Diffeomorphic Image Registration with Neural Velocity Field

Kun Han[1] Shanlin Sun[1] Xiangyi Yan[1] Chenyu You[2] Hao Tang[1]
Junayed Naushad[1] Haoyu Ma[1] Deying Kong[1] Xiaohui Xie[1]
[1]University of California, Irvine, USA [2] Yale University, USA
{khan7,shanlins,xiangyy4,htang6,jnaushad,haoyum3,deyingk,xhx}@uci.edu
{chenyu.you}@yale.edu

Abstract

Diffeomorphic image registration, offering smooth transformation and topology preservation, is required in many medical image analysis tasks. Traditional methods impose certain modeling constraints on the space of admissible transformations and use optimization to find the optimal transformation between two images. Specifying the right space of admissible transformations is challenging: the registration quality can be poor if the space is too restrictive, while the optimization can be hard to solve if the space is too general. Recent learning-based methods, utilizing deep neural networks to learn the transformation directly, achieve fast inference, but face challenges in accuracy due to the difficulties in capturing the small local deformations and generalization ability. Here we propose a new optimization-based method named DNVF (Diffeomorphic Image Registration with Neural Velocity Field) which utilizes deep neural network to model the space of admissible transformations. A multilayer perceptron (MLP) with sinusoidal activation function is used to represent the continuous velocity field and assigns a velocity vector to every point in space, providing the flexibility of modeling complex deformations as well as the convenience of optimization. Moreover, we propose a cascaded image registration framework (Cas-DNVF) by combining the benefits of both optimization and learning based methods, where a fully convolutional neural network (FCN) is trained to predict the initial deformation, followed by DNVF for further refinement. Experiments on two large-scale 3D MR brain scan datasets demonstrate that our proposed methods significantly outperform the state-of-the-art registration methods.

1. Introduction

Image registration is an essential task used in many medical image analysis applications [17, 32], such as assessing disease progression over time, merging and comparing different image modalities, and shape analysis. By maximizing the image similarity, such as intensity correlation, image registration provides the correspondence and non-linear transformation between pairs of images. Diffeomorphic image registration offers more desirable properties such as smooth deformation, topology preservation, and transformation invertibility.

Traditional methods, such as elastic-type models [4, 36], B splines [34], LDDMM [7, 60] and SyN [3], solve the image registration problem by optimizing the deformation fields. These methods typically make certain model assumptions. For example, LDDMM assumes the diffeomorphic deformation can be obtained by solving the flow-based ordinary differential equation (ODE) with certain regularization constraints. SyN symmetrizes the cross-correlation Euler-Lagrange equations within the space of diffeomorphic maps. However, these methods usually generate high-accuracy results at the cost of slow speed and intensive computation, and the performance may vary under different modeling assumptions [42].

Rapid advance in learning-based methods [5, 13, 35, 25, 24, 62, 37, 26] have achieved promising results in the image registration task. With deep neural networks, learning-based methods can efficiently estimate the transformation between two medical images. As more attention is focused on learning-based methods, many new techniques and complex network structures [9, 27, 16, 40] have been applied to chase better performance. However, the accuracy improvement is only modest, mainly because the representations learned from neural networks are not able to predict sophisticated deformations and dense correspondences for each pair of images in dataset. Moreover, the generalizability is still a major challenge for these methods which limits the performance for the out-of-distribution image pairs.

The recent development of neural fields provides a class of coordinate-based neural networks which parameterize the physical properties of objects across space and time [52]. Neural fields have shown their great potential in modeling general dynamic scenes [30, 20, 43] which fit the observed time-variant views with great detail through the op-

timization of a neural network. Therefore, our question of curiosity is hence: *can we use neural fields to represent the dynamics of diffeomorphic image registration?*

In this paper, we propose to realize diffeomorphic image registration by optimizing an implicit neural representation of a continuous velocity field. Specifically, we parameterize the continuous velocity field as a multilayer perceptron (MLP), whose input is a 3D spatial coordinate (x, y, z) and output is the corresponding 3D velocity vector (v_x, v_y, v_z). With periodic sinusoidal activation functions, the MLP can efficiently represent the high-frequency content [41] and therefore improve its ability to model the small and complex deformations in the registration problem. The diffeomorphic deformation can be obtained through integration over the neural velocity field, which is realized by the scaling and squaring (SS) method [1] in our work. SS follows the Lie algebra in group theory and the deformation is produced by the exponential of the velocity field through a spatial transform layer.

Moreover, we propose a cascaded framework called Cas-DNVF which combines the benefits of learning-based methods and DNVF. In the first stage, we pretrain a fully convolutional neural network to predict an initial deformation with a short inference time and simplify the search space of optimal deformation for the following DNVF. Based on that, DNVF can optimize a residual deformation specifically for each pair of images. By combining the benefits of two different methods, Cas-DNVF has better generalizability and can achieve the accurate alignment between two images within a short running time. Our experiments shows the DNVF can be integrated with different learning-based registration methods under the framework of Cas-DNVF.

