of Xu and Niethammer (2019) where DeepAtlas was developed to learn the two tasks simultaneously by using Voxelmorph as the registration network. We extend the DeepAtlas approach by using the proposed Dual-PRNet^{++} as the registration network. Notice that the pre-trained registration network is fixed during the training of segmentation network.

Details of the unified framework of segmentation and registration are presented in Fig. 5. Given a pair of moving (M) and fixed volumes (F), a registration network is adopted to estimate a deformation field (Φ), which is then used to warp the available segmentation labels from the source volume to the unlabelled (target) one, e.g., from the moving volume to the fixed one. Taking the moving and fixed volumes as input, the segmentation network generates two segmentation maps, denoted as S_M and S_F. The source volumes which have ground-truth labels (G) are utilized to train the segmentation network in a regular supervised manner, while the unlabeled volumes can be used to train the same segmentation network, by leveraging the generated labels warped from the corresponding volumes having labels. Specifically, when the moving volumes are labeled and the fixed volumes are unlabeled, the segmentation loss ($\mathcal{L}_{seg}$) for the unified framework can be computed as:

$$\mathcal{L}_{seg} = \lambda_M \mathcal{L}_s(S_M, G_M) + \lambda_F \mathcal{L}_u(S_F, G_M \circ \Phi) \tag{5}$$

where $\mathcal{L}_s$ and $\mathcal{L}_u$ are the segmentation losses computed from the moving volumes (with labels) and the fixed volumes (without labels) respectively. In this paper, we adopt a Dice loss for the segmentation task by following DeepAtlas (Xu and Niethammer, 2019). λ_M and λ_F are the weights that balance the impact of labelled and unlabeled data. Conversely, the two segmentation losses can be computed reversely when we use the moving volumes as unlabeled and the fixed ones as labeled:

$$\mathcal{L}_{seg} = \lambda_M \mathcal{L}_s(S_F, G_F) + \lambda_F \mathcal{L}_u(S_M \circ \Phi, G_F) \tag{6}$$

5. Experimental results and comparisons

Datasets. The proposed Dual-PRNet and Dual-PRNet^{++} are evaluated on 3D brain MRI registration on two public datasets,

LPBA40 (Shattuck et al., 2008) and Mindboggle101 (Klein and Tourville, 2012). The LPBA40 (Shattuck et al., 2008) contains 40 T1-weighted MR images, each of which was annotated with 56 subcortical ROIs. The Mindboggle101 (Klein and Tourville, 2012) has 101 T1-weighted MR images, which were annotated with 25 cortical regions or 31 cortical regions, and can be used to evaluate registration results with more fine and detailed structure of the brain.

Experimental settings. Our experiments on unsupervised 3D brain MRI registration were conducted by following (Kuang and Schmah, 2018). Specifically, on the LPBA40, we train our models on 30 subjects, generating 30×29 volume pairs, and test on the remaining 10 subjects. We follow (Kuang and Schmah, 2018) with the provided code[1], and merge 56 labels into 7 regions specified as: Frontal Lobe, Parietal Lobe, Occipital Lobe, Temporal Lobe, Cingulate Lobe, Putamen, and Hippocampus, which were defined by the major clinical structures of the brain (e.g., each cortical lobe, plus three more regions). Then we center-crop the volumes into a size of $160 \times 192 \times 160$. On the Mindboggle101, we adopt the 25 cortical regions in our experiments on the registration task, and further merge the 25 cortical regions into five large regions corresponding to five anatomical structures of the brain: Frontal lobe, Parietal lobe, Occipital lobe, Temporal lobe, and Cingulate lobe, again by following the implementation details of Kuang and Schmah (2018). The data was divided into 42 subjects (with 1722 pairs) for training, and 20 subjects with 380 pairs for testing. All volumes were cropped with size of $160 \times 192 \times 160$.

For joint segmentation and registration, we conducted experiments on the Mindboggle101 with 31 cortical regions by following Xu and Niethammer (2019). λ_M and λ_F in Eqs. (5) or (6) are set to 1 in our experiments.

The proposed Dual-PRNet and Dual-PRNet^{++} were implemented in Pytorch and trained on 4 Titan Xp GPUs. Batch size is set to 4, due to the limitation of GPU memory. We adopt Adam optimization with a learning rate of 1e-4. The results of VoxelMorph were produced by running the codes provided by the original authors.

Measurements. We adopted the Dice score, Average Symmetric Surface Distance (ASD in mm), and Symmetric Hausdorff Distance(HD in mm), by following (de Vos et al., 2019), as evaluation metrics. The Dice score measures the degree of overlap at the voxel level.

$$\mathcal{D}ice = \frac{2|L_W \cap L_F|}{|L_W| + |LF|} \tag{7}$$

where L_W and L_F denote the labels of warped volume and fixed volume. ASD and HD calculate a surface distance between the moved label and fixed label, which are sensitive to outliers of registration results. Given $\mathcal{R}_W$ and $\mathcal{R}_F$ as the surface point sets of the warped label and fixed label, we can computed the ASD as follows:

$$\mathcal{A}SD = \frac{\sum_{x\in\mathcal{R}_W} D(x, \mathcal{R}_F) + \sum_{y\in\mathcal{R}_F} D(y, \mathcal{R}_W)}{|\mathcal{R}_W| + |\mathcal{R}_F|} \tag{8}$$

where $D(x, \mathcal{R}_F)$ denote a minimal distance of one point x to another point in $\mathcal{R}_F$. Additionally, we adopt the definition of HD as:

$$\mathcal{H}D = \max\{D_h(\mathcal{R}_W, \mathcal{R}_F), D_h(\mathcal{R}_F, \mathcal{R}_W)\} \tag{9}$$

where

$$\mathcal{D}_h(\mathcal{R}_W, \mathcal{R}_F) = \max_{x\in\mathcal{R}_W} \min_{y\in\mathcal{R}_F} D(x, y) \tag{10}$$

