PC-Reg: A pyramidal prediction–correction approach for large deformation image registration

Yin, W.; Sonke, J.J.; Gavves, E.

DOI
10.1016/j.media.2023.102978

Publication date
2023

Document Version
Final published version

Published in
Medical Image Analysis

Medical Image Analysis 90 (2023) 102978

Contents lists available at ScienceDirect

Medical Image Analysis

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

PC-Reg: A pyramidal prediction–correction approach for large deformation image registration

Wenzhe Yin [a,*], Jan-Jakob Sonke [b], Efstratios Gavves [a]

[a] Informatics Institute, University of Amsterdam, The Netherlands
[b] Department of Radiation Oncology, the Netherlands Cancer Institute, Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Image registration
Diffeomorphic registration
Multi-scale registration
Deep learning

ABSTRACT

Deformable image registration plays an important role in medical image analysis. Deep neural networks such as VoxelMorph and TransMorph are fast, but limited to small deformations and face challenges in the presence of large deformations. To tackle large deformations in medical image registration, we propose PC-Reg, a pyramidal Prediction and Correction method for deformable registration, which treats multi-scale registration akin to solving an ordinary differential equation (ODE) across scales. Starting with a zero-initialized deformation at the coarse level, PC-Reg follows the predictor–corrector regime and progressively predicts a residual flow and a correction flow to update the deformation vector field through different scales. The prediction in each scale can be regarded as a single step of ODE integration. PC-Reg can be easily extended to diffeomorphic registration and is able to alleviate the multiscale accumulated upsampling and diffeomorphic integration error. Further, to transfer details from full resolution to low scale, we introduce a distillation loss, where the output is used as the target label for intermediate outputs. Experiments on inter-patient deformable registration show that the proposed method significantly improves registration not only for large but also for small deformations.

1. Introduction

Deformable image registration is a fundamental task in medical image analysis with the objective of finding geometric correspondence between two images. Given an image pair, classical methods (Beg et al., 2005; Cao et al., 2005; Avants et al., 2008; Klein et al., 2009) iteratively optimize the transformation parameters by minimizing a similarity function between a warped moving image and a fixed image. However, this optimization process is time-consuming and undesirable when fast registration is required. Recently, deep neural networks have been widely investigated in deformable medical image registration. The advantage of deep neural networks is representation learning, that is learning from data high-level features of images. Strong feature representations can then be beneficial for accurate deformable image registration. Also, the inference time of the learning-based model is fast, which is desired in some medical applications.

One of the most well-known learning-based methods is VoxelMorph (Balakrishnan et al., 2019). It uses a U-Net architecture (Ronneberger et al., 2015) to estimate the displacement vector field (DVF) and trains the model in an end-to-end fashion. VoxelMorph is an effective and flexible framework and can be extended to diffeomorphic and probabilistic registration (Dalca et al., 2019). The diffeomorphic registration predicts the velocity field, which is then integrated to obtain the deformation vector field. However, VoxelMorph suffers with large deformations due to the high degrees of freedom in the transformation parameters (Mok and Chung, 2020). TransMorph (Chen et al., 2022) improves VoxelMorph by using Transformer (Vaswani et al., 2017) blocks instead of convolutions for its strong ability to model long-range spatial dependency. Swin Transformer (Liu et al., 2021) blocks are used to model the spatial correspondences between the moving and fixed images. However, recent studies (Pegios and Czolbe, 2022) have shown that TransMorph still faces challenges with large displacements. Rühaak et al. (2017) and Heinrich and Hansen (2020) address the large displacement problem by considering the sparse keypoints correspondence and discrete search space.

Intuitively, a natural way to address large deformations is to consider a multi-scale problem setting, where we can model longer and longer-range dependencies on lower and lower scales. LapIRN (Mok and Chung, 2020) is such a multi-scale approach for medical image registration and uses a pyramidal similarity loss. Three identical registration networks are trained progressively on pyramidal images with different scales. This multi-scale design has shown great success in

* Corresponding author.
E-mail address: w.yin@uva.nl (W. Yin).

https://doi.org/10.1016/j.media.2023.102978
Received 9 January 2023; Received in revised form 10 September 2023; Accepted 22 September 2023
Available online 28 September 2023

optical flow estimation, which is a 2D variant of medical image registration in natural images. In learning-based optical flow estimation, it has been shown that the multi-scale pyramidal feature with a single neural network can bring more benefit than the pyramidal images in optical flow estimation (Sun et al., 2018). Hence, a reasonable hypothesis is that registration would also improve performance by using the pyramidal feature instead of pyramidal images. While multi-scale modeling for registration is beneficial, it has three important weaknesses. First, the blur artifacts in the predicted deformation vector field caused by the bilinear/trilinear interpolation (Luo et al., 2021) lead to accumulated error in the final prediction. Second, the integration error for the diffeomorphic registration is accumulated through scales. Third, the multi-scale estimation leads to the "small objects move fast" problem (Lu et al., 2020) where small objects cannot be seen in lower resolution estimation and cannot be recovered in higher resolution.

In this paper, we first draw inspiration from optical flow methods to model geometric registration, specifically PWC-Net (Sun et al., 2018), because of its good performance on 2D optical flow estimation. To address the aforementioned problems with optical flow, however, we note the similarity between the parameterization of deformations as displacement vector fields and the Euler discretization in Euler-based numerical solvers in differential equation modeling. Inspired by Predictor–Corrector (Lapidus and Seinfeld, 1971; Butcher, 2016), a classical method for numerically solving ODEs, we propose a novel end-to-end trained multi-scale registration framework, *PC-Reg*. PC-Reg models multi-scale prediction as integrating an ODE. Unlike previous methods (e.g. LapIRN), our framework allows the integration of ODE across various scales and formulates the registration prediction on each scale as one step of solving an ODE, using multi-scale neural networks, ensuring consistency across steps. We, thereafter, propose a novel error correction module as part of the ODE step across scales to diminish the accumulation of diffeomorphic integration and upsampling errors through scales.

PC-Reg utilizes the pyramidal feature extracted by a neural network such that the model is able to capture high-level semantic long-range dependencies on a coarse scale. Then, we start with a zero-initialized deformation vector field in the lowest scale and predict residual flows in each scale by a prediction network. The deformation vector field is progressively upsampled by trilinear interpolation and added with the residual flow in each scale. Moreover, on each scale, we do one-step correction by a correction network. The final prediction on each scale is the combination of the outputs of prediction and correction networks. The two steps of prediction and correction can balance the predictions of small deformation and large deformation. Our method has a formulation similar to that of the predictor–corrector method for differential equations. In this work, however, we integrate the formulation with the multi-scale neural network in a novel way, where each ODE step represents one scale, and the ODE is integrated through different scales. By contrast, the original method requires a specific physical function and operates with low-dimensional data. It accumulates the deformation vector field from coarse to fine and consistently refines the vector field. Last, due to the intermediate-level predictions lacking details for the full resolution, we constrain the intermediate-level predictions by a distillation loss (Luo et al., 2021) that takes the final output as the target for intermediate-level predictions. As the final prediction in the full resolution has the richest low-level contextual details of deformation, it provides self-guided supervision for the lower-scale predictions. Lastly, we are the pioneers to show the benefits of introducing the distillation loss to medical image registration.

Our pyramidal prediction correction framework can also be extended to allow for diffeomorphic registration to obtain a smooth, invertible deformation vector field that preserves topology. Specifically, the diffeomorphic variant of our method, *PC-Reg-diff*, estimates the velocity field per scale rather than directly predicting the deformation vector field. This can alleviate the integration error on each scale, which avoids damage to the full-resolution prediction.

In this work, we make the following contributions:

1. We develop a novel end-to-end trained multi-scale registration framework, which models the multi-scale prediction as an integrated ODE across scales and considerably improves diffeomorphic registration performance on large deformation datasets.
2. By merging the classical numerical ODE method with multi-scale neural networks, we propose a novel error correction module as part of the ODE step across scales to diminish the accumulation of diffeomorphic integration and upsampling errors through scales.
3. We introduce a distillation loss for self-guided training in pyramidal medical image registration.
4. Our proposed approach significantly surpasses other methods across two distinct datasets. It particularly excels in diffeomorphic registration within the large deformation dataset (Abdomen CT to CT registration), while concurrently maintaining superior performance on small deformations (OASIS dataset).

2. Related work

2.1. Deformable image registration

Supervised Methods Deep learning has boosted the research of medical imaging registration. Early supervised methods (Miao et al., 2016; Cao et al., 2017; Sokooti et al., 2017; Eppenhof and Pluim, 2018) have achieved good results by using convolution neural networks. However, supervised methods commonly require laborious manual annotation. Hence, most methods either use the deformation vector field obtained by classical methods on intensity image and segmentation masks (Cao et al., 2017) as ground-truth or generate the training pairs by applying synthetic transformations that can be used as target labels (Sokooti et al., 2017; Eppenhof and Pluim, 2018; Yin et al., 2022). Beekman et al. (2021) proposed a learning-based diffeomorphic registration approach in a supervised manner. These methods are limited by the performance of the chosen classical methods or the quality of synthetic transformations. Therefore, the generated ground truth is an upper bound of the model, which does not fully reflect the inherent geometric deformation in image pairs.

In this work, we focus on unsupervised image registration. In addition, we conduct experiments with a Dice loss based on the segmentation masks for semi-supervised training.

Unsupervised Methods In order to learn the deformation directly from the data, the unsupervised methods commonly use neural networks to estimate the displacement field and transform the moving image to the fixed image by spatial transformers (Jaderberg et al., 2015). Then, a similarity function is used to evaluate the transformation quality that enables end-to-end training. Often regularization terms are added to the similarity term to obtain plausible results. The most used similarity metrics are mean squared error (MSE) (Beg et al., 2005), sum of squared differences (SSD) (Wolberg and Zokai, 2000), normalized cross-correlation (NCC) (Avants et al., 2008), mutual information (MI) (Viola and Wells III, 1997), and modality independent neighborhood descriptor (MIND) (Heinrich et al., 2012). DIRNet (Vos et al., 2017) proposed a patch-based unsupervised framework where a B-spline transformer is used. De Vos et al. (2019) further extend this work by cascading multiple transformer networks together, which is better on large displacement. These models are based on control points transformation (B-spline).