Our contributions of this work are as follows:

- We propose neural registration method, called DNVF, for diffeomorphic image registration, utilizing MLP with sinusoidal activation to represent continuous velocity field and model diffeomorphic deformation through integral curves of velocity field. Optimal registration is discovered by tuning parameters of MLP.
- We further propose a cascaded framework (Cas-DNVF) to incorporate learning to DNVF, where a fully convolutional net is trained to predict the initial deformation, followed by DNVF for further refinement.
- Extensive experiments on two 3D brain MR datasets demonstrate that the proposed methods achieve state-of-the-art performance while preserving desirable diffeomorphic properties.

2. Related works

2.1. Pair-wise optimization method

Extensive works have been conducted to tackle the task of image registration. Traditional methods model image registration as an optimization problem and minimize the energy function iteratively for each pair of images [42]. These methods typically enforce transformation regularity through certain model assumptions. Several studies directly optimize the deformable displacement field including elastic-type models [4, 36], statistical parametric mapping [2], free-form deformations with b-splines [34] and optical flow based Demons [47]. Besides, many other studies focus on the registration problem within the space of diffeomorphic maps to ensure the desirable diffeomorphic properties, such as topology preservation and transformation invertibility. Popular methods include Large Deformation Diffeomorphic Metric Mapping (LDDMM) [7, 60] and symmetric image normalization method (SyN) [3]. LDDMM models the diffeomorphic deformation by considering the velocity over time according to the Lagrange transport equation [12, 14]. And SyN develops a novel symmetric diffeomorphic optimizer for maximizing the cross-correlation in the space of topology preserving maps [3].

The proposed DNVF is also a pair-wise optimization-based diffeomorphic image registration method, however, there are no strong assumptions about the dynamics of the registration since using a neural network to model the deformation provides greater flexibility. Moreover, DNVF can utilize deep learning packages for efficient inference and optimization.

2.2. Learning-based method

Medical researches have shown the promising progress brought by the recent learning methods [33, 10, 11, 45, 55, 59, 46, 58, 57, 56]. In image registration, learning-based methods [5, 13, 35, 38, 39, 26, 25, 24, 62, 9, 27, 16, 40, 61] achieve higher accuracy and efficiency. By learning a common representation for a collection of images, the extracted features can be used to perform registration with fast inference speed. VoxelMorph [5] directly regresses deformation fields by minimizing the dissimilarity between input and target images. The multi-resolution strategy was introduced in LapIRN [25] to avoid local minima during optimization. SYMNet [24] symmetrically warps images regarding the middle of the geodesic path and predicts the diffeomorphic deformation. A recursive cascaded network was proposed in [63] to boost the performance of registration by iteratively applying the registration network to the warped moving image and fixed image. A transformer block was deployed over the CNN backbone to capture the semantic contextual relevance in DTN [62]. To better solve large deformations, MS-ODENet [53] proposed to use neural ODE on image registration to refine the estimated transformation,

by modeling the dynamics of the parameters of registration models. However, the representation power of neural networks is limited by the network structure and training data might not be able to generate complex deformations and capture dense correspondences for all pairs of images in a dataset. The generalization gap between training data and test data also restricts the performance of a pre-trained neural network during inference time.

2.3. Neural Field

The recent advance of neural fields enables the parameterization of physical properties and dynamics through coordinate-based neural networks. [29] formulated the generative shape-conditioned 3D modeling with a continuous implicit surface. [41] introduced sinusoidal representation networks to model the 2D image and 3D scene with fine details. [28] learned a temporally and spatially continuous vector field to perform dense 4D reconstruction from images or sparse point clouds. [50] proposed to perform high-resolution MR image reconstruction via implicit neural representation. [43] applied neural field to model the diffeomorphic transformation on 3D shapes. Some recent works focus on pair-wise image registration problems, such as IDIR [49] and NODEO [51].

Unlike them, DNVF uses a simple multilayer perceptron to represent the continuous neural velocity field and model the diffeomorphic deformation. The proposed Cas-DNVF further combines the benefit of learning-based and optimization-based methods with better generalizability, matching accuracy and time efficiency. Experimental results in Section 5 demonstrate the advantages of our proposed methods.

3. Preliminaries

3.1. Deformable registration

Deformable image registration denotes warping one (moving) image to align it with the second (fixed or target) image by maximizing the similarity between the registered images under some regularization constraints. The displacement field returned from the registration defines the dense mapping between points in the moving image and corresponding points in the fixed image. The typical deformable image registration can be formulated as :

$$\phi^* = \arg\min_{\phi} \mathcal{L}_{\text{sim}}\left(I_f, \phi \circ I_m\right) + \mathcal{L}_{reg}(\phi) \tag{1}$$

where ϕ^* represents the optimal displacement field ϕ, I_f and I_m denote the fixed and moving images, $\phi \circ I_m$ represents I_m warped by ϕ, $\mathcal{L}_{sim}$ measures the image similarity between the fixed image and warped image, and $\mathcal{L}_{reg}$ represents the smoothness regularization function.