In addition, to measure the smoothness of the estimated deformation field, we further compute folding fractions of Jacobian determinant on the field (Ashburner, 2007). The Jacobian determinate $|J_\Phi|$ on a deformation field indicates the relative changes in a local area. Specifically, $|J_\Phi(p)| \leq 0$ means the folding has occurred around the location p of Φ, which means Φ is non-smooth and not physically realistic. Therefore, we adopt the fraction of folding on $|J_\Phi|$ to evaluate the regularity of deformation field.

[1] https://github.com/dykuang/Medical-image-registration.

5.1. Comparisons with the state-of-the-art approaches

We compare our Dual-PRNet and Dual-PRNet^{++} with a number of recent approaches: affine registration, SyN (Shattuck et al., 2008), VoxelMorph (Balakrishnan et al., 2018; 2019), FAIM (Kuang and Schmah, 2018), PMRNet (Liu et al., 2019), LapIRN (Mok and Chung, 2020), Contrastive Registration (CReg) (Liu et al., 2020) and CycleMorph (Kim et al., 2021) on both LPBA40 and Mindboggle101 datasets. As discussed in Section 2, VoxelMorph estimates a single deformation field with a single-stream network. FAIM extends VoxelMorph by designing a new penalty loss on negative Jacobian determinants. PMRNet computes average multi-resolution deformation fields, which are obtained from a dual-stream encoder, as the final deformation field. CReg estimates a single transformation field from feature pyramids by using a contrastive loss and a single-stream decoder. Two registration networks were designed in CycleMorph (Kim et al., 2021), with a cycle consistency, which takes inverse order volumes as inputs. We implemented the affine registration and SyN by using ANTs (Avants et al., 2011). For VoxelMorph and FAIM, we used the codes and models provided by the original authors. For CycleMorph, we utilized the released code, and trained the models on the datasets used in our experiment, by using the same experimental settings as Kim et al. (2021)).

Results and comparisons. On the LPBA40 dataset, as shown in Fig. 7 and Table 1, Dual-PRNet improves the average Dice score to 0.778, and outperforms SyN, VoxelMorph, and CycleMoprh considerably. With the enhanced PR^{++} module, Dual-PRNet^{++} can further increase the average Dice score by 2.0% (Dice 0.798), and achieves the best performance in the term of average Dice score. However, Dual-PRNet^{++} is outperformed by LapIR in the terms of ASD and HD. With the penalty on both the size and nonsmoothness of the deformation field, LapIR is able to obtain the lowest HD and folding fraction on the determinate of the Jacobian.

The results on the Mindboggle101 are shown in Table 2, where our Dual-PRNet^{++} consistently outperforms the other methods on Dice score, and achieves the best performance on all five regions. *It reaches a high average Dice score of 0.748, surpassing the closest one - 0.629 of CReg and CycleMorph, by a large margin. Notice that the new PR^{++} modules lead to a large improvement of* 0.631 → 0.748 *over the original Dual-PRNet, demonstrating its ability to learn detailed brain structure.* Furthermore, Dual-PRNet^{++} achieves an ASD of 0.849, which is the best result among all methods, and a comparable HD with LapIRN. This indicates that our method is able to generate less outliers when performing registration. Notice that our methods can achieve excellent Dice scores by simply using CC loss and smooth loss, but did not explored an additional penalty on the deformation fields as did by LapIRN, which would reduce the HD and ASD.

Table 1
The results of different methods on LPBA40, in the terms of average Dice Score (Avg Dice), Symmetric Hausdorff Distance (HD in mm), and Average Symmetric Surface Distance (ASD in mm).

	Avg Dice↑	HD↓	ASD↓
Affine	0.669	13.283	2.469
VoxelMorph	0.683	14.575	2.238
FAIM	0.664	12.935	1.790
CycleMorph	0.733	12.961	1.886
LapIRN	0.796	**11.824**	**1.504**
Dual-PRNet	0.778	13.549	2.096
Dual-PRNet^{++}	**0.798**	12.983	1.724

Table 2
The results of different methods on Mindboggle101, in the terms of average Dice Score (Avg Dice), Symmetric Hausdorff Distance (HD in mm), and Average Symmetric Surface Distance (ASD in mm).

Region	Frontal	Parietal	Occipital	Temporal	Cingulate	Avg Dice↑	HD↓	ASD↓
Affine	0.455	0.406	0.354	0.469	0.450	0.427	16.27	1.433
SyN	0.558	0.496	0.446	0.578	0.549	0.525	–	–
VoxelMorph	0.532	0.459	0.480	0.585	0.499	0.511	16.977	1.261
FAIM	0.572	0.551	0.537	0.469	0.508	0.527	16.544	1.018
PMRNet	0.579	0.559	0.430	0.544	0.546	0.532	–	–
LapIRN	0.543	0.634	0.477	0.632	0.627	0.583	**15.920**	1.069
CReg	0.644	0.620	0.537	0.703	0.640	0.629	–	–
CycleMorph	0.695	0.612	0.526	0.683	0.628	0.629	16.255	1.009
Dual-PRNet	0.602	0.690	0.550	0.695	0.618	0.631	16.826	1.395
Dual-PRNet^{++}	**0.735**	**0.810**	**0.667**	**0.802**	**0.724**	**0.748**	16.080	**0.849**

Table 3
Ablation study on different components of Dual-PRNet $^{++}$ on Mindboggle101 (5 regions) and LPBA40, with average Dice scores reported.