VoxelMorph (Balakrishnan et al., 2019) proposed to directly predict the dense vector field. The framework can be easily extended to efficient diffeomorphic registration (Dalca et al., 2019) by using the scaling and squaring technique (Arsigny et al., 2006). VoxelMorph uses a U-Net (Ronneberger et al., 2015) architecture for single-scale prediction. TransMorph (Chen et al., 2022) improves VoxelMorph by using Swin Transformer (Liu et al., 2021) as the basic block of U-Net and achieves state-of-the-art performance. Xmorph (Shi et al., 2022) also uses Transformer blocks and adopts cross-attention to fuse the feature of moving

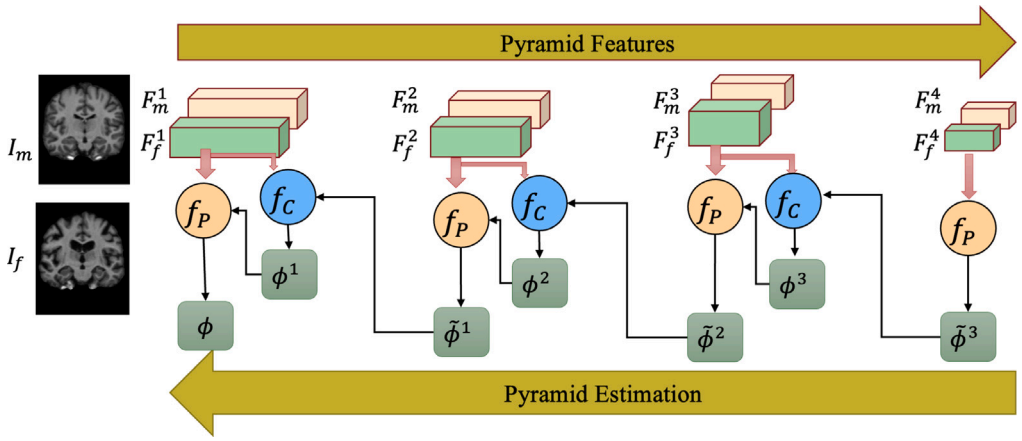


Fig. 1. The framework of the proposed PC-Reg method. Features of both moving and fixed images are extracted in a pyramidal way. F^1, F^2, F^3, F^4 are features at scales $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, respectively. f_P represents the prediction module. f_C is a correction module. ϕ is the predicted deformation vector field. The deformation vector field is first updated by the prediction module and then refined by the correction module.

and fixed images. To better model the large deformation, Mok and Chung (2020) learns the registration networks in different scales and cascades them together. Dual-PRNet (Hu et al., 2019) also adopts the multi-scale design. Different from LapIRN, it predicts the displacement vector field based on the features of moving and fixed images in each scale. In addition, cycle-consistency constraint (Kim et al., 2021) can be added to improve the training of registration networks. The cycle consistency enhances image registration performance by providing an implicit regularization to preserve topology during the deformation.

Several research efforts have been made to tackle the large deformation problem, primarily rooted in the large displacement diffeomorphic metric mapping (LDDMM) model (Beg et al., 2005). Yang et al. (2017) utilized a deep encoder–decoder network for patch-wise prediction of deformation by leveraging the LDDMM model to estimate momentum. Shen et al. (2019a) proposed an end-to-end deep-learning framework for 3D medical image registration, combining affine registration with a vector momentum-parameterized stationary velocity field (vSVF) model. In another work, Shen et al. (2019b) estimated spatiotemporal velocity fields for spatial transformations by using a spatially-varying regularizer attached to deforming objects. More recently, Greer et al. (2021) and Tian et al. (2023) integrated a cycle-consistency constraint in their approach.

Our work employs multi-scale prediction to address the large deformation problem while maintaining good performance on small deformation. Our deformation vector field prediction per scale is based on the features extracted by a neural network rather than pyramidal images. Different from LapIRN and Dual-PRNet, we consider using a correction module to alleviate the accumulated error in each scale, which is critical for reducing the integration error in diffeomorphic registration. In addition, unlike Dual-PRNet which only uses displacement vector field parameterization, our method can be extended to diffeomorphic registration with a good performance.

For LapIRN, we observe that LapIRN has a worse performance for the diffeomorphic registration than the displacement prediction variant (LapIRN-disp), especially on the large deformation dataset (ABD50). We believe this is for the following reasons. (1) The diffeomorphic integration and upsampling errors are accumulated through scales; (2). The three subnetworks of LapIRN are trained sequentially for different scales which makes the errors cannot be corrected during training.

By contrast to (1) and (2), our multi-scale registration framework is trained end-to-end. Also, we formulate the registration prediction per scale as one step of solving an ODE, allowing for consistent and continuous integration steps of the ODE through scales for the velocity field. Also, we utilize a novel correction module to alleviate the accumulated errors. From an ODE perspective, LapIRN can be viewed as solving three ODEs separately and summing up the results together. The above innovations make our method outperform other baseline methods in diffeomorphic registration. Furthermore, our registration module is based on the fusion of extracted features instead of the fusion of images (LapIRN), which makes that our method can better comprehend the spatial relationship of fixed and moving images.

Optical Flow Learning-based optical flow methods can also be used to solve the medical image registration task. Optical flow seeks the dense correspondence between two images. FlowNet (Dosovitskiy et al., 2015) is the first work that introduces neural networks into optical flow estimation. Also, it introduced a cross-correlation module which became the fundamental component in the following works. PWC-Net (Sun et al., 2018) improved FlowNet by pyramidal processing, where warping and cost volume are computed in each feature pyramid level. It iteratively predicts and refines the flow field from coarse to fine. The number of PWC-Net refinement iterations is limited by the number of scales. Thus, RAFT (Teed and Deng, 2020) proposed to use a recurrent network to update the flow field in many iterations. RAFT uses a recurrent scheme to iteratively refine the flow field in 1/8 scale. The pyramidal models have the problem of accumulated error due to the multiple bilinear upsampling. To this end, UPFlow (Luo et al., 2021) proposed a self-guided upsample module and a pyramid distillation loss to address the interpolation blur problem caused by bilinear upsampling between pyramid levels.

In medical image registration, the 1/8 scale prediction will lose detailed complex anatomical structure information. Also, the recurrence at a fine-scale level will bring high computation costs in 3D images. Hence, our method uses a pyramidal way to estimate the deformation vector field which is similar to PWC-Net (Sun et al., 2018). Different from PWC-Net (Sun et al., 2018), we work on the more challenging 3D medical images with complex anatomical deformation. Also, PWC-Net does not integrate the intermediate outputs into the final prediction, while we accumulate the predicted vector field from each scale for modeling the large deformation. Moreover, we additionally use a correction module to alleviate the accumulated upsampling error. Finally, we extend the method to pyramidal diffeomorphic registration, which is not considered in PWC-Net.

2.2. Vision Transformer

Transformer (Vaswani et al., 2017) was developed to model the sequential data in the natural language processing (NLP) area. The basic building block in a transformer is the self-attention module. It is capable to model the long-range relation in sequential data. Vision Transformer (ViT) (Dosovitskiy et al., 2020) extends this concept into

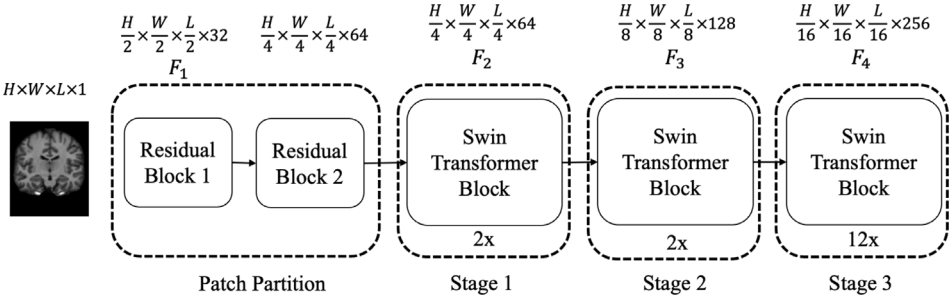


Fig. 2. Feature extraction by Swin Transformer blocks. Two residual blocks are used for partitioning the image into patches, followed by three stages of patch merging and Swin Transformer blocks. There are 2, 2, and 12 Transformer blocks in each stage, respectively. Between every two stages, features are downsampled. Features at different scales(F_1, F_2, F_3 and F_4) are extracted.

the vision community. ViT splits the whole image into a sequence of patches and employs self-attention to replace the convolution. Based on this, Swin Transformer (Liu et al., 2021) proposed a window-based self-attention module and shift window mechanism to produce the feature maps hierarchically. Recently, Transformer has been used in various medical imaging tasks, including segmentation (Chen et al., 2021; Xie et al., 2021), and registration (Liu et al., 2021; Shi et al., 2022). In this work, we use Swin Transformer (Cao et al., 2021; Chen et al., 2022) blocks in 3D as our feature extractor.

3. Predictor–corrector registration networks

Given an image pair comprising a moving image I_m, and a fixed image I_f, image registration aims to find a transformation ϕ that minimizes the similarity function between the transformed image and the fixed image with regularization on the smoothness of the deformation vector field:

$$\min_{\phi} L = L_{sim}(I_m \circ \phi, I_f) + \mathcal{R}(\phi), \tag{1}$$

where L_{sim} is the similarity loss, $\mathcal{R}$ is a regularization term on deformation vector field ϕ, and $I_m \circ \phi$ is the transformed moving image.

Diffeomorphic parameterization. In the diffeomorphic deformation parameterization, the deformation vector field ϕ is parameterized by the velocity field v:

$$\frac{\mathrm{d}\phi^t}{\mathrm{d}t} = v^{(t)}(\phi^{(t)}), \tag{2}$$

such that to minimize

$$v^* = \arg\min_{v} \frac{1}{2}\int_0^1 \|v(t)\|_V^2 \, dt + L_{sim}(I_m \circ \phi, I_f), \tag{3}$$

where $\|v(t)\|_V$ is an appropriate Sobolev norm on the velocity field (Beg et al., 2005). The final deformation vector field is obtained by integrating the velocity field (Dalca et al., 2019). The diffeomorphic formulation is smooth, invertible, and preserves topology, which is important in a clinical setting. Unfortunately, state-of-the-art methods (Balakrishnan et al., 2019; Chen et al., 2022; Mok and Chung, 2020) do not adapt well in the diffeomorphic registration setting when applied on large deformation datasets.

Displacement parameterization. Following LapIRN (Mok and Chung, 2020), we also present a displacement, non-diffeomorphic parameterization, based on the displacement vector field $u(x)$ (Ashburner, 2007):

$$\phi(x) = x + u(x), \tag{4}$$

where x is the identity transformation. With this parameterization, the neural network directly models the displacement vector field $u(x)$. While displacement parameterization usually has better performance on Dice score (Balakrishnan et al., 2019; Mok and Chung, 2020), it does not guarantee invertibility or the preservation of the anatomical topology of organs.

In the following work, we start with the displacement field parameterization and then introduce our diffeomorphic variant.

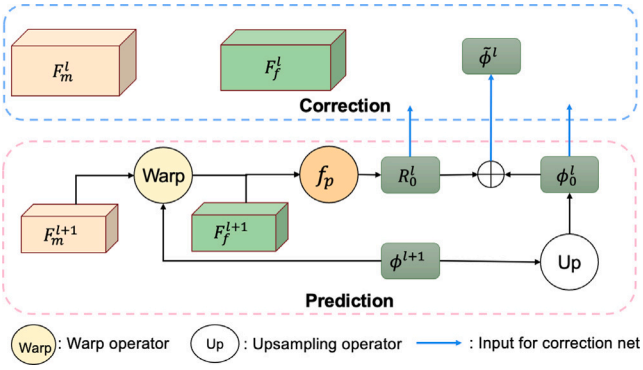


Fig. 3. Prediction module. First, the deformation vector field from last scale ϕ^{l+1} is used to warp the feature F_m^{l+1}. Then, the prediction network f_p estimates a residual vector field R_0^l to update the deformation vector field.