3.2. Diffeomorphic registration

Diffeomorphic image registration not only aligns two images but also preserves the topology and maintains transformation invertibility [7]. The diffeomorphic deformation ϕ is calculated through the integral of the velocity field $\boldsymbol{v}$ (assume Lipschitz continuity) following the ordinary differential equation (ODE):

$$\frac{\partial \boldsymbol{\phi}^{(t)}}{\partial t} = \boldsymbol{v}(\boldsymbol{\phi}^{(t)}) \tag{2}$$

where $\boldsymbol{\phi}^{(0)} = I$ is the identity transformation. In this paper, we assume the velocity field $\boldsymbol{v}$ is stationary over $t = [0, 1]$ and the final deformation is taken to be $\boldsymbol{\phi}^{(1)}$.

4. Method

Let I_f, I_m be the fixed image and moving image that need to be aligned. In this paper, we focus on 3D image registration where I_f and I_m are defined on 3D spatial domain $\Omega \subset \mathcal{R}^3$. I_f and I_m are affinely registered in the preprocessing step, therefore, we only need to model the non-linear displacement between two images.

Figure 1 presents an overview of our methods. DNVF is an optimization-based model which utilizes a MLP to represent the neural velocity field v_θ where θ are the parameters of the MLP. Unlike previous image registration methods, DNVF takes the 3D spatial coordinates as the input, rather than the image intensities. For each spatial point $\boldsymbol{p} \in \Omega$, v_θ provides the corresponding velocity vector $\boldsymbol{v} = v_\theta(\boldsymbol{p})$ at that point. The diffeomorphic deformation ϕ_θ is then calculated through the integral over the neural velocity field v_θ as described in Sec. 3.2. Inside DNVF, we use scaling and squaring to do the integration and the details will be described in Sec. 4.2. We optimize the parameters θ of the neural velocity field and find the optimal $\hat{\theta}$ by minimizing the loss function:

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}_{\text{sim}}\left(I_f, \phi_\theta \circ I_m\right) + \mathcal{L}_{reg}(\phi_\theta) \tag{3}$$

Based on DNVF, we propose a cascaded framework Cas-DNVF which combines the benefit of the learning-based methods and DNVF. First, we train a fully convolutional neural network (FCN) to model a function $g_\beta(I_f, I_m) = \phi^{init}$ which provides an initial deformation for a given pair of images. We train the g_β by minimizing the loss function similar to Eq.3:

$$\hat{\beta} = \arg\min_{\beta} [\mathbb{E}_{(I_f, I_m)\sim D}[\mathcal{L}_{\text{sim}}\left(I_f, g_\beta(I_f, I_m) \circ I_m\right) + \mathcal{L}_{reg}(g_\beta(I_f, I_m))]] \tag{4}$$

where $\mathcal{D}$ is the dataset distribution, β are the parameters of the FCN, I_f and I_m are the sampled volume pairs from

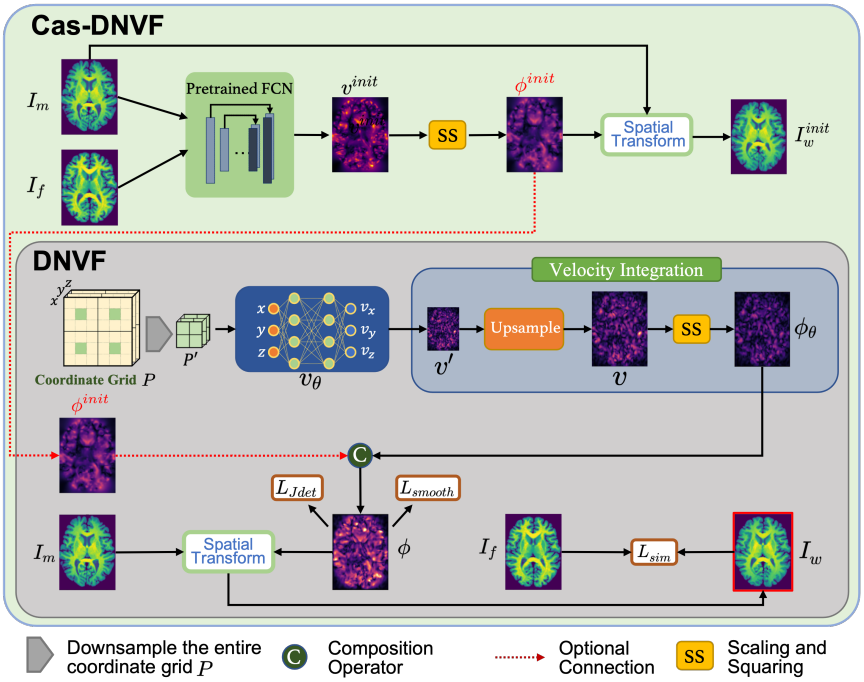

Figure 1. Overall framework of the proposed DNVF and Cas-DNVF. I_m, I_f and I_w denote the moving image, fixed image and warped image. DNVF uses neural representations to model continuous velocity field v_θ which takes 3D position $\boldsymbol{p} \in \mathbb{R}^3$ as input and assigns corresponding velocity vector $\boldsymbol{v} \in \mathbb{R}^3$. The velocity field v_θ is represented by a MLP with periodic sinusoidal activation functions. The velocity integration inside DNVF will be introduced in Sec.4.2. During the optimization, DNVF implicitly captures the dense correspondence between two input images. Cas-DNVF combines learning-based method and DNVF. A FCN is firstly pretrained to predict the initial deformation and simplify the search space of optimal deformation for the following DNVF so that it can focus on modeling the small local deformation with high accuracy and efficiency. The framework is designed for 3D registration, we use 2D image for simplicity.