Dual-stream	PR	Res.	Cor.	Mind101	LPBA40
×	×	×	×	0.511	0.683
✓	×	×	×	0.582	0.767
✓	✓	×	×	0.631	0.778
✓	✓	✓	×	0.694	0.785
✓	✓	✓	✓	0.748	0.798

In addition, we also computed folding fractions of Jacobian determinant on deformation fields which measure the smoothness of the deformation fields. Our Dual-PRNet++ obtained a folding faction of 1.725 on the Mindboggle101, outperforming VoxelMorph with 2.274, while FAIM and CycleMorph achieved higher performance of 0.983 and 1.142 respectively. Notice that both FAIM and CycleMorph have a regularity loss designed specifically to encourage the smoothness of the estimated deformation fields in an explicit manner. For example, negative Jacobian determinants were directly measured as a loss in FAIM, while CycleMorph computes a regularization function and a cycle constraint loss to achieve it.

5.2. Ablation study

We provide ablation studies to further verify the efficiency of individual technical components developed in our Dual-PRNet^{++}. We assessed the benefit of the dual-stream design, sequential pyramid registration with PR modules, and the improved PR^{++} modules. Results from these ablation experiments on the Mindboggle101 are presented in Table 3. To compare our dual-stream architecture with the single-stream design in VoxelMorph, we apply the dual-stream architecture for estimating a single deformation field as VoxelMorph, which achieved an average Dice score of 0.582 on the Mindboggle101 and 0.767 on the LPBA40, improving the single-stream counterpart by +7.1% and +8.4%, respectively. By integrating our sequential pyramid registration with PR modules, the results can be further increased, with +4.9% and +1.1% further improvements on the Mindboggle101 and LPBA40. To further enhance the local details in the learned deep features, we develop the PR^{++} modules which aggregate richer local details by explicitly computing the local correlation features, with residual convolutions for further enhancement. This results in a large further improvement on the Mindboggle101: 0.631→0.748 on Dice score, as shown in Table 3. Particularly, the residual convolutions and 3D correlation have independent improvements of 0.631→0.694 and 0.694→0.748 respectively, making comparable contribution in our design.

5.3. Results on joint segmentation and registration

We further evaluate the performance of joint segmentation and registration framework by using the proposed Dual-PRNet^{++}, which can be integrated into a 3D segmentation network. To make a fair comparison, we replace Voxelmorph with our Dual-PRNet^{++} as the registration network in the joint framework of DeepAtlas (Xu and Niethammer, 2019). Our Dual-PRNet^{++} estimates a deformation field (the final one as described in Section 2.5) which is then used to warp the available segmentation labels from the source volume to the target one, to guide the learning. Notice that in this implementation, we only optimize the segmentation network by fixing the registration one (also referred as semi-DA in Xu and Niethammer, 2019), due to limited GPU memory. Higher performance can be expected with a fully joint learning of the two tasks.

Table 4
Performance of joint segmentation and registration on Mindboggle101 (31 regions).

Model	N=1	N=21	N=65
Segmentation	– –	73.48	81.31
DeepAtlas + VoxelMorph	61.19	75.63	– –
DeepAtlas + Dual-PRNet^{++}	**66.86**	**78.04**	– –

Experiments were conducted on the Mindboggle101. By following DeepAtlas (Xu and Niethammer, 2019), we use 31 labeled regions in the experiments which are different from the 25 regions used in previous registration experiments. Again, with same experimental settings as Xu and Niethammer (2019), the joint networks are trained with N labeled volumes and the remained $65 - N$ volumes unlabeled. It is a fully supervised learning when $N = 65$, which is the total number of the volumes in the dataset. We use $N = 21$ in our experiments by following DeepAtlas, and results are compared in Table 4.

In the case of $N = 21$, DeepAtlas with our Dual-PRNet^{++} obtains an average Dice score of 78.04%, clearly outperforming the pure segmentation network (73.48%) which is only trained on 21 labelled volumes. This demonstrates that the joint registration network is greatly helpful to improve the performance of segmentation network, by leveraging the additional unlabelled volumes. Furthermore, as the joint registration network, our Dual-PRNet^{++} can provide more accurate warped anatomical labels than Voxelmorph used by DeepAtlas (Xu and Niethammer, 2019), resulting in large performance improvements on the joint framework, e.g., 61.19% → 66.86% in one-shot learning ($N = 1$), and 75.63% → 78.04% when $N = 21$. Notice that our result (78.04%) is also closed to the result (81.31%) of fully supervised learning where all labelled volumes (65 in total) are used for training. The results suggest that our Dual-PRNet^{++} can work more effectively in the joint segmentation and registration framework, and is helpful to training the segmentation network with limited data and labels provided.

6. Discussion

In this section, we further study the robustness of our methods in the cases of large displacements and large splice spaces.