3.1. Model architecture

Our model can be divided into three components: a feature extractor, a predictor, and a corrector. The overall framework, inspired by predictor–corrector methods (Butcher, 2016) in numerical differential equations, is shown in Fig. 1. As an overview, we first use a feature extractor to obtain features from the moving and fixed images in different scales. In each scale, we then use a predictor to estimate the residual vector field to update the deformation field and a corrector module to refine the deformation field. The deformation field is progressively updated through different scale predictions.

3.1.1. Feature extractor

We use a variant of Swin transformer (Liu et al., 2021) as our feature extractor. Given an image pair, I_m and I_f, we first extract pyramidal features F_m^l and F_f^l, $l = 1, \ldots, 4$, where l denotes the scale level of features. Different from the original Swin transformer and the one used in Chen et al. (2022), we first partition the 3D input volumes into different patches by two convolution residual blocks with feature dimensions 32 and 64, respectively. Each block consists of two 3D convolution layers with instance normalization layers. The partitioned patch has the size of $4 \times 4 \times 4$. Then, three stages of patch partition and Swin Transformer blocks are used. Patch partition is for downsampling the features, while each Swin Transformer block consists of one multi-head attention (MSA) and one shifted window-based self-attention (SW-MSA). There are 2, 2, and 12 Swin Transformer blocks in

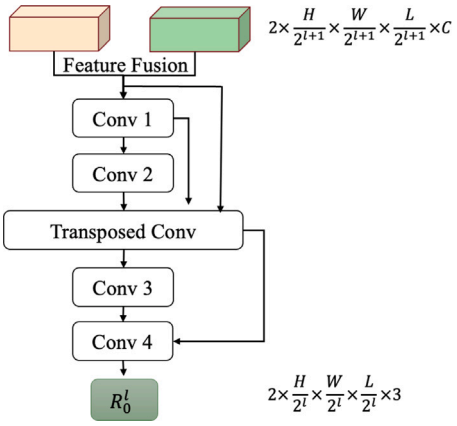


Fig. 4. Module architecture of predication network f_P.

each stage, respectively. We extract features for two images separately and obtain the feature maps at 1/2, 1/4, 1/8, and 1/16 scales.

As shown in Fig. 2, the input images I_m and I_f have the size of $H \times W \times L$, and the extracted features F^l have the size of $\frac{H}{2^l} \times \frac{W}{2^l} \times \frac{L}{2^l} \times (32 * 2^l), l = 1, \ldots, 4$. H, W, L are the height, width, and length of the 3D image. Here, we assume isotropic voxel size. Feature maps at scale l have a double spatial size as the one at scale level $l + 1$.

3.1.2. Residual vector field prediction

After extracting the features of moving and fixed images, we predict the coarse-to-fine deformation vector fields. The pyramidal optical flow methods commonly use bilinear interpolation to upsample the estimated flow field between every two scales. The upsampled flow field is then used to warp the moving image feature for predicting a new flow field in the current scale. This works well in 2D natural images (Sun et al., 2018). However, 3D medical images have more complex structures and may have large deformations, leading to large cumulative errors when upsampling, as well as during the integration of the velocity fields in the diffeomorphic setting. To alleviate this issue, we introduce the prediction–correction model. In this section, we start with the prediction module.

Starting from a zero-initialized deformation vector field ϕ^{l+1} at the lowest scale ($l + 1 = 4$), we progressively upsample ϕ and use a prediction module f_P^l to predict a residual vector field to refine it at each scale. Specifically, the prediction network relies on a transposed convolution layer to predict a residual vector field in higher scales with learnable parameters.

$$\tilde{\phi}^l = \mathrm{Up}(\phi^{l+1}) + f_P(\phi^{l+1}, F_m^{l+1}, F_f^{l+1}) \tag{5}$$

$$\tilde{\phi}^l = \phi_0^l + R_0^l. \tag{6}$$

where we denote $\phi_0^l = \mathrm{Up}(\phi^{l+1})$, and $R_0^l = f_P(\phi^{l+1}, F_m^{l+1}, F_f^{l+1})$. $\tilde{\phi}^l$ is the updated deformation vector field, and Up is the trilinear upsampling operation.

The inputs for the prediction module are the deformation vector field at scale $l+1$, ϕ^{l+1}, and the moving and fixed image features (F_m^{l+1} and F_f^{l+1}). As shown in Fig. 3, the deformation vector field ϕ^{l+1} is first used to warp F_m^{l+1}. Then, we concatenate $F_m^{l+1} \circ \phi^{l+1}$ with F_f^{l+1} as the input of prediction network f_P. The architecture of f_P is illustrated in Fig. 4. The module consists of four convolution layers and one transposed convolution layer. The transposed convolution is a learnable upsampling operation. By using the stride of transposed convolution as 2, we predict the residual vector field in higher resolution. For the rest of the convolution layers, the kernel size is 3 with stride 1. To allow for better gradient flows during training, we adopt the DenseNet (Huang et al., 2017) design, where each layer is densely connected to other layers.

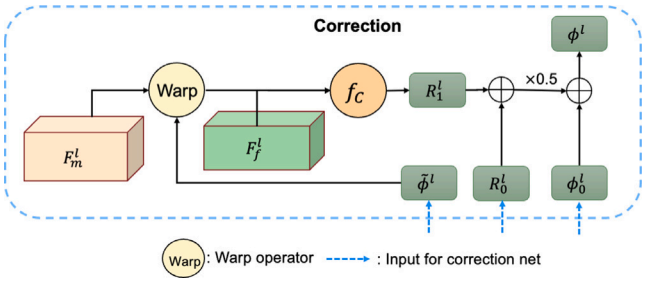


Fig. 5. Correction module. $\tilde{\phi}^l$, R_0^l, and ϕ_0^l are outputs from prediction module. The predicted vector field R_1^l is combined with R_0^l to update the deformation vector field ϕ^l.

3.1.3. Correction module

In Fig. 3, we show how to obtain $\tilde{\phi}^l$ from the features and deformation vector field at scale $l + 1$. The input for correction module is the predicted deformation vector field $\tilde{\phi}^l$, the initial deformation vector field ϕ_0^l at scale l, and the residual vector field $R_0^l = f_P(\phi^{l+1}, F_m^{l+1}, F_f^{l+1})$. Our correction module is inspired by the predictor–corrector method (Butcher, 2016) in numerical methods for differential equations. For a differential equation

$$y' = f(t, y), \tag{7}$$

with initial condition $y(t_0) = y_0$ and step size h, we can solve it numerically by:

$$\tilde{y}_{i+1} = y_i + h f(t_i, y_i) \tag{8}$$

$$y_{i+1} = y_i + \frac{1}{2} h (f(t_i, y_i) + f(t_{i+1}, \tilde{y}_{i+1})) \tag{9}$$

Our prediction step is an analogy to Eq. (8) with step size $h = 1$. Inspired by this method, we propose a correction network f_C to do one step further correction on the deformation vector field that is presented in Fig. 5. We rewrite Eq. (6) and define our prediction–correction step as:

$$\tilde{\phi}^l = \phi_0^l + R_0^l \tag{10}$$

$$\phi^l = \phi_0^l + \frac{1}{2}(R_0^l + R_1^l) \tag{11}$$

$$R_1^l = f_C(\tilde{\phi}^l, F_m^l, F_f^l) \tag{12}$$

The intuitive interpretation is that we want to use the features from the current scale to enrich the details and correct the predicted deformation vector field $\tilde{\phi}^l$. As $\tilde{\phi}^l$ is predicted based on the lower resolution features, F_m^l and F_f^l, it has a better perception of the global information but may lack detailed local information. Hence, the correction step (Eq. (11)) comprises $\tilde{\phi}^l$ with the higher resolution features for updating the deformation vector field. This correction step is for alleviating the error caused by the upsampling or the velocity field numerical integration.

We implement the correction network f_C as a four-layer convolutional network without transposed convolution layer and with dense skip connections. We use Leaky ReLu activations and instance normalization between each layer. The input of f_C is the concatenation of the transformed moving feature by $\tilde{\phi}^l$, $F_m^l \circ \tilde{\phi}^l$ and the feature of fixed image F_f^l. The output is a residual vector field R_1^l.

Overall, we perform coarse-to-fine deformation vector field prediction and correction. The predicted deformation vector field is progressively refined. Our final output is estimated by one-step prediction without correction as we do not have features in full resolution. In our experiments, we evaluate the effectiveness of the proposed correction module through an ablation study.

3.2. Diffeomorphic registration

Next, we extend the prediction–correction registration network to support diffeomorphic registration and thus predict deformation vector fields that are smooth and invertible. Instead of the deformation vector fields, the diffeomorphic variant of our method (PC-Reg-diff) predicts the velocity fields in each scale. Starting with a zero-initialized velocity field, PC-Reg-diff uses prediction and correction modules to update and refine the velocity field at each scale iteratively. We accumulate the velocity fields. As the prediction and correction modules are based on the transformed feature of the moving image, we apply the scaling-and-squaring approach (Dalca et al., 2019) to obtain the deformation vector field in each scale. Then, the obtained deformation vector fields are used to transform the moving image.

In diffeomorphic parameterization, we need to integrate the velocity field v in Eq. (2) to obtain the deformation vector field. In this work, we simply follow the previous works (Arsigny et al., 2006; Dalca et al., 2019; Mok and Chung, 2020) by using a computationally efficient scaling-and-squaring approach for integration. The scaling-and-squaring approach considers the velocity field as stationary. It integrates the velocity field from time 0 to 1 with time steps T. The deformation vector field is represented as a Lie algebra member that is exponentiated to generate a time 1 deformation $\phi^{(1)} = exp(v)$. The scaling-and-squaring recurrence starts with:

$$\phi^{(1/2^T)} = p + \frac{v(p)}{2^T}, \tag{13}$$

where T is the number of steps for integration, and p denotes the spatial locations. Here, we use $T = 7$. The $\phi^{(1)}$ can be obtained using the recurrence:

$$\phi^{(1/2^{t-1})} = \phi^{(1/2^t)} \circ \phi^{(1/2^t)}, \tag{14}$$

Thus, $\phi^{(1)} = \phi^{(1/2)} \circ \phi^{(1/2)}$.

3.3. Loss

Our overall loss consists of three parts: a similarity loss L_{sim}, a regularization loss $L_{\mathcal{R}}$, and a distillation loss L_{DT}. The overall unsupervised loss is defined by:

$$L = L_{sim} + \lambda_R L_{\mathcal{R}} + \lambda_{DT} L_{DT}, \tag{15}$$

where λ_R and λ_{DT} are hyperparameters for weighing different parts.