$\mathcal{D}$. During inference, for the input image pair (I_f, I_m), we first use the pretrained g_β to predict an initial deformation ϕ^{init}. In the second stage, we use DNVF to optimize a residual deformation ϕ_θ^{res} where the overall deformation is calculated using a spatial transform layer to combine ϕ^{init} and ϕ_θ^{res}:

$$\hat{\theta} = \arg\min_\theta \mathcal{L}_{\text{sim}}\left(I_f, \phi_\theta^{res} \circ \phi^{init} \circ I_m\right) + \mathcal{L}_{reg}(\phi_\theta^{res} \circ \phi^{init}) \tag{5}$$

By analyzing the performance shown in Section 5, the initial deformation ϕ^{init} predicted by g_β helps DNVF to achieve faster convergence while alleviating the generalizability issue of learning-based method and providing a more precise dense matching between the pair of input images.

4.1. Neural Velocity Field Representation

Inspired by recent works on neural rendering [41, 22], we model the neural representation of the continuous velocity field v_θ using a MLP where θ are the parameters of the neural network. The neural velocity field can be viewed as a function of a spatial 3D point: $\boldsymbol{v} = v_\theta(\boldsymbol{p})$, which outputs the corresponding velocity vector $\boldsymbol{v}$ given 3D spatial coordinate $\boldsymbol{p}$.

In this paper, we focus on the diffeomorphic registration between two 3D brain MR scans. The complex structure of the human brain requires sophisticated deformation field ϕ_θ to achieve a precise matching. Therefore, we expect the neural velocity field v_θ to be able to model the high frequency function, otherwise, the deformation ϕ_θ integrated over v_θ will be too smooth and can not provide an accurate mapping. However, classic MLPs have difficulty learning high frequency functions because of the "spectral bias" [44]. As shown in [31, 6], the deep networks have a learning bias towards low frequency functions.

Therefore in this work, instead of using classic ReLU activation function, we choose to use periodic sinusoidal function enabling fitting of high-frequency content as indicated by SIREN [41] and adopt its weight initialization scheme for deep structure. The neural velocity field v_θ is designed as follows:

$$\begin{aligned} f_0(\mathbf{x}_0) &= \mathbf{W}_0\mathbf{x}_0 + \mathbf{b}_0 \mapsto x_1 \\ f_i(\mathbf{x}_i) &= \mathbf{W}_i \sin(\mathbf{x}_i) + \mathbf{b}_i \mapsto x_{i+1} \end{aligned} \tag{6}$$

where the 3D coordinate $\boldsymbol{p}$ is firstly mapped to a high dimensional embedding by $f_0 : \mathbb{R}^3 \mapsto \mathbb{R}^N$. Then the i^{th} layer of network f_i can be viewed as a Fourier frequency mapping [8] with the learnable parameters $\mathbf{W}_i$ and $\mathbf{b}_i$. By involving the frequency information in the network, the neural tangent kernel [44] is modified accordingly such that the neural velocity field v_θ can have a good representation of

the high-frequency details and be able to model the sophisticated deformation ϕ_θ. In our implementation, we use a 5-layer MLP to represent the neural velocity field v_θ with 512 hidden units.

4.2. Diffeomorphic Deformation as Integration

Inside DNVF, after defining the neural velocity field v_θ, the diffeomorphic deformation ϕ_θ is calculated by integrating over v_θ according to (2). Following [13, 24, 25, 62], the velocity field is assumed to be stationary over time. We use the scaling and squaring method to calculate the diffeomorphic deformation as shown in Figure 1:

Scaling and Squaring [1] When the velocity field is stationary, the exponential map $\boldsymbol{\phi}^{(t)} = \exp(v_\theta)$ defines one-parameter subgroup of diffeomorphisms. The final deformation $\boldsymbol{\phi}^{(1)} = \exp(v_\theta)$ can be solved more efficiently by utilizing the group actions. Specifically, the initial deformation is $\phi^{(1/2^T)} = \boldsymbol{p} + v_\theta(\boldsymbol{p})/2^T$ where T is the total time step. We recursively compute $\phi^{(1/2^{t-1})} = \phi^{(1/2^t)} \circ \phi^{(1/2^t)}$ through a spatial transform layer, and the final deformation $\phi^{(1)}$ is obtained by $\phi^{(1)} = \phi^{(1/2)} \circ \phi^{(1/2)}$. No additional learnable parameters are introduced during the recursive operation. The deformations are calculated over a fixed grid with linear interpolation. The entire step is differentiable. In our implementation, we chose $T = 7$.

It is infeasible to feed all 3D coordinates ($D \times H \times W \times 3$) into DNVF due to GPU memory constraints. Therefore in our implementation, we empirically downsample the original coordinate grid with a scale 1/3 and upsample the resulting velocity field to recover the full resolution for velocity integration as shown in Figure 1.