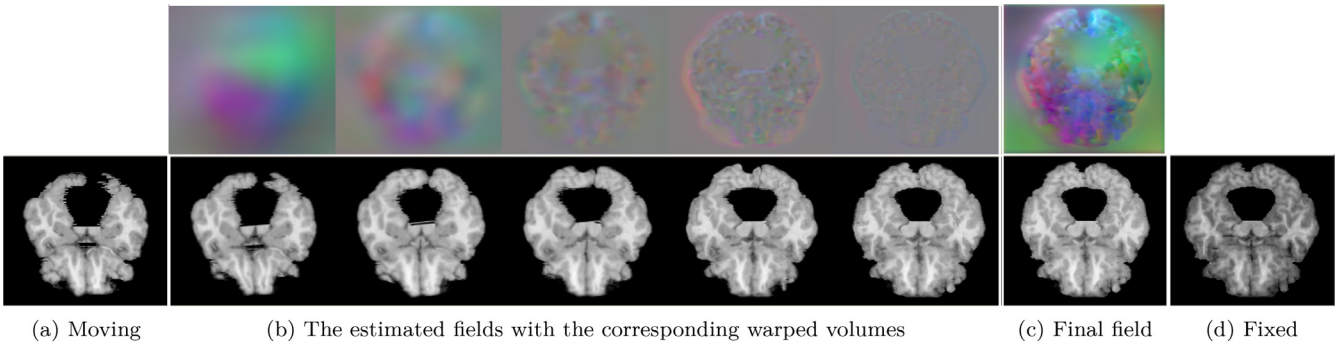


Fig. 6. Visualization of the estimated deformation fields with the corresponding warped volumes: (a) the moving volume, (b) the estimated deformation fields with increasing resolutions (top: left→right), with the corresponding warped volumes (bottom: left→right). Sequential warping is implemented with all the preceding fields, and thus the last warped volume is exactly the same as the volume warped by the final field. (c) the final deformation field with the warped volume, and (d) the fixed volume.

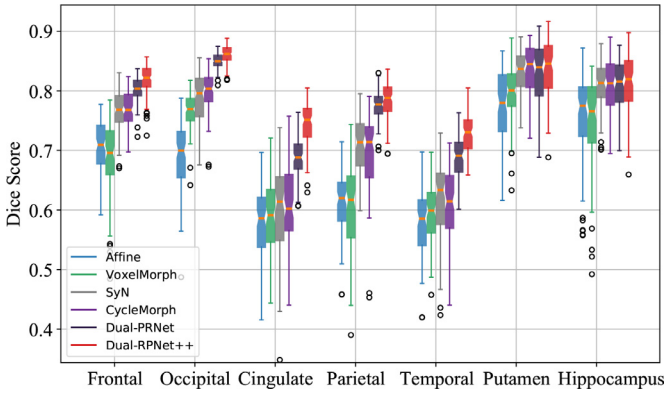


Fig. 7. Dice scores of different methods on LPBA40 (7 regions). The average scores are: 0.669 (Affine), 0.683 (VoxelMorph), 0.731 (SyN), 0.733 (CycleMorph), 0.778 (Dual-PRNet), and 0.798 (Dual-PRNet^{++}).

Then we analyze the performance on detailed structure and cross-dataset learning, with a discussion on the limitation of our methods, which might accumulate interpolation artifacts.

6.1. Robustness

On large displacement. We first visualize the generated multi-resolution deformation fields in Fig. 6 (top). As can be found, a deformation field generated from a lower-resolution layer contains coarse and high-level context information, which is able to encode the high-level semantic information by warping a volume at a larger scale. Conversely, the deformation field estimated from a higher-resolution layer can capture more detailed features. Fig. 6 (bottom) shows the warped images by using the corresponding deformation fields presented. The sequential deformation fields are refined gradually to generate the final field, which can warp the moving image toward the fixed one more accurately, by aggregating more detailed structural information from the preceding fields via sequential warping, as shown in Fig. 6(c).

We investigate the capability of our methods for handling large spatial displacements, and compare our registration results against that of VoxelMorph in Fig. 8. Our Dual-PRNet and Dual-PRNet^{++} can align the image more accurately than VoxelMorph, especially on the regions containing large spatial displacements, as indicated in green or red regions. In addition, as can be found, Dual-PRNet^{++} has an improvement over the original Dual-PRNet, by using the enhanced PR^{++} modules. The design of 3D correlations with more convolutional layers in the PR++ modules can enlarge the receptive fields, which in turn further enhances the ability to handle large displacements.

To further verify the ability of our Dual-PRNet^{++} to handle large displacements quantitatively, we assume that the large displacements more likely happen on the regions where an affine registration can not perform well, such as the "Occipital" and "Parietal" regions on Mindboggle101, which have low Dice scores of 0.354 and 0.406 respectively. Our Dual-PRNet^{++} can have large *relative* improvements of 88%-100% in these two regions (where VoxelMorph only has 13%-36% *relative* improvements), compared to 60%-71% *relative* improvements on the regions where the affine registration achieves a higher Dice score over 0.450. On the LPBA40, our Dual-PRNet^{++} has *relative* improvements of 26%-29% on the regions of "Cingulate" and "Temporal" which have low Dice scores of 0.576 and 0.578 by the affine registration, while only achieving about 8% *relative* improvements on the regions of "Hippocampus" and "Putamen" where the affine registration performs better with 0.753 and 0.775 Dice scores.