Similarity loss. Following Balakrishnan et al. (2019) and Chen et al. (2022), we use the local normalized cross-correlation (LNCC) loss as the basis for the similarity loss

$$L_{NCC}(I_m, I_f, \phi) = \sum_{p \in \omega} \frac{\sum_{p_i}(I_f(p) - \bar{I}_f(p))([I_m \circ \phi](p_i) - [\bar{I}_m \circ \phi](p))^2}{(\sum_{p_i}(I_f(p_i) - \bar{I}_f(p))^2)(\sum_{p_i}([I_m \circ \phi](p_i) - [\bar{I}_m \circ \phi](p))^2)}, \tag{16}$$

where p is the voxel coordinates, ω is the image domain, and $\bar{I}$ denotes the mean voxels within the window centered at voxel p. L_{NCC} measures the local similarity within sliding windows. To compute the losses for all intermediate outputs from different scales, we use a multi-scale similarity loss L_{sim} that is similar to LapIRN (Mok and Chung, 2020):

$$L_{sim} = \sum_i \frac{1}{2^{i-1}} L_{NCC}(\text{Pool}(I_m), \text{Pool}(I_f), \phi^i), \tag{17}$$

where Pool is the average pooling operation. Different from the multi-scale similarity loss in LapIRN (Mok and Chung, 2020), where the predicted deformation vector field in different scales is used separately in different stages, we aggregate our multi-scale predictions simultaneously and train the model in an end-to-end manner.

Regularization loss. In displacement field parameterization, training with a sole similarity loss may lead to nonsmooth and unrealistic deformations. To avoid this, we add a regularization loss over the deformation vector field. We adopt the diffusion regularizer from Balakrishnan et al. (2019) as our smoothness regularization. Like the similarity loss, we define the smoothness loss over deformation vector field predictions from all scales:

$$L_{\mathcal{R}} = \sum_i \frac{1}{2^{i-1}} \sum_{p \in \omega} \|\nabla \phi^i(p)\|^2. \tag{18}$$

Distillation loss. In addition, as our prediction relies on the intermediate deformation vector field estimations, it is crucial to provide correct guidance for coarse predictions during training. One way is to define the similarity loss over intermediate outputs as we mentioned above. Besides that, we introduce a distillation loss that uses the final deformation vector field output as the target to supervise the intermediate outputs. The intermediate flows are encouraged to resemble the predicted flow in full resolution. By transferring the information from full-resolution to low-resolution predictions, the intermediate predictions are able to perceive the moving fast small objects by the feedback from the fine-grained output. The idea is similar to knowledge distillation. Similarly to Luo et al. (2021), the distillation loss is defined by the L1 loss between the upsampled deformation vector field prediction V_i at scale i and the final prediction V:

$$L_{DT} = \sum_i \alpha_i (\|\phi - \text{Up}(\phi^i)\|), \tag{19}$$

where α_i is weighting hyperparameter in scale i, and Up is the trilinear upsampling function. We also conduct an ablation study on the distillation loss to show its effectiveness.

Optionally, for best performance, we conduct experiments in a semi-supervised setting. To this end, we use an additional Dice loss together with the L_{NCC} loss as the similarity loss. The Dice loss is calculated between the warped moving segmentation mask by the estimated deformation vector field and the fixed segmentation. The Dice loss is defined by:

$$L_{DSC} = 1 - DSC(S_f, S_m), \tag{20}$$

$$DSC(S_f, S_m) = \frac{1}{K} \sum_{k=1}^{K} \frac{|S_f^k \cap (S_m^k \circ \phi)|}{|S_f^k| + |S_m^k \circ \phi|}, \tag{21}$$

where K is the number of labels, S_f and S_m are the segmentation masks of fixed and moving images. We set the weight of Dice loss as 1.

4. Experiments

4.1. Dataset

We conduct our experiments on OASIS dataset (Marcus et al., 2007), a manually annotated subset of AbdomenCT-1K dataset (Ma et al., 2021), Mindboggle-101 (Klein and Tourville, 2012), and Learn2Reg Abdomen CT-CT (Hering et al., 2022a).

OASIS In this study, we perform inter-patient registration experiments on 414 T1-weighted brain MRI scans from the OASIS dataset (Marcus et al., 2007), which includes scans of Alzheimer patients with mild to moderate symptoms. We use the pre-processed OASIS dataset provided by the Learn2Reg challenge (Hering et al., 2022a). The scans were pre-processed using the standard techniques provided by FreeSurfer (Fischl, 2012), including bias correction, normalization, and skull stripping. The images are resampled into the affinely-aligned common template space. We use the min–max normalization to normalize each scan to $[0, 1]$. Each scan has a spatial size of $160 \times 192 \times 224$ and an isotropic voxel size of 1 mm^3. In addition, segmentation masks of 35 anatomical structures are available. We randomly select the dataset into 294 images for training, 20 for validation, and 100 for

Table 1
Quantitative evaluation results of *unsupervised* setting on OASIS and ABD50 datasets. **Conventional** indicates the conventional methods, while **Displacement** and **Diffeomorphic** are learning-based methods with displacement field and diffeomorphic parameterization, respectively. All methods and their diffeomorphic variants are compared. ↑: the higher the better, ↓: the lower the better.

	Method	OASIS			ABD50		
		DSC ↑	Hausdorff ↓	$\lvert J_\phi\rvert \le 0$ ↓	DSC ↑	Hausdorff ↓	$\lvert J_\phi\rvert \le 0$ ↓
	Unregistered	0.572±0.071	3.91±1.04	–	0.180±0.063	34.38±7.50	–
Conventional	SyN	0.739±0.038	2.38±0.64	<0.0001	0.281±0.079	27.82±6.09	<0.0001
	NiftReg	0.677±0.069	2.96±0.97	<0.0001	0.297±0.073	26.90±5.59	0.001±0.002
Displacement	VoxelMorph	0.785±0.037	2.27±0.66	0.009±0.002	0.204±0.070	32.92±7.07	0.009±0.006
	TransMorph	0.804±0.025	1.95±0.47	0.027±0.008	0.295±0.109	31.49±8.81	0.035±0.015
	LapIRN-disp	0.815±0.026	1.85±0.46	0.003±0.001	0.530±0.081	**20.45**±5.41	0.019±0.006
	PC-Reg	**0.824**±0.024	**1.79**±0.45	0.008±0.002	**0.568**±0.081	22.79±5.54	0.063±0.016
Diffeomorphic	VoxelMorph-diff	0.789±0.034	2.12±0.61	<0.0001	0.208±0.072	32.96±7.14	<0.0001
	TransMorph-diff	0.802±0.026	1.92±0.47	<0.0001	0.306±0.111	30.86±8.99	<0.0001
	LapIRN	0.766±0.044	2.31±0.72	<0.0001	0.353±0.108	26.46±7.27	<0.0001
	PC-Reg-diff	**0.818**±0.023	**1.80**±0.44	<0.0001	**0.571**±0.080	**19.68**±5.77	<0.0001

Table 2
Quantitative evaluation results of *semi-supervised* setting on OASIS and ABD50 datasets. **Displacement** and **Diffeomorphic** are learning-based methods with displacement field and diffeomorphic parameterization, respectively. All methods and their diffeomorphic variants are compared. ↑: the higher the better, ↓: the lower the better.

	Method	OASIS			ABD50		
		DSC ↑	Hausdorff ↓	$\lvert J_\phi\rvert \le 0$ ↓	DSC ↑	Hausdorff ↓	$\lvert J_\phi\rvert \le 0$ ↓
	Unregistered	0.572±0.071	3.91±1.04	–	0.180±0.063	34.38±7.50	–
Displacement	VoxelMorph	0.838±0.022	1.76±0.46	0.011±0.002	0.379±0.065	22.56±4.79	0.010±0.003
	TransMorph	0.852±0.014	1.49±0.32	0.008±0.002	0.510±0.071	18.13±4.74	0.024±0.008
	LapIRN-disp	0.860±0.015	1.56±0.36	0.002±0.001	0.690±0.074	15.20±4.69	0.021±0.006
	PC-Reg	**0.888**±0.013	**1.34**±0.31	0.009±0.002	**0.718**±0.075	**15.16**±5.16	0.053±0.016
Diffeomorphic	VoxelMorph-diff	0.839±0.021	1.65±0.40	<0.0001	0.365±0.064	23.86±5.00	<0.0001
	TransMorph-diff	0.850±0.016	1.49±0.33	<0.0001	0.511±0.067	17.13±4.27	<0.0001
	LapIRN	0.830±0.018	1.78±0.41	<0.0001	0.535±0.094	20.14±5.63	<0.0001
	PC-Reg-diff	**0.875**±0.015	**1.39**±0.30	<0.0001	**0.670**±0.081	**13.64**±4.78	<0.0001

testing. For the testing set, we select every two consecutive images as pairs, resulting in 99 pairs.

ABD50 For CT-CT registration, we use a subset of AbdomenCT data (Ma et al., 2021) as this subset provides manually annotated segmentation masks. It has 50 cases with 12 annotated organs: liver, kidney, spleen, pancreas, esophagus, gallbladder, stomach, aorta, celiac trunk, inferior vena cava, right adrenal gland, and left adrenal gland. Due to varying spatial size, we resample each volume to 224 × 224 × 96. Moving images were resampled based on the voxel size of fixed images. Affine pre-registration is not used. The min–max normalization is used to normalize each scan to $[0, 1]$. We randomly select 35, 5, and 10 volumes for training, validation, and test sets, respectively. Due to the limited number of data, we combine every two volumes in the test set for a comprehensive evaluation, which gives us 90 testing pairs. initial overlap (DSC) of test pairs in ABD50 is low, which means it is initially largely deformed. Hence, we use this dataset to evaluate the model performance on large deformation.

Mindboggle-101 To have a comprehensive evaluation of the proposed method, we additionally use Mindboggle-101 (Klein and Tourville, 2012). Mindboggle-101 has 101 T1-weighted MR images, which are manually annotated with 31 cortical regions. Following Kuang and Schmah (2019) and Hu et al. (2019), we group the 31 cortical regions into five large regions that correspond to five anatomical structures of brains: the frontal lobe, Parietal lobe, Occipital lobe, Temporal lobe, and Cingulate lobe. From this dataset, we use 62 MRI images from OASIS-TRT-20, NKI-TRT-20, and NKI-RS-22, since they are already pre-affinely registered into the MNI152 space. Each image has a voxel grid of 160 × 192 × 160 with isotropic voxel sizes of $1 \times 1 \times 1$ mm^3. We normalize each image into $[0, 1]$. For training, we randomly select 42 images for training, 5 images (20 pairs) for validation, and 15 images (210 pairs) for testing.