4.3. Cascaded Registration

Initial Deformation As shown in Figure 1, we parameterize the function $g_\beta(I_f, I_m) = \phi^{init}$ in (4) with a fully convolutional neural network (FCN), a scaling and squaring layer, and a spatial transform layer. FCN adopts the Unet-like network structure as shown in Figure 2 which takes the concatenation of moving image and fixed image as input, and directly outputs the velocity field. By maximizing the similarity between the warped image and fixed image, the FCN is trained to predict a initial deformation ϕ^{init}.

Moreover, g_β can also be parameterized by other learning-based models such as [13, 24, 25, 62]. Ablation study in Sec. 5.5 shows that DNVF can provide consistent boost in performance for different learning-based methods.

Optimization of Residual Deformation Based on the initial deformation predicted by the trained g_θ, we use DNVF to optimize the residual deformation ϕ_θ^{res} for each pair of images following (5). The ϕ^{init} and ϕ_θ^{res} are combined by a spatial transform layer as the overall deformation ϕ.

Because the predicted ϕ^{init} usually provides a good mapping for large deformations, DNVF will focus more on the local small deformations which are difficult for learning-based methods. The velocity vector $\boldsymbol{v}$ output from v_θ accounts for the local small deformations between the warped moving image ($\phi^{init} \circ I_m$) and fixed image I_f. In our implementation, instead of directly combining two deformations, we empirically rescale the velocity vector $\boldsymbol{v}$ in DNVF by a factor 0.1 to improve the stability during optimization.

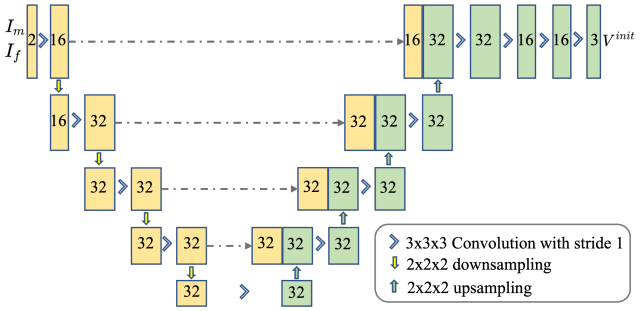

Figure 2. Structure of fully convolutional network (FCN) used in Cas-DNVF. The input is the concatenation of moving image and fixed image, and the output is the velocity field.

4.4. Optimization

The loss function $\mathcal{L}$ consists of two components: $\mathcal{L}_{sim}$ penalizing the misalignment between the warped moving image I_w and fixed image I_f, $\mathcal{L}_{reg}$ regularizing the deformation smoothness ($\mathcal{L}_{smooth}$) and local orientation consistency ($\mathcal{L}_{Jdet}$).

$\mathcal{L}_{sim}$ We use local normalized cross-correlation (NCC) as the metric to measure the similarity between the warped moving image I_w and fixed image I_f:

$$NCC(I_w, I_f) = \sum_{\boldsymbol{p} \in \Omega} \frac{\left(\sum_{\boldsymbol{p}_i} \left(I_w\left(\boldsymbol{p}_i\right) - \bar{I}_w(\boldsymbol{p})\right)\left(I_f\left(\boldsymbol{p}_i\right) - \bar{I}_f(\boldsymbol{p})\right)\right)^2}{\sum_{\boldsymbol{p}_i} \left(I_w\left(\boldsymbol{p}_i\right) - \bar{I}_w(\boldsymbol{p})\right)^2 \sum_{\boldsymbol{p}_i} \left(I_f\left(\boldsymbol{p}_i\right) - \bar{I}_f(\boldsymbol{p})\right)^2} \tag{7}$$

where $I_w = \phi \circ I_m$, ϕ is calculated by the velocity integration via SS, $\boldsymbol{p}_i$ denotes the points over a local window of size n^3 around $\boldsymbol{p}$, $\bar{I}_w(\boldsymbol{p})$ and $\bar{I}_f(\boldsymbol{p})$ are the mean intensity of that local window. High value of NCC represents a precise matching between images. Therefore, we use negative NCC as the similarity loss: $\mathcal{L}_{sim} = -NCC(\phi \circ I_m, I_f)$ with the window size set to 9.

$\mathcal{L}_{Jdet}$ In order to secure local orientation consistency, we follow [24] to impose a selective Jacobian determinant regularization. If the Jacobian determinant at a given point $\boldsymbol{p}$ is positive, then the deformation field preserves the orientation near $\boldsymbol{p}$. Otherwise, the orientation in the neighborhood is reversed and the topology is destroyed. With a ReLU function, we can penalize the local region with a negative Jacobian determinant:

$$\mathcal{L}_{Jdet} = \frac{1}{N} \sum_{\boldsymbol{p} \in \Omega} relu\left(-|J_\phi(\boldsymbol{p})|\right) \tag{8}$$

where the Jacobian matrix J_ϕ is defined as:

$$J_\phi(\boldsymbol{p}) = \begin{bmatrix} \frac{\partial \phi_x(\boldsymbol{p})}{\partial x} & \frac{\partial \phi_x(\boldsymbol{p})}{\partial y} & \frac{\partial \phi_x(\boldsymbol{p})}{\partial z} \\ \frac{\partial \phi_y(\boldsymbol{p})}{\partial x} & \frac{\partial \phi_y(\boldsymbol{p})}{\partial y} & \frac{\partial \phi_y(\boldsymbol{p})}{\partial z} \\ \frac{\partial \phi_z(\boldsymbol{p})}{\partial x} & \frac{\partial \phi_z(\boldsymbol{p})}{\partial y} & \frac{\partial \phi_z(\boldsymbol{p})}{\partial z} \end{bmatrix} \tag{9}$$