On large slice space. We further evaluate the robustness of Dual-PRNet to large slice space. Experiments were conducted on LPBA40, by reducing the slices of the moving volumes from 160×192×160 to 160×24×160. Specifically, we preform the slice reduction on moving volumes by simply removing the slices to 160×24×160, and then interpolate the reduced volumes to the original size (160×192×160) with a spline interpolation (order=1). Then we perform our methods on the reduced-interpolated volumes which have the same size of the original moving volumes. By this way, we can verify different levels of slice reductions between the moving volume and the fixed volume, while keeping the fixed volumes unchanged. During testing, the estimated final deformation field is applied to the labels of the moving volume using zero-order interpolation. With a large reduction of slices from 192 to 24, our Dual-PRNet can still obtain a high average Dice score of 0.711, which even outperforms 0.683 of VoxelMorph (Balakrishnan et al., 2018; 2019) using the original non-reduced volumes. This demonstrates the strong robustness of our model against the large spacing displacements.

6.2. On detailed structure

Compared with LPBA40 dataset, the Mindboggle101 is annotated with the cortical structure, which contains more complicated anatomical structure of the brain, and often requires more accurate local detailed information to identify subtle difference. Fig. 9 demonstrates the registration results on a number of MRI examples from the Mindboggle101. As can be found, VoxelMorph does not provide accurate results on the detailed brain structure. In addition, as demonstrated in ablation studies, our sequential pyramid registration with either PR modules or PR^{++} modules has larger improvements on the Mindboggle101 than that of the LPBA40. Our sequential pyramid registration performs coarse-to-

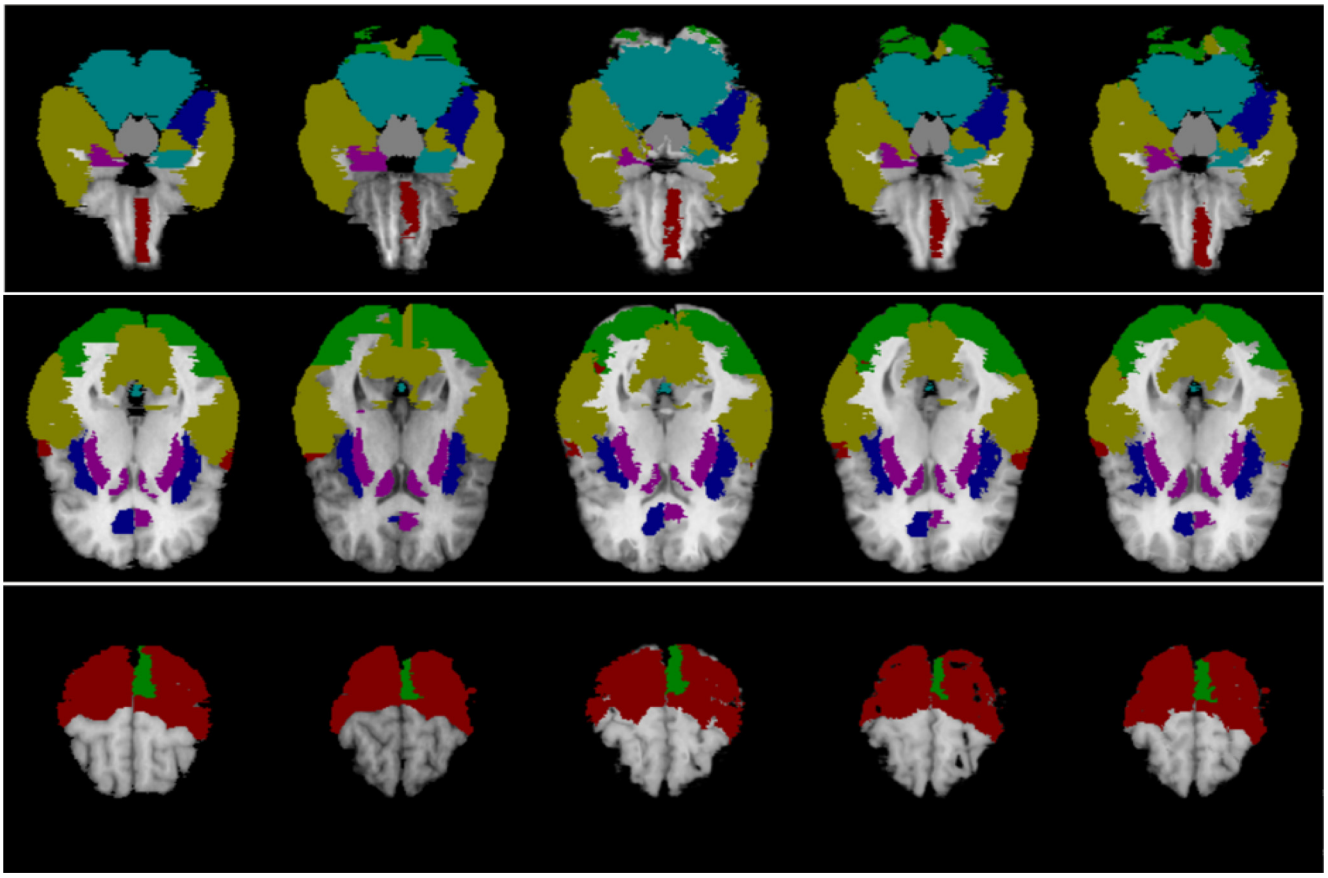


Fig. 8. Registration results on large spatial displacements. From left to right: the moving image, the fixed image, results of VoxelMorph, Dual-PRNet and Dual-PRNet^{++}.

fine refinements of the deformation fields via sequential warping, which naturally aggregate more detailed information from multi-layer feature pyramids. Furthermore, the enhanced Dual-PRNet^{++} can achieve further higher performance by using the improved PR^{++} modules, which are able to compute the local correlations explicitly and thus encodes more detailed structural information. The sequential warping allows the model to propagate the strong high-level context information gradually through the decoding layers, which enhances both the high-level semantic context and the local detailed structure.