Learn2Reg Abdomen CT-CT We use the Abdomen CT to CT dataset from Learn2Reg challenge (Hering et al., 2022a). The dataset contains 50 preprocessed images with 13 manually annotated anatomical labels. The preprocessing includes canonical affine pre-alignment, cropping, padding, and resampling. Each image has the same size of 192 × 160 × 256, and the voxel size is $2 \times 2 \times 2$ mm^3. We normalize each image into $[0, 1]$. Due to only the labels of the original training set being given, for convenient evaluation of all the methods, we select 5 images from the original training set for validation and 10 images (90 pairs) for testing. The rest 35 images are used for training. This dataset is challenging due to the large deformation of abdomen organs between different patients.

4.2. Measurements

We evaluate the model performance by Dice score (DSC) (Dice, 1945), Hausdorff Distance and log Jacobian determinant score, which are also used in previous works (Balakrishnan et al., 2019; Mok and Chung, 2020) and Learn2Reg challenge (Hering et al., 2022a). The final score (mean and standard deviation) is obtained by averaging all the test pairs.

DSC Due to most datasets not having ground truth correspondence (dense/sparse point-wise relationship) available, DSC (Eq. (21)) is used to measure the overlap between the segmentation maps of the deformed moving images and fixed images. We provide the mean and standard deviation of the Dice score.

Hausdorff Distance Hausdorff Distance is a metric to evaluate the maximum distance of a set to the nearest point in the other set. It is defined by:

$$d_H(S_f, \hat{S}_m) = \max\{\max_{x \in S_f} \min_{y \in \hat{S}_m} d(x, y), \max_{y \in \hat{S}_m} \min_{x \in S_f} d(x, y)\}, \tag{22}$$

where $\hat{S}_m = S_m^k \circ \phi$ and d is the Euclidean distance. Following the protocol of Learn2Reg challenge (Hering et al., 2022a), instead of

using the original Hausdorff distance, we adopt a robust variant, 95% Hausdorff distance. It measures the 95th percentile of the distances between boundaries of warped moving image segmentation and fixed image segmentation. We account for the voxel size and use the voxel size of the fixed image during evaluation.

Jacobian Determinant Similarly to other learning-based methods (Balakrishnan et al., 2019; Chen et al., 2022), we use $|J_\phi(p)| = |\nabla\phi(p)| \leq 0$ to evaluate the percentage of folding voxels of the deformation vector field, where p denotes the voxel location and $|\cdot|$ is the determinant.

4.3. Baseline methods

We compare our methods with different state-of-the-art conventional methods and learning-based models.

Conventional Methods We used the conventional method, Symmetric Normalization (SyN) (Avants et al., 2008) as our baseline. For the implementation, we use the public package, Advanced Normalization Tools (ANTs) (Avants et al., 2011). In addition, NiftyReg (Modat et al., 2010), open-source software was used as the second baseline.

Learning-based Methods We compare our model with three learning-based methods, VoxelMorph (Balakrishnan et al., 2019),1u LapIRN (Mok and Chung, 2020) and TransMorph (Chen et al., 2022). For all three models, we use the local normalized cross-correlation loss and the smoothness regularization loss with weights λ_R as 1. For semi-supervised training, we add a DSC loss for each method. We compare the model performance with and without a diffeomorphic parameterization. LapIRN is originally a diffeomorphic registration method with a displacement vector field variant, LapIRN-disp. For VoxelMorph and TransMorph, a scaling-and-squaring integration module is added on top of the prediction for the diffeomorphic registration. We denote the diffeomorphic variants as VoxelMorph-diff and TransMorph-diff.

For TransMorph, we use the default TransMorph-Large setting from Chen et al. (2022), which has a 128 initial embedding dimension. In the ABD50 dataset, $7 \times 7 \times 4$ window size is used to make the TransMorph compatible with the image size in the dataset. As for LapIRN, it has a three-stage training strategy on scales $\frac{1}{4}$, $\frac{1}{2}$, and full resolution, respectively. To have a fair comparison, we train LapIRN with the same iterations in the third stage as other methods. In the semi-supervised setting, the DSC loss is added in the third stage with the full resolution.

4.4. Implementation

In our experiments, we adopt Adam optimizer (Kingma and Ba, 2014) with a learning rate $1e{-}4$. The batch size is 1, and we empirically set $\lambda_R = 1$, $\lambda_{DT} = 2$. For the similarity loss, the window size of L_{NCC} is 9 in full resolution. We train PC-Reg and PC-Reg-diff in both unsupervised and semi-supervised manners. For the semi-supervised training, the weight for DSC loss is 1. We use three stages of Swin Transformer blocks in all our experiments. For the window size, we use the size of $5 \times 6 \times 7$ in the OASIS dataset and $7 \times 7 \times 4$ in the ABD50 dataset to make the model compatible with the data. Our experiments are conducted in an A100 GPU with 40 GB memory.

4.5. Experimental results

In this section, we present the quantitative and qualitative performance of the proposed methods, PC-Reg and PC-Reg-diff, on different datasets.

Unsupervised comparison The unsupervised results of different methods on OASIS and ABD50 are shown in Table 1. We first present the results with displacement vector field parameterization. In both datasets, our method, PC-Reg has the best performance on both the Dice score and the Hausdorff score. The experimental results show that multi-scale approaches, LapIRN-disp and PC-Reg are advantageous compared to single-scale approaches (VoxelMorph and TransMorph).

Table 3
Quantitative evaluation results of **unsupervised** setting on the Mindboggle-101 dataset.

Method	Mindboggle-101		
	DSC ↑	Hausdorff ↓	$\lvert J_\phi\rvert \leq 0$ ↓
Unregistered	0.458±0.016	4.67±0.69	–
VoxelMorph	0.675±0.017	3.77±0.83	0.011±0.001
TransMorph	0.688±0.016	3.37±0.87	0.009±0.001
LapIRN-disp	0.724±0.015	3.35±0.90	0.005±0.001
LapIRN	0.694±0.016	3.34±0.83	0.002±0.0004
PC-Reg	**0.756**±0.013	**2.74**±0.68	0.009±0.001
PC-Reg-diff	**0.735**±0.014	**3.05**±0.80	<0.0001

Table 4
Quantitative evaluation results of **unsupervised** setting on the Learn2Reg Abdomen CT-CT dataset.

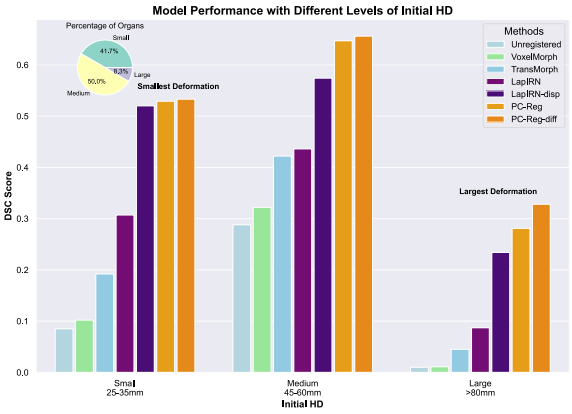
Method	Learn2Reg Abdomen CT-CT		
	DSC ↑	Hausdorff ↓	$\lvert J_\phi\rvert \leq 0$ ↓
Unregistered	0.243±0.064	37.06±7.78	–
VoxelMorph	0.315±0.076	35.34±8.11	0.034±0.021
TransMorph	0.393±0.085	34.38±8.42	0.024±0.012
LapIRN-disp	0.430±0.088	30.56±7.86	0.013±0.008
LapIRN	0.382±0.081	33.28±8.44	0.0001±0.0002
PC-Reg	**0.460**±0.080	**30.02**±7.34	0.056±0.027
PC-Reg-diff	**0.449**±0.087	**31.06**±8.12	0.004±0.007

Specifically, on the large displacement dataset, ABD50, our PC-Reg outperforms VoxelMorph and TransMorph by large margins of up to 20 and 26 percentage points of Dice score, respectively. In ABD50, the initial mean Dice score of 12 annotated organs is around 0.18. Hence, the dataset is challenging due to the large displacements. This shows that the multi-scale design can better capture the long-range displacement. Within multi-scale approaches, PC-Reg has a better performance than LapIRN by 1% Dice score on the OASIS dataset and around 4% in the ABD50 dataset.

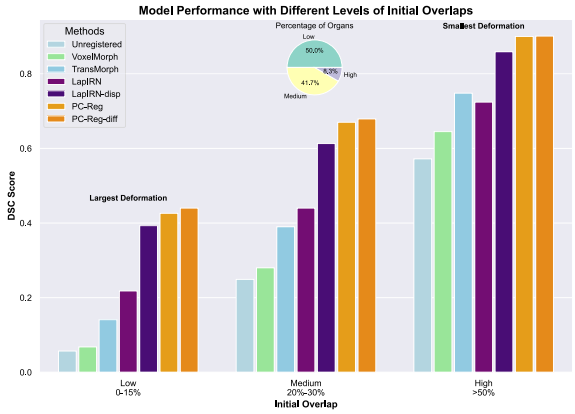
Moreover, as far the diffeomorphic variants, the proposed method, PC-Reg-diff, outperforms all baselines. In ABD50, we notice that our PC-Reg-diff improves PC-Reg around 0.3% Dice score and 3 Hausdorff score, while the counterpart multi-scale approach, LapIRN, observes a decreased performance. More generally, PC-Reg-diff outperforms all other diffeomorphic methods by 20–37 percentage points, including LapIRN. We believe this is due to our correction step alleviating the integration error in multiple scales. The reason is that adding the velocity field from different scales can be regarded as the Euler method of integration in numerical methods for differential equations. By contrast, the predictor–corrector method for integration is more numerically stable. Furthermore, we conducted the paired Wilcoxon tests between PC-Reg(-diff) and other learning-based methods. The p-value is less than 0.005.

In addition to OASIS and ABD50 datasets, we use Mindboggle-101 and Learn2Reg Abdomen CT-CT to evaluate the unsupervised registration performance. The experimental results are shown in Tables 3 and 4. The proposed methods, PC-Reg and PC-Reg-diff, outperform all other methods by a large margin on both Mindboggle-101 and Learn2Reg Abdomen CT-CT datasets. For the multi-scale diffeomorphic prediction, PC-Reg-diff outperforms LapIRN by around 4 percentage of DSC score on the Mindboggle-101 dataset and approximate 6 percentage on the Learn2Reg Abdomen CT-CT dataset. This implies the advantage of our method by modeling the multi-scale registration as an integrated ODE with the prediction–correction scheme.

Semi-supervised comparison Since we are interested in finding out which method can attain the best accuracies overall, we also conducted semi-supervised experiments of different methods on OASIS and ABD50, which are presented in Table 2. In the OASIS dataset, PC-Reg and PC-Reg-diff have the best Dice and Hausdorff performance, outperforming all other learning-based baselines and their diffeomorphic



(a) Model performance with different levels of initial **Hausdorff Distance**. The **Small** initial HD ($25-35mm$) indicates the small deformation, while the **Large** initial HD ($>80mm$) indicates the large deformation.