$\boldsymbol{\mathcal{L}_{smooth}}$ In order to avoid oddly skewed deformations, a spatial gradient is used to constrain the smoothness of the deformation field as a regularization term. A large spatial gradient means the radical change of deformation in the local area which is not desired in the registration problem. Therefore, the smoothness loss is defined as:

$$\mathcal{L}_{smooth} = \sum_{\boldsymbol{p} \in \Omega} \|\nabla \phi(\boldsymbol{p})\|^2 \tag{10}$$

We present the complete loss function as follows, where λ_1 and λ_2 control the weight of orientation consistency loss and deformation smoothness loss:

$$\mathcal{L} = \mathcal{L}_{sim} + \mathcal{L}_{reg} = \mathcal{L}_{sim} + \lambda_1 \mathcal{L}_{Jdet} + \lambda_2 \mathcal{L}_{smooth} \tag{11}$$

5. Experiment

5.1. Dataset and Preprocessing

We evaluate our method on two public 3D brain MR datasets: the OASIS [21] and the Mindboggle101 [19]. **OASIS** dataset contains 416 T1-weighted MR scans aging from 18 to 96 with 100 of them diagnosed with mild to moderate Alzheimer's disease. Subcortical segmentation maps of 35 anatomical structures serve as the ground truth for the evaluation of our method. **Mindboggle101** consists of 101 T1-weighted MR scans from 5 datasets, e.g. HLN-12, MMRR-21 and NKI-RS. We followed [54] to remove the images with incorrect labels and evaluated the performance on 31 cortical regions. Standard preprocessing methods were carried out on two datasets. Having skull stripped, all scans were resampled to same resolution $(1mm \times 1mm \times 1mm)$. For each dataset, images were aligned to MNI 152 space by affine transformation. The final images were cropped to size $(162 \times 192 \times 144)$ and normalized by the maximum intensity of each volume.

5.2. Experimental Setting

The proposed method is evaluated on the atlas-based image registration task same as [24]. We compare our method with traditional optimization-based methods: SyN[3] and NiftyReg[23], and state-of-the-art learning-based methods VoxelMorph(VM)[5] and SYMNet[24]. For each dataset, we randomly sampled 20 scans as moving images and 3 scans as the atlases, resulting in 60 image pairs, and evaluate the results using all anatomical labels. To make a fair comparison, we also conduct instance-specific optimization for the learning-based methods and compare the optimization results. We also compare DNVF with two recent independently proposed methods, IDIR [49] and NODEO [51] (released at roughly the same time as the preprint of this work) following NODEO's data setting. Both methods also use neural nets to model deformation but with some major differences: IDIR models deformation field instead of velocity field, while NODEO involves CNN.

We follow the parameter setting of SyN in VoxelMorph with gradient step size 0.25 and gaussian parameter [0.9, 0.2]. Both SyN and NiftyReg use cross-correlation as the cost function. For learning-based methods, we use 86 and 250 images as the training data for Mindboggle and OASIS dataset respectively. Different iteration settings were used to analyze the running time and performance of optimization-based methods.

During the optimization of DNVF, λ_1 and λ_2 are empirically set to 100 and 0.1 for the local orientation consistency and the smoothness of deformation. The FCN of Cas-DNVF is trained with λ_1 and λ_2 being 10 and 4. The network parameters are optimized using the Adam algorithm with a learning rate of $1e^{-4}$. Our model is implemented using PyTorch and evaluated on a machine with a RTX 2080 Ti GPU and an Intel i7-7700K CPU.

5.3. Evaluation Metric

The goal of diffeomorphic image registration is to generate spatial correspondences between pairs of images while maintaining the topology. We evaluate the performance of registration methods using Dice Similarity Coefficient, Jacobian Determinant, and the Structural Similarity following [13, 24, 25, 62]. Dice Similarity Coefficient (**DSC**) measures the overlap between the segmentation of the fixed image and the warped segmentation of the moving image based on the deformation field. Negative Jacobian Determinant ($|\boldsymbol{J_{<0}}|$) represents local distortion in the neighborhood as discussed in Sec. 4.4. We report the ratio of $|J_{<0}|$ to evaluate to performance of topology preservation. Structural Similarity (**SSIM**) [48] measures the similarity between fixed image and warped moving image by taking texture into account.