We further perform more quantitative analysis. First, we can measure the fineness of the brain structures roughly by computing the ratio of labeled voxels to the total number of voxels in the volumes. The ratios of the labeled voxels on the Mindboggle101 and LPBA40 are 35.83% and 60.02% respectively, which demonstrate that the clinical regions defined in the Mindboggle101 are more fineness. Second, the difficulty of the brain structure on the two datasets can be demonstrated clearly by the performance of an Affine Registration, which has a 0.427 Avg Dice on the Mindboggle101 and a 0.669 Avg Dice on the LPBA40. Therefore, brain structures presented in the Mindboggle101 are more challenging, and our Dual-PRNet++ achieved an over 75% relative improvement, compared to 19% relative improvement obtained on the LPBA40.

6.3. On cross-dataset learning

We further evaluate the generalization capability of the proposed Dual-PRNet and Dual-PRNet^{++} by conducting external cross-dataset validation. For example, we train on the LPBA40 and test on the Mindboggle101, and vice versa. Dual-PRNet^{++} obtains an average Dice score of 0.788 on the LPBA40, and 0.739 on the Mindboggle101, which are slightly lower than the original performance: 0.798 and 0.748, respectively. Similarly, the cross-data performance of Dual-PRNet are 0.747 and 0.581 on the two databases, compared to the original 0.778 and 0.631 respectively. Therefore, the cross-data performance of both Dual-PRNet and Dual-PRNet^{++} are compared favorably against the original results of VoxelMorph (with 0.683 and 0.511 respectively), demonstrating the improved generalization ability of the proposed methods over different datasets.

6.4. Limitations

Dual-PRNet has its limitation by preforming sequential warping on deformation fields, which might result in an accumulation of interpolation artifacts. Thus it yields the highest folding fraction. However, by integrating our PR^{++} module, the folding fraction of deformation field on Dual-PRNet^{++} can be reduced considerably, reaching a higher performance compared favorably against VoxelMorph. However, we note that the current implementation of Dual-PRNet and Dual-PRNet++ do not have an explicit mechanism to regulate the amount of regularization as it is the case for FAIM (Kuang and Schmah, 2018) or LapIRN (Mok and Chung, 2020). Nonetheless, inspired by these works, we note that adding a regularization term based on the determinant of the Jacobian matrix of the deformation is straightforward in our methods, and worth investigating in line with regularizing the solution where trading accuracy does not affect the downstream clinical task. Besides, we further preformed our Dual-PRNet++ using additional supervision of segmentation masks, which can improve the performance from 0.748 to 0.752 on the Mindboggle101. However, this improvement is relatively limited when compared to that of VoxelMorph. (Balakrishnan et al., 2019)