(b) Model performance with different levels of initial overlaps. **Low** initial overlap ($0-15\%$ DSC score) may correspond to harder deformations, although not necessarily since overlap is also related to organ size. **High** initial overlap corresponds to $>50\%$ DSC score.

Fig. 6. Model performance with different levels of displacements.

variants, that is LapIRN(-disp), TransMorph(-diff), and VoxelMorph(-diff), by 2%–5% in Dice. Furthermore, in the ABD50 dataset, PC-Reg-diff significantly outperforms(p-value < 0.005) other methods by 10%–30% in Dice. Overall, PC-Reg-diff has a stable and relatively good performance on both OASIS (small deformation) and ABD50 (large deformation) datasets.

Performance on each region In order to demonstrate the strengths of the proposed method on organs that undergo large deformations, we further present results per organ of ABD50 in Fig. 7. PC-Reg has a better performance in extreme cases where the organs almost do not overlap initially. Another interesting point is that large deformation often corresponds to organs with very small sizes and could be seen as "fast-small objects". Importantly, PC-Reg and PC-Reg-diff improve accuracy with such small objects undergoing large deformations, as well as with large objects undergoing medium deformations. More details on other datasets are presented in Appendix B.

Peformance for Large Displacement Registration We use ABD50 to evaluate the model performance for large displacement because the organs in ABD50 have low initial overlaps between each pair. To make this clear, we grouped the 12 organs into three categories based on the initial Hausdorff Distance (HD) and the initial DSC score of the unregistered pairs. Hausdorff Distance is a straightforward indicator to reveal the degree of the initial displacement, while the initial DSC score can implicitly reveal the hardness of the deformation.

First, in the ABD50 dataset, based on the initial Hausdorff distance (100 percent), the 12 organs of test pairs are divided into **Low** with 25–35 mm initial HD, **Medium** with 45–60 mm initial HD, and **High** with >80 mm initial HD. The **High** HD indicates the large deformation. The results are presented in Fig. 6(a). Second, we grouped the organs into **Small** with 0–15% initial overlap, **Medium** with 20–30% initial overlap, and **Large** with >50% initial overlap. The small overlap may suggest large displacement, although not always since overlap also depends on organ size. The results are shown in Fig. 6(b). Both PC-Reg and PC-Reg perform the best among all other methods in three categories. Specifically, in the large deformation group, the proposed methods outperform all other methods by a considerable margin. Although LapIRN-disp improves LapIRN a lot with displacement prediction, it cannot preserve the diffeomorphism property we desire. The performance gap between LapIRN-disp and LapIRN is considerable in the large displacement dataset ABD50, and can also be observed in the original LapIRN paper. We suspect that this phenomenon is due to the integration error in LapIRN being accumulated. In contrast, our diffeomorphic prediction PC-Reg-diff has a similar or even better performance as PC-Reg.

Qualitative results The qualitative results of our method on the OASIS dataset are illustrated in Fig. 8. The segmentation contour of four sub-regions, including the ventricles, third ventricle, thalami, and hippocampi are shown in the visualization. It can be seen our diffeomorphic method has a smoother deformation vector field. We demonstrate the qualitative results of our methods on OASIS and ABD50 datasets. The diffeomorphic variant of our method makes the deformation vector field much smoother, especially in the ABD50 dataset (see Fig. 9).

Table 5
The effect of the correction module and distillation loss of PC-Reg on ABD50. Our method with the correction module and the distillation loss achieves consistently better performance.

Correction	Distillation	DSC↑	Hausdorff ↓
✗	✗	$0.503_{\pm 0.097}$	$25.44_{\pm 6.72}$
✗	✓	$0.512_{\pm 0.095}$	$24.87_{\pm 6.40}$
✓	✗	$0.562_{\pm 0.087}$	$23.79_{\pm 6.76}$
✓	✓	$\mathbf{0.568}_{\pm 0.081}$	$\mathbf{22.79}_{\pm 5.54}$

Table 6
The effect of the correction module of PC-Reg-diff on ABD50. Without the correction module, the diffeomorphic registration performance has a considerable drop.

Correction	PC-Reg-diff	
	DSC↑	Hausdorff ↓
✗	$0.475_{\pm 0.113}$	$23.76_{\pm 7.32}$
✓	$\mathbf{0.571}_{\pm 0.080}$	$\mathbf{19.68}_{\pm 5.77}$

4.6. Ablation study

In order to evaluate the effectiveness of our proposed correction module and distillation loss, we conducted an ablation study over these two modules with displacement field parameterization. As shown in Table 5, the correction module brings 6% improvement of the Dice score. Moreover, the distillation loss with weight 2 gives a further improvement on both Dice and Hausdorff scores with and without the correction module. Hence, the correction module is crucial for multi-scale prediction. Also, the distillation is beneficial for transferring detailed textual information from the full-resolution prediction. However, we show that even without distillation loss, purely prediction–correction pyramidal design can outperform other baseline methods. We would like to highlight that the proposed prediction–correction design and the end-to-end pyramidal architectural improvement contribute to better performance. We additionally add an ablation study of the correction module on the diffeomorphic prediction in Table 6 to show that the correction of the accumulated error through scales is crucial for good performance on the large deformation dataset.

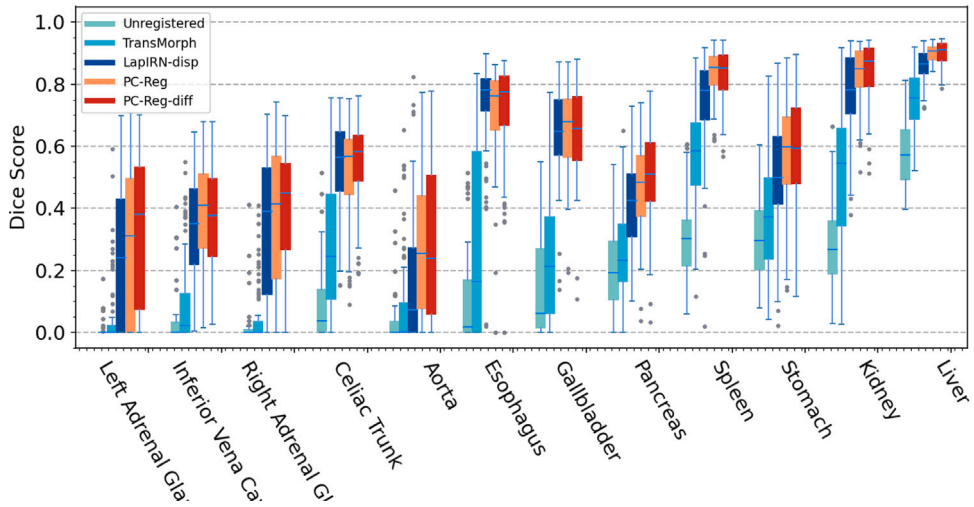


Fig. 7. Quantitative performance of different methods on each sub-region of ABD50 dataset. Sub-regions on the x-axis are ordered by the region size (from low to high). PC-Reg(-diff) outperforms other methods on most sub-regions.

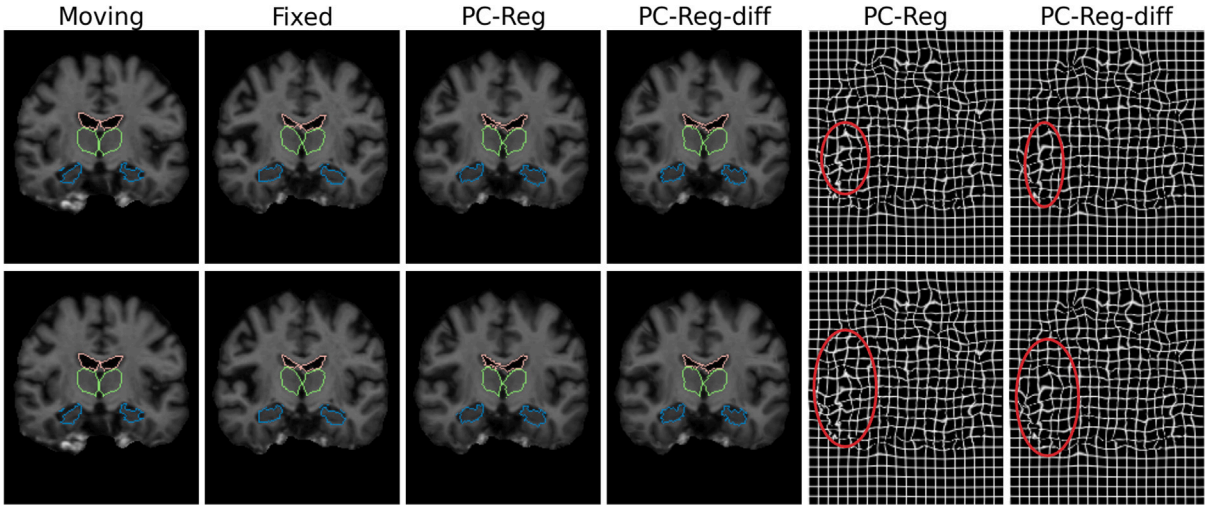


Fig. 8. Qualitative results on OASIS dataset. The first row shows the unsupervised results, while the second row shows the semi-supervised results. The contours of the ventricles, third ventricle, thalami, and hippocampi are shown in the figure.

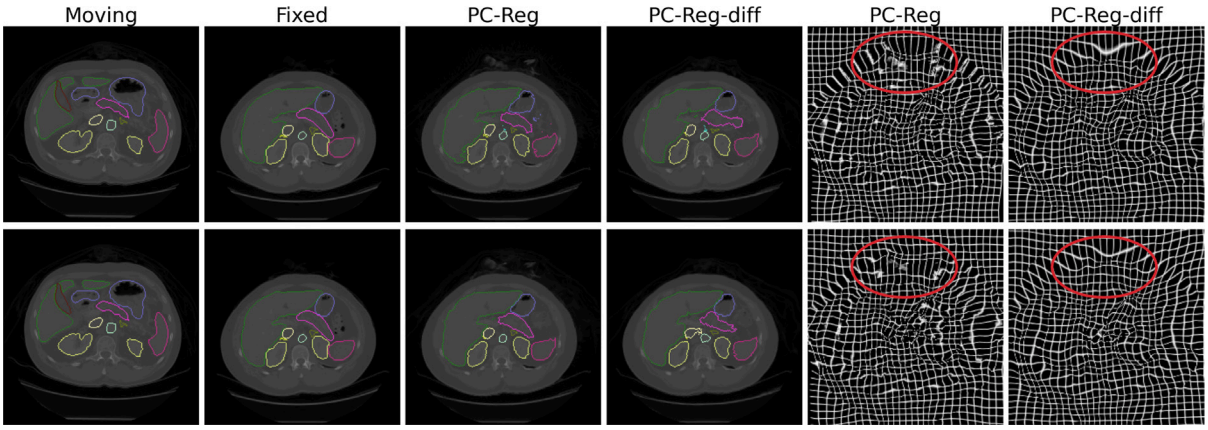


Fig. 9. Qualitative results on ABD50 dataset. The first row shows the unsupervised results, while the second row shows the semi-supervised results. The diffeomorphic variant, PC-Reg-diff, has a smoother deformation vector field with less folding voxels.