5.4. Results Comparison and Discussion

Accuracy The evaluation results of our methods, compared to both traditional optimization-based and learning-based methods, are summarized in Table 1. We report the results of SyN with iteration setting [600,600,300] and the results of NiftyReg with maximal level and iteration [5, 1000], which give better results than their default settings. All available anatomical masks are used in the evaluation. DNVF outperforms the traditional methods on Mindboggle and OASIS dataset in terms of DSC and SSIM, and still achieves the low ratio of $|J_{<0}|$. The evaluation results show the benefit of flexibility brought by neural velocity

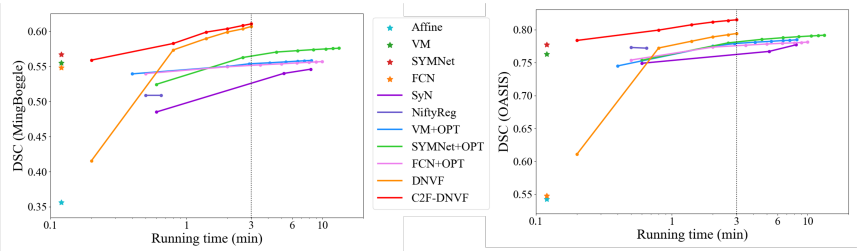

Figure 4. DSC versus Running time with different iteration setting.

implementation is not available. $|J_{<0}|$ reflects local distortion in deformation field. The diffeomorphism realized by velocity integration provides desired properties such as deformation smoothing and topology preservation for DNVF and NODEO. Therefore they have better performance in $|J_{<0}|$ than IDIR which doesn't realize diffeomorphism but models the displacement field. NODEO involves the convolutional layer and gaussian kernel to enhance the spatial interaction and reduce the ratio of $|J_{<0}|$. However, it provides the lowest matching accuracy in terms of DSC which is consistent with the result in Ablation study 2. Moreover, NODEO utilizes Neural ODE solver to do the integration requiring longer time to converge, which is not desired in real medical application. DNVF achieves the highest DSC score and lowest ratio of $|J_{<0}|$ which demonstrate the benefit of using MLP-based neural field and periodic sinusoidal activation functions to model the diffeomorphic deformation. Besides, DNVF only requires one-pass of the neural velocity field and utilizing scaling and squaring does not introduce additional learnable parameters, therefore it defines a simpler deformation space than NODEO and reduces the difficulty in finding the optimal solution via limited steps of optimization.

5.5. Ablation Study

In this section, we conduct ablation studies to measure the impact of components in our proposed method in terms of DSC which is averaged on all anatomical structures.

Activation function In Sec. 4.1, we state the benefit of using periodic sinusoidal activation to model small local deformation in diffeomorphic image registration. In this study, we conduct experiments on Mindboggle and OASIS datasets using classic ReLU activation with or without positional encoding (PE) as shown in Table 3. Though the fixed PE performed well in [22] with the supervision of ground truth, it didn't help to capture the dynamic dense correspondence and deformation field in unsupervised image registration task. The results show that the nested sinusoidal activation function provides better capacity in capturing the correspondence and modeling deformation.

Dataset	DNVF	MLP+ReLU	MLP+ReLU+PE
Mindboggle	**0.606**	0.440	0.397
OASIS	**0.794**	0.708	0.683

Table 3. Results using sinusoidal and ReLU activation functions.

Velocity field representation Table 4 shows the results with different velocity field representation: I) **Grid**: We use volumetric learnable parameter with size $(D\times H\times W\times 3)$ to represent the entire velocity field and update this parameter via optimization with additional regularization term (TV) suggested by [15]; II) **CNN**: Because the input to DNVF is a downsampled coordinate grid with shape $(\frac{D}{3}\times\frac{H}{3}\times\frac{W}{3}\times 3)$, we replace the MLP with a 3D convolution neural network. The evaluation results demonstrate that the fourier mapping expressed by the MLP in (6) provides the high deformation representation power and matching accuracy.

Dataset	DNVF	Grid	Conv+Sin	Conv+ReLU
Mindboggle	**0.606**	0.572	0.515	0.480
OASIS	**0.794**	0.755	0.739	0.731

Table 4. Results with different velocity field representations.

Choice of pretrained FCN for Cas-DNVF The proposed DNVF can also work with SOTA learning-based methods as shown in Table 5. In this study, we replace the original FCN with VoxelMorph(VM) and SYMNet. The results demonstrate that the Cas-DNVF is a generalized framework and provide consistent performance improvement to different learning-based methods.

Dataset	FCN+DNVF	VM+DNVF	SYMNet+DNVF
Mindboggle	**0.612**	0.609	0.606
OASIS	0.815	**0.816**	0.812

Table 5. Results with different pretrained FCN in Cas-DNVF.

6. Conclusion

In this paper, we propose a neural field model to represent the continuous velocity field and model deformation for solving diffeomorphic image registration. The validation experiments demonstrate the significant advantages brought by proposed methods. The proposed DNVF and Cas-DNVF methods offer a new framework for classical image registration problem. However, our current methods still have some limitations. First, the model has a relatively large memory footprint due to scaling and squaring component. A future direction is to improve the derivation of deformation field from velocity field. Second, the two stages of Cas-DNVF is decoupled. A future improvement is to integrate them and train the learning model in an end-to-end manner.