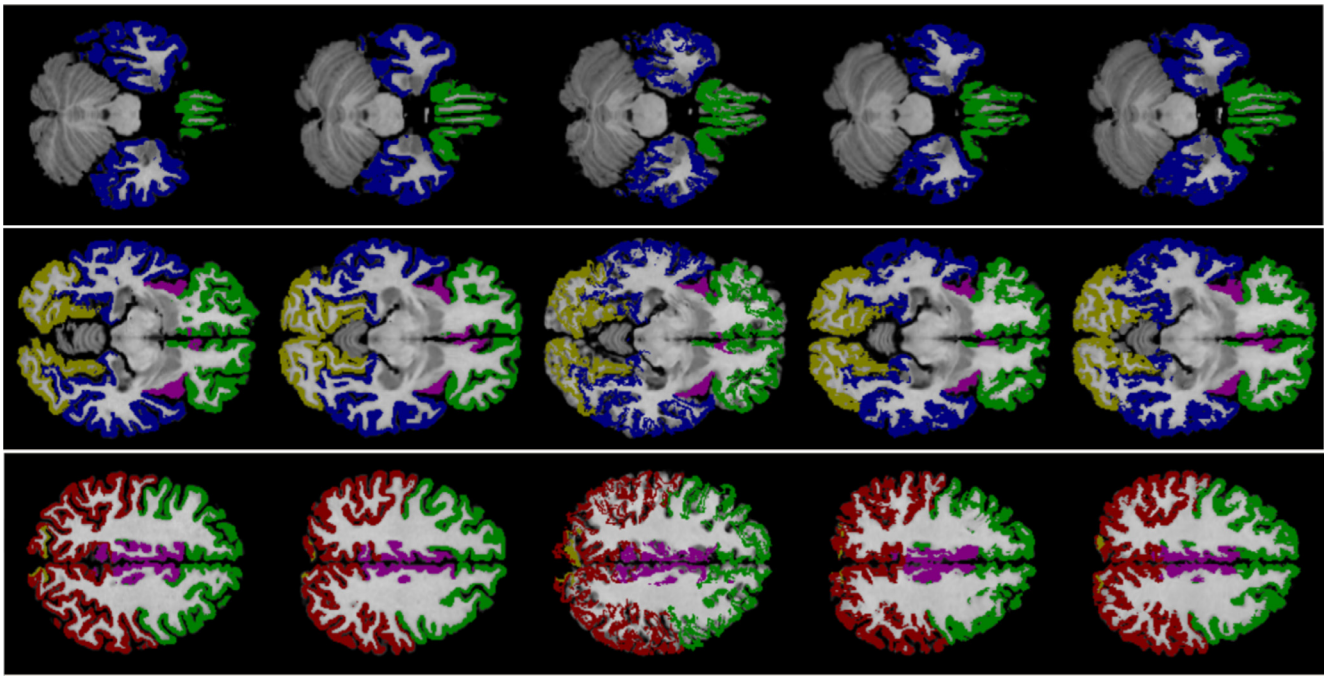


Fig. 9. Registration results on the Mindboggle101. From left to right: the moving image, the fixed image, results of VoxelMorph, Dual-PRNet and Dual-PRNet^{++}.

Besides, in this work, we only performed our methods on 3D brain image (MRI) registration and segmentation. We expect that they can be further applied for or extended to more general 3D medical images, for registration or segmentation, but more experiments and evaluations are required, which can be kept as our future work.

7. Conclusion

We have presented our Dual-Stream Pyramid Registration Network (Dual-PRNet), with its extension, Dual-PRNet^{++}, for unsupervised 3D medical image registration. Our Dual-PRNet has two-stream architecture by design, which allows it to compute two convolutional feature pyramids separately from two input volumes. Then sequential pyramid registration with a set of PR modules is proposed to estimate a sequence of registration fields, which can refine the learned pyramid features gradually in a coarse-to-fine manner via sequential warping. The PR module is further enhanced by computing local correlation features with further enhancement by residual convolutions, resulting in an enhanced Dual-PRNet^{++}. The proposed methods can be integrated into a 3D segmentation framework for joint registration and segmentation, where we demonstrate that it can greatly facilitate the segmentation task by accurately warping the voxel-level labels. Extensive experiments were conducted on LPBA40 and Mindboggle101 databases, where the proposed Dual-PRNet^{++} can outperform the stage-of-the-art methods considerably on unsupervised brain MRI registration.

Credit Author Statment

Miao Kang and Xiaojun Hu contributed to idea developments, implementations and drafted the initial version of the submission. Particularly, Miao Kang was working on performing all experiments on this extension of MICCAI paper to current Journal version.

Weilin Huang contributed to main supervision of the whole work, including improving the idea, advising and organizing the experiments with related analysis and discussion. Organizing writing the paper and making significant revisions.

Matthew R. Scott and Mauricio Reyes provided suggestions and reviewed the paper.

Declaration of Competing Interest

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

Acknowledgments

We would like to thank Zhenlin Xu for sharing the implementation details of joint 3D segmentation framework (DeepAtlas), with image post-processing details on Mindboggle101 dataset.

References

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. Neuroimage 38 (1), 95–113.
Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12, 26–41.
Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ants similarity metric performance in brain image registration. Neuroimage 54, 2033–2044.
Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2018. An unsupervised learning model for deformable medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9252–9260.
Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging 38 (8), 1788–1800.
Boveiri, H.R., Khayami, R., Javidan, R., Mehdizadeh, A.R., 2020. Medical image registration using deep neural networks: a comprehensive review. Comput. Electr. Eng. 87, 106767.
Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 424–432.
Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T., 2015. FlowNet: learning optical flow with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2758–2766.
Eppenhof, K.A., Lafarge, M.W., Veta, M., Pluim, J.P., 2019. Progressively trained convolutional neural networks for deformable image registration. IEEE Trans. Med. Imaging 39 (5), 1594–1604.