In addition, we compare our model performance with respect to the number of the scales of our prediction on the ABD50 dataset in Table 7. Our default model uses three stages of Swin Transformer blocks and predicts the deformation in four scales, including the full resolution. Decreasing/increasing the stages of Swin Transformer blocks will result in fewer or more features, which changes the number of prediction

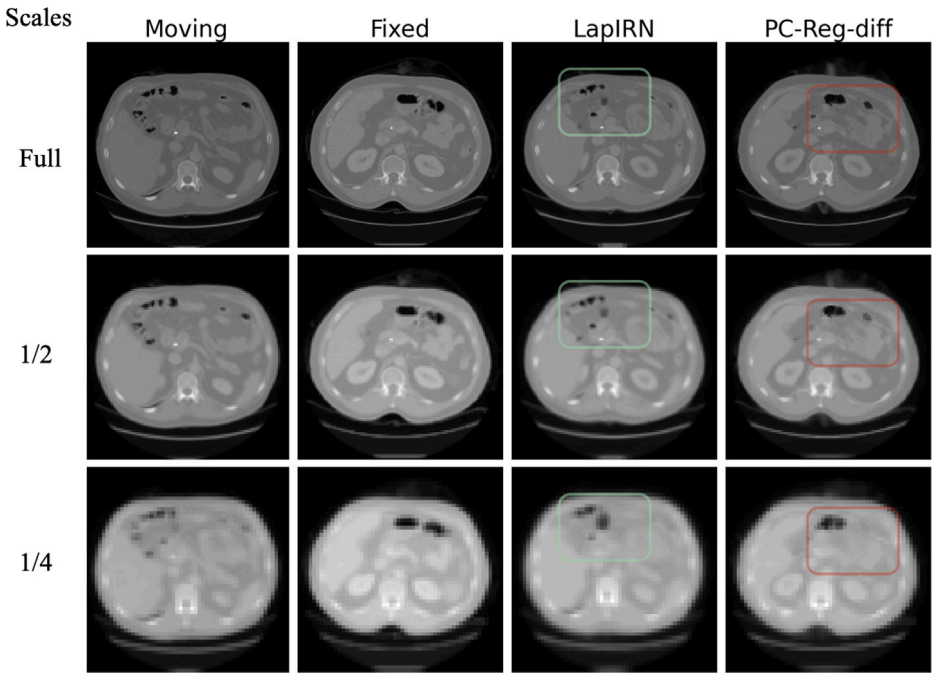


Fig. 10. Qualitative comparison between PC-Reg-diff and LapIRN in each scale on the ABD50 dataset. Full resolution, 1/2, and 1/4 results are shown from top to bottom. PC-Reg-diff outperforms LapIRN in each scale and is able to correct the errors from the lower scales (marked as red). The errors of LapIRN (marked as green) in the low scale (1/4) are not corrected in the higher scales.

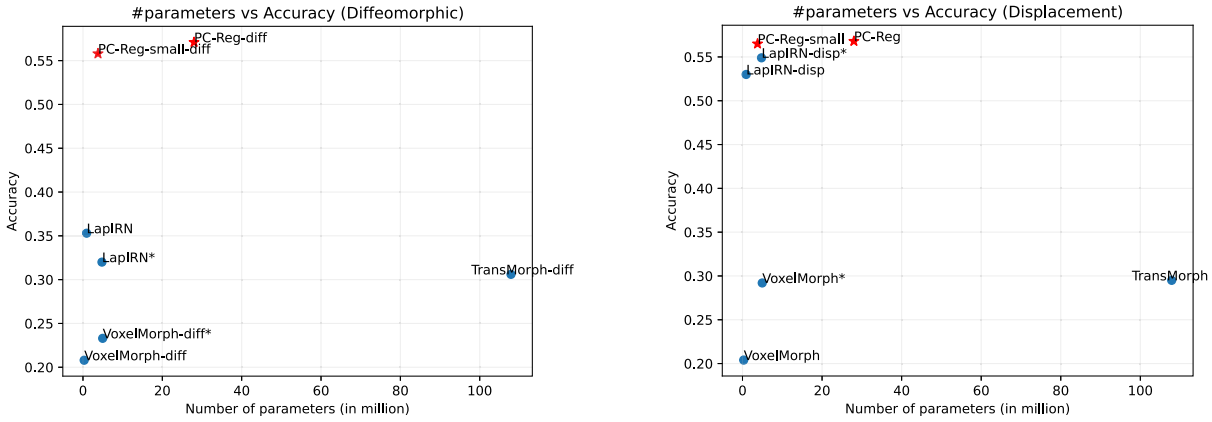


Fig. 11. The number parameters vs accuracy of diffeomorphic registration (left) and displacement (right).

Table 7
The effect of the number of scales on ABD50. Our model with 5 scale predictions has the best performance.

#of scales	DSC↑	Hausdorff ↓
3	$0.494_{\pm 0.103}$	$25.32_{\pm 6.22}$
4	$0.568_{\pm 0.081}$	$22.79_{\pm 5.54}$
5	$\mathbf{0.581}_{\pm 0.075}$	$\mathbf{21.95}_{\pm 5.36}$

scales. However, as our model relies on the feature extractor, fewer scales mean fewer layers/depths of our feature extractor network, which reduces the receptive field of the feature and gives less representative features. Hence, we can observe a considerable decrease in the performance with three scales prediction. Similarly, with more scales, we have a deeper feature extractor and more intermediate constraints, which leads to an even better performance of around 0.581 Dice score. For all experiments, we use the same $\lambda_{DT} = 2$. PC-Reg achieves the best performance with five scale predictions on the ABD50 dataset. As the number of scales that we can obtain depends on the image size, we use 4 scales in all other experiments for general usage.

To have a clear comparison between the proposed PC-Reg-diff and the counterpart multi-scale method, LapIRN, we visualize the deformation results at each scale in Fig. 10. PC-Reg-diff consistently has a better performance in each scale. The errors of LapIRN (marked as **green**) in the low scale (1/4) are not corrected in the higher scales. In contrast, PC-Reg-diff has a considerably better prediction in the lowest scale, and the errors that are not revealed in the lowest scale are alleviated subsequently (marked as **red**). This implies that PC-Reg-diff is able to reduce the accumulated integration error and achieves the best performance.

4.7. Model size

To have a fair comparison, we increased the number of parameters of VoxelMorph and LapIRN (indicated with $*$) and decreased the parameters of the proposed method (indicated with "-small").

Table A.8
Average inference time for each method on one test pair. Learning-based methods are tested on a GPU and have a superior inference speed compared with conventional methods. PC-Reg is able to do the inference within one second.

Methods	Time (s)
Syn	725.732
NiftReg	24.275
VoxelMorph	0.128
TransMorph	0.119
LapIRN	0.006
PC-Reg	0.438

PC-Reg-small and PC-Reg-small-diff can maintain competitive performance with considerably reduced model sizes while outperforming other baseline methods with similar model sizes, which is shown in Fig. 11.

5. Discussion and conclusion

In this paper, we proposed a pyramidal framework, PC-Reg, for deformable medical image registration, which is particularly apt for large deformations. PC-Reg is based on pyramidal feature extraction and uses prediction and correction networks for deformation vector field estimation in each scale. Our method progressively predicts a residual vector field to update the deformation vector field and a correction vector field to refine it through different scales. Furthermore, we introduce a distillation loss to provide self-guidance using the full-resolution prediction during training. PC-Reg can also be extended to allow for diffeomorphic registration (PC-Reg-diff) to obtain a smooth and invertible deformation vector field. PC-Reg-diff has a superior advantage on the large deformation dataset. The experimental results show that our method performs well on OASIS and ABD50 datasets.

Although the proposed PC-Reg and the diffeomorphic variant have a good performance, it has a few limitations. Firstly, we only conducted inter-patient registration experiments. Also, deformation accuracy is only evaluated on contours with limited anatomical regions. In the future, we will evaluate our method on comprehensive real-world medical data, including intra-patient data and data with anatomical key points. Secondly, PC-Reg uses the concatenation operation to combine two features for predicting the deformation vector field because of memory limitation. It would be interesting to investigate using attention modules to efficiently and comprehensively exploit the spatial correlation between two features.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Efstratios Gavves reports financial support was provided by Elekta AB. Efstratios Gavves reports financial support was provided by University of Amsterdam. Jan Jakob Sonke reports a relationship with Elekta AB that includes: funding grants. Wenzhe Yin reports a relationship with University of Amsterdam that includes: employment.

Data availability

We use public data have provided information about the public data we use in the paper. Regarding the code, code-sharing requests will be handled on an individual basis.

Acknowledgments

This work was financially supported by RVO, The Netherlands funding with grant number PPS2102 and Elekta Oncology AB, Sweden.

Appendix A. Inference time

In Table A.8, we report the average inference time of each method on ABD50. We do inference on one test pair from ABD50 for 10 times and calculate the average inference time in seconds. Learning-based methods have superior inference speed compared to conventional methods. Also, PC-Reg can predict the deformation vector field within one second.

Appendix B. Performance on each sub-region

In this section, we report the unsupervised performance of the proposed method and the baseline methods on each sub-region of OASIS, Mindboggle-101, and Learn2Reg Abdomen CT-CT datasets (see Figs. B.12–B.14).

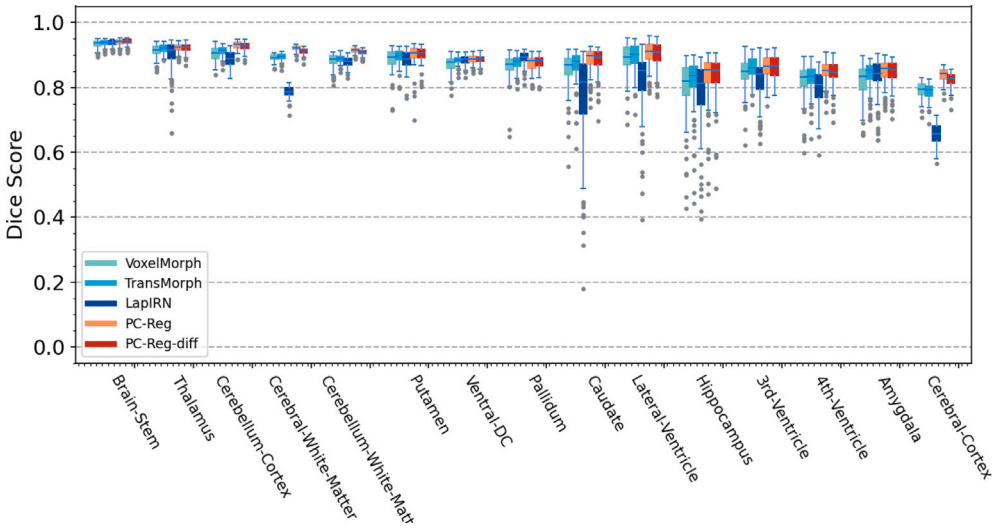


Fig. B.12. Quantitative performance of different methods on each sub-region of OASIS dataset. Following Balakrishnan et al. (2019), we select and group the sub-regions into 15 regions for visualization. PC-Reg outperforms other baselines on most sub-regions.