Estienne, T., Vakalopoulou, M., Christodoulidis, S., Battistela, E., Lerousseau, M., Carre, A., Klausner, G., Sun, R., Robert, C., Mougiakakou, S., et al., 2019. U-ReSNet: ultimate coupling of registration and segmentation with deep nets. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 310–319.
Fu, Y., Lei, Y., Wang, T., Curran, W.J., Liu, T., Yang, X., 2020. Deep learning in medical image registration: a review. Phys. Med. Biol. 65 (20), 20TR01.
Glaunès, J., Qiu, A., Miller, M.I., Younes, L., 2008. Large deformation diffeomorphic metric curve mapping. Int. J. Comput. Vis. 80, 317.
Haskins, G., Kruger, U., Yan, P., 2020. Deep learning in medical image registration: a survey. Mach. Vis. Appl. 31 (1-2).
Hering, A., van Ginneken, B., Heldmann, S., 2019. mlVIRNET: multilevel variational image registration network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 257–265.
Hering, A., Kuckertz, S., Heldmann, S., Heinrich, M.P., 2019. Memory-efficient 2.5D convolutional transformer networks for multi-modal deformable registration with weak label supervision applied to whole-heart CT and MRI scans. Int. J. Comput. Assist.Radiol. Surg. 14, 1901–1912.
Hu, S., Wei, L., Gao, Y., Guo, Y., Wu, G., Shen, D., 2017. Learning-based deformable image registration for infant MR images in the first year of life. Med. Phys. 44 (1), 158–170.
Hu, X., Kang, M., Huang, W., Scott, M.R., Wiest, R., Reyes, M., 2019. Dual-stream pyramid registration network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 382–390.
Hu, Y., Gibson, E., Barratt, D.C., Emberton, M., Noble, J.A., Vercauteren, T., 2019. Conditional segmentation in lieu of image registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 401–409.
Huang, W., Yang, H., Liu, X., Li, C., Zhang, I., Wang, R., Zheng, H., Wang, S., 2021. A coarse-to-fine deformable transformation framework for unsupervised multi-contrast mr image registration with dual consistency constraint. IEEE Trans. Med. Imaging.
Hui, T.-W., Tang, X., Change Loy, C., 2018. LiteFlowNet: a lightweight convolutional neural network for optical flow estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8981–8989.
Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks. Adv. Neural Inf. Process. Syst. 2017–2025.
Jiang, Z., Yin, F.-F., Ge, Y., Ren, L., 2020. A multi-scale framework with unsupervised joint training of convolutional neural networks for pulmonary deformable image registration. Phys. Med. Biol. 65 (1), 015011.
Kim, B., Kim, D.H., Park, S.H., Kim, J., Lee, J.-G., Ye, J.C., 2021. CycleMorph: cycle consistent unsupervised deformable image registration. Med. Image Anal..
Klein, A., Tourville, J., 2012. 101 Labeled brain images and a consistent human cortical labeling protocol. Front. Neurosci. 6, 171.
Krebs, J., Delingette, H., Mailhé, B., Ayache, N., Mansi, T., 2019. Learning a probabilistic model for diffeomorphic registration. IEEE Trans. Med. Imaging 38 (9), 2165–2176.
Kuang, D., Schmah, T., 2018. FAIM-a convnet method for unsupervised 3D medical image registration. arXiv:1811.09243.
Kuckertz, S., Papenberg, N., Honegger, J., Morgas, T., Haas, B., Heldmann, S., 2020. Deep-learning-based CT-CBCT image registration for adaptive radio therapy 11313, 113130Q.
Lei, Y., Fu, Y., Wang, T., Liu, Y., Patel, P., Curran, W.J., Liu, T., Yang, X., 2020. 4D-CT deformable image registration using multiscale unsupervised deep learning. Phys. Med. Biol. 65 (8), 085003.
Lewis, K. M., Balakrishnan, G., Rost, N. S., Guttag, J., Dalca, A. V., 2018. Fast learning-based registration of sparse clinical images. arXiv:1812.06932.
Liu, L., Aviles-Rivero, A. I., Schönlieb, C.-B., 2020. Contrastive registration for unsupervised medical image segmentation. arXiv:2011.08894.
Liu, L., Hu, X., Zhu, L., Heng, P.-A., 2019. Probabilistic multilayer regularization network for unsupervised 3D brain image registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 346–354.
Miao, S., Wang, Z.J., Liao, R., 2016. A CNN regression approach for real-time 2D/3D registration. IEEE Trans. Med. Imaging 35 (5), 1352–1363.
Mok, T.C., Chung, A.C., 2020. Large deformation diffeomorphic image registration with Laplacian pyramid networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 211–221.
Nielsen, R.K., Darkner, S., Feragen, A., 2019. TopAwaRe: topology-aware registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 364–372.
Ranjan, A., Black, M.J., 2017. Optical flow estimation using a spatial pyramid network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4161–4170.
Risheng, L., Zi, L., Xin, F., Chenying, Z., Hao, H., Zhongxuan, L., 2021. Learning deformable image registration from optimization: perspective, modules, bilevel training and beyond. arXiv:2004.14557.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241.
Shattuck, D.W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K.L., Poldrack, R.A., Bilder, R.M., Toga, A.W., 2008. Construction of a 3D probabilistic atlas of human cortical structures. Neuroimage 39, 1064–1080.
Sokooti, H., De Vos, B., Berendsen, F., Lelieveldt, B.P., Išgum, I., Staring, M., 2017. Nonrigid image registration using multi-scale 3D convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 232–239.
Sun, D., Yang, X., Liu, M.-Y., Kautz, J., 2018. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8934–8943.
Vercauteren, T., Pennec, X., Perchant, A., Ayache, N., 2009. Diffeomorphic demons: efficient non-parametric image registration. NeuroImage 45, 61–72.
de Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Išgum, I., 2019. A deep learning framework for unsupervised affine and deformable image registration. Med. Image Anal. 52, 128–143.
de Vos, B.D., Berendsen, F.F., Viergever, M.A., Staring, M., Išgum, I., 2017. 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, pp. 204–212.
Wang, S., Cao, S., Wei, D., Wang, R., Ma, K., Wang, L., Meng, D., Zheng, Y., 2020. LT-Net: label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9162–9171.
Xu, Z., Niethammer, M., 2019. DeepAtlas: joint semi-supervised learning of image registration and segmentation. arXiv:1904.08465.
Yang, X., Kwitt, R., Styner, M., Niethammer, M., 2017. Quicksilver: fast predictive image registration - a deep learning approach. Neuroimage 158, 378.
Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V., 2019. Data augmentation using learned transformations for one-shot medical image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8543–8553.
Zhao, S., Dong, Y., Chang, E.I., Xu, Y., et al., 2019. Recursive cascaded networks for unsupervised medical image registration. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 10600–10610.
Zhao, S., Lau, T., Luo, J., Chang, E.I.-C., Xu, Y., 2020. Unsupervised 3D end-to-end medical image registration with volume tweening network. IEEE J. Biomed. Health Inf. 24 (5), 1394–1404.
Zhu, Z., Cao, Y., Qin, C., Rao, Y., Ni, D., Wang, Y., 2020. Unsupervised 3D end–to-end deformable network for brain MRI registration. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 1355–1359.