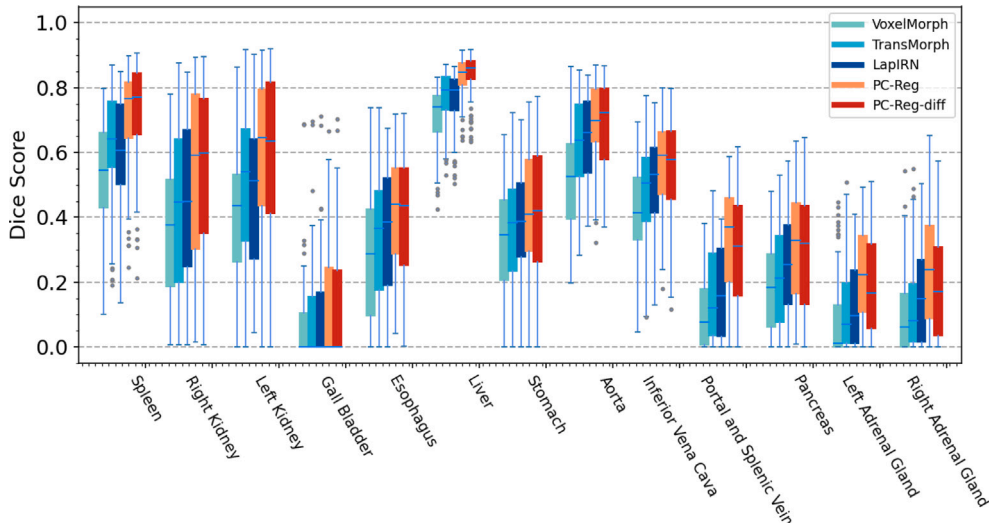


Fig. B.13. Quantitative performance of different methods on each sub-region of Learn2Reg Abdomen CT-CT dataset. PC-Reg outperforms other baselines on all the sub-regions.

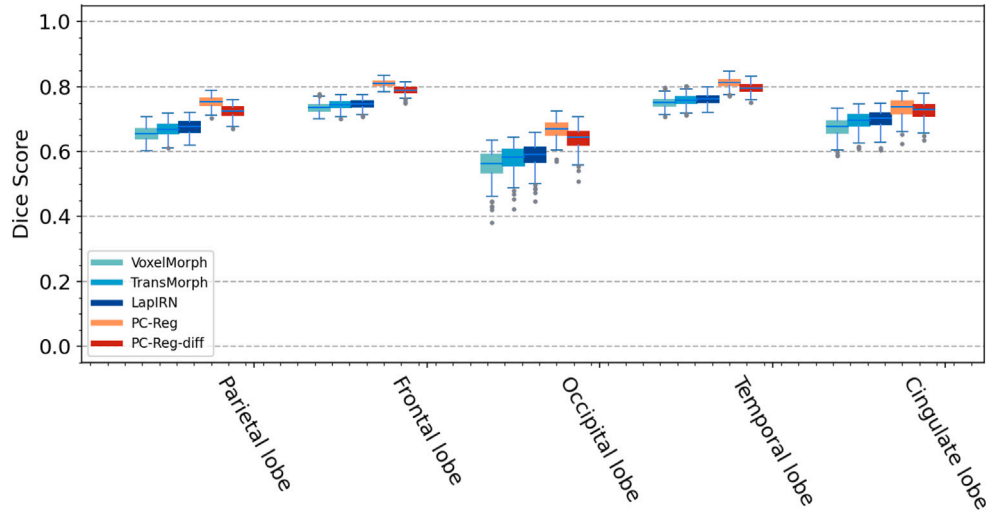


Fig. B.14. Quantitative performance of different methods on each sub-region of Mindboggle-101 dataset. PC-Reg outperforms other baselines on all the sub-regions.

References

Arsigny, V., Commowick, O., Pennec, X., Ayache, N., 2006. A log-euclidean framework for statistics on diffeomorphisms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 924–931.

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 (1), 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 (3), 2033–2044.

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.

Beekman, C., Schaake, E., Sonke, J.-J., Remeijer, P., 2021. Deformation trajectory prediction using a neural network trained on finite element data—application to library of CTVs creation for cervical cancer. Phys. Med. Biol. 66 (21), 215004.

Beg, M.F., Miller, M.I., Trouvé, A., Younes, L., 2005. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vis. 61 (2), 139–157.

Butcher, J.C., 2016. Numerical Methods for Ordinary Differential Equations. John Wiley & Sons.

Cao, Y., Miller, M.I., Winslow, R.L., Younes, L., 2005. Large deformation diffeomorphic metric mapping of vector fields. IEEE Trans. Med. Imaging 24 (9), 1216–1230.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2021. Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537.

Cao, X., Yang, J., Zhang, J., Nie, D., Kim, M., Wang, Q., Shen, D., 2017. Deformable image registration based on similarity-steered CNN regression. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 300–308.

Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y., 2022. Transmorph: Transformer for unsupervised medical image registration. Med. Image Anal. 82, 102615.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

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

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.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

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., Pluim, J.P., 2018. Pulmonary CT registration through supervised learning with convolutional neural networks. IEEE Trans. Med. Imaging 38 (5), 1097–1105.

Fischl, B., 2012. FreeSurfer. Neuroimage 62 (2), 774–781.

Greer, H., Kwitt, R., Vialard, F.-X., Niethammer, M., 2021. Icon: Learning regular maps through inverse consistency. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3396–3405.

Heinrich, M.P., Hansen, L., 2020. Highly accurate and memory efficient unsupervised learning-based discrete CT registration using 2.5 D displacement search. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 190–200.

Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, M., Schnabel, J.A., 2012. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. Med. Image Anal. 16 (7), 1423–1435.

Hering, A., Hansen, L., Mok, T.C.W., Chung, A.C.S., Siebert, H., Häger, S., Lange, A., Kuckertz, S., Heldmann, S., Shao, W., Vesal, S., Rusu, M., Sonn, G., Estienne, T., Vakalopoulou, M., Han, L., Huang, Y., Yap, P.-T., Brudfors, M., Balbastre, Y., Joutard, S., Modat, M., Lifshitz, G., Raviv, D., Lv, J., Li, Q., Jaouen, V., Visvikis, D., Fourcade, C., Rubeaux, M., Pan, W., Xu, Z., Jian, B., De Benetti, F., Wodzinski, M., Gunnarsson, N., Sjölund, J., Grzech, D., Qiu, H., Li, Z., Thorley, A., Duan, J., Großbröhmer, C., Hoopes, A., Reinertsen, I., Xiao, Y., Landman, B., Huo, Y., Murphy, K., Lessmann, N., Van Ginneken, B., Dalca, A.V., Heinrich, M.P., 2022a. Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. IEEE Trans. Med. Imaging.

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. Springer, pp. 382–390.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.

Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K., 2015. Spatial transformer networks. Adv. Neural Inf. Process. Syst. 28.

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. 71, 102036.

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

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P., 2009. Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging 29 (1), 196–205.

Klein, A., Tourville, J., 2012. 101 Labeled brain images and a consistent human cortical labeling protocol. Front. Neurosci. 6, 171.

Kuang, D., Schmah, T., 2019. Faim–a convnet method for unsupervised 3d medical image registration. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 646–654.

Lapidus, L., Seinfeld, J.H., 1971. Numerical Solution of Ordinary Differential Equations. Academic Press.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Lu, Y., Valmadre, J., Wang, H., Kannala, J., Harandi, M., Torr, P., 2020. Devon: Deformable volume network for learning optical flow. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2705–2713.

Luo, K., Wang, C., Liu, S., Fan, H., Wang, J., Sun, J., 2021. Upflow: Upsampling pyramid for unsupervised optical flow learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1045–1054.

Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., Cao, S., Zhang, Q., Liu, S., Wang, Y., Li, Y., He, J., Yang, X., 2021. AbdomenCT-1K: Is abdominal organ segmentation a solved problem? IEEE Trans. Pattern Anal. Mach. Intell. http://dx.doi.org/10.1109/TPAMI.2021.3100536.

Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. 19 (9), 1498–1507.

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.

Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S., 2010. Fast free-form deformation using graphics processing units. Comput. Methods Programs Biomed. 98 (3), 278–284.

Mok, T.C., Chung, A., 2020. Large deformation diffeomorphic image registration with laplacian pyramid networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 211–221.

Pegios, P., Czolbe, S., 2022. Can transformers capture long-range displacements better than CNNs? In: Medical Imaging with Deep Learning.

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. Springer, pp. 234–241.

Rühaak, J., Polzin, T., Heldmann, S., Simpson, I.J., Handels, H., Modersitzki, J., Heinrich, M.P., 2017. Estimation of large motion in lung CT by integrating regularized keypoint correspondences into dense deformable registration. IEEE Trans. Med. Imaging 36 (8), 1746–1757.

Shen, Z., Han, X., Xu, Z., Niethammer, M., 2019a. Networks for joint affine and non-parametric image registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4224–4233.

Shen, Z., Vialard, F.-X., Niethammer, M., 2019b. Region-specific diffeomorphic metric mapping. Adv. Neural Inf. Process. Syst. 32.

Shi, J., He, Y., Kong, Y., Coatrieux, J.-L., Shu, H., Yang, G., Li, S., 2022. XMorpher: Full transformer for deformable medical image registration via cross attention. arXiv preprint arXiv:2206.07349.

Sokooti, H., Vos, B.d., 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. Springer, 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.

Teed, Z., Deng, J., 2020. Raft: Recurrent all-pairs field transforms for optical flow. In: European Conference on Computer Vision. Springer, pp. 402–419.

Tian, L., Greer, H., Vialard, F.-X., Kwitt, R., Estépar, R.S.J., Rushmore, R.J., Makris, N., Bouix, S., Niethammer, M., 2023. GradICON: Approximate diffeomorphisms via gradient inverse consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18084–18094.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Viola, P., Wells III, W.M., 1997. Alignment by maximization of mutual information. Int. J. Comput. Vis. 24 (2), 137–154.

Vos, B.D.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. Springer, pp. 204–212.

Wolberg, G., Zokai, S., 2000. Robust image registration using log-polar transform. In: Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101), Vol. 1. IEEE, pp. 493–496.

Xie, Y., Zhang, J., Shen, C., Xia, Y., 2021. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 171–180.

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

Yin, W., Chagin, V.O., Cardoso, M.C., Rohr, K., 2022. Non-rigid registration of temporal live cell microscopy image sequences using deep learning. In: Medical Imaging 2022: Image Processing, Vol. 12032. SPIE, pp. 353–358.