References

[1] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 924–931. Springer, 2006.

[2] John Ashburner and Karl J Friston. Voxel-based morphometry—the methods. *Neuroimage*, 11(6):805–821, 2000.

[3] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, 2008.

[4] Ruzena Bajcsy and Stane Kovačič. Multiresolution elastic matching. *Computer vision, graphics, and image processing*, 46(1):1–21, 1989.

[5] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9252–9260, 2018.

[6] Ronen Basri, Meirav Galun, Amnon Geifman, David Jacobs, Yoni Kasten, and Shira Kritchman. Frequency bias in neural networks for input of non-uniform density. In *International Conference on Machine Learning*, pages 685–694. PMLR, 2020.

[7] M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61(2):139–157, 2005.

[8] Nuri Benbarka, Timon Höfer, Andreas Zell, et al. Seeing implicit neural representations as fourier series. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2041–2050, 2022.

[9] Junyu Chen, Yufan He, Eric C Frey, Ye Li, and Yong Du. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. *arXiv preprint arXiv:2104.06468*, 2021.

[10] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[11] Xuming Chen, Shanlin Sun, Narisu Bai, Kun Han, Qianqian Liu, Shengyu Yao, Hao Tang, Chupeng Zhang, Zhipeng Lu, Qian Huang, et al. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. *Radiotherapy and Oncology*, 160:175–184, 2021.

[12] Gary E Christensen, Richard D Rabbitt, and Michael I Miller. Deformable templates using large deformation kinematics. *IEEE transactions on image processing*, 5(10):1435–1447, 1996.

[13] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 729–738. Springer, 2018.

[14] Paul Dupuis, Ulf Grenander, and Michael I Miller. Variational problems on flows of diffeomorphisms for image matching. *Quarterly of applied mathematics*, pages 587–600, 1998.

[15] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022.

[16] Jing Hu, Ziwei Luo, Xin Wang, Shanhui Sun, Youbing Yin, Kunlin Cao, Qi Song, Siwei Lyu, and Xi Wu. End-to-end multimodal image registration via reinforcement learning. *Medical Image Analysis*, 68:101878, 2021.

[17] Mariarosaria Incoronato, Marco Aiello, Teresa Infante, Carlo Cavaliere, Anna Maria Grimaldi, Peppino Mirabelli, Serena Monti, and Marco Salvatore. Radiogenomic analysis of oncological data: a technical survey. *International journal of molecular sciences*, 18(4):805, 2017.

[18] David N Kennedy, Christian Haselgrove, Steven M Hodge, Pallavi S Rane, Nikos Makris, and Jean A Frazier. Candishare: a resource for pediatric neuroimaging data, 2012.

[19] Arno Klein and Jason Tourville. 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in neuroscience*, 6:171, 2012.

[20] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021.

[21] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

[22] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020.

[23] Marc Modat, Gerard R Ridgway, Zeike A Taylor, Manja Lehmann, Josephine Barnes, David J Hawkes, Nick C Fox, and Sébastien Ourselin. Fast free-form deformation using graphics processing units. *Computer methods and programs in biomedicine*, 98(3):278–284, 2010.

[24] Tony CW Mok and Albert Chung. Fast symmetric diffeomorphic image registration with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4644–4653, 2020.

[25] Tony CW Mok and Albert Chung. Large deformation diffeomorphic image registration with laplacian pyramid networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 211–221. Springer, 2020.

[26] Tony CW Mok and Albert Chung. Conditional deformable image registration with convolutional neural network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 35–45. Springer, 2021.

[27] Tony CW Mok and Albert Chung. Affine medical image registration with coarse-to-fine vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20835–20844, 2022.

[28] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5379–5389, 2019.

[29] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.

[30] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

[31] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019.

[32] Petter Risholm, Alexandra J Golby, and William Wells. Multimodal image registration for preoperative planning and image-guided neurosurgical procedures. *Neurosurgery Clinics*, 22(2):197–206, 2011.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[34] Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach, and David J Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging*, 18(8):712–721, 1999.

[35] Ameneh Sheikhjafari, Michelle Noga, Kumaradevan Punithakumar, and Nilanjan Ray. Unsupervised deformable image registration with fully connected generative neural network. 2018.

[36] Dinggang Shen and Christos Davatzikos. Hammer: hierarchical attribute matching mechanism for elastic registration. *IEEE transactions on medical imaging*, 21(11):1421–1439, 2002.

[37] Zhengyang Shen, Jean Feydy, Peirong Liu, Ariel H Curiale, Ruben San Jose Estepar, Raul San Jose Estepar, and Marc Niethammer. Accurate point cloud registration with robust optimal transport. *Advances in Neural Information Processing Systems*, 34:5373–5389, 2021.

[38] Zhengyang Shen, Xu Han, Zhenlin Xu, and Marc Niethammer. Networks for joint affine and non-parametric image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4224–4233, 2019.

[39] Zhengyang Shen, François-Xavier Vialard, and Marc Niethammer. Region-specific diffeomorphic metric mapping. *Advances in Neural Information Processing Systems*, 32, 2019.

[40] Jiacheng Shi, Yuting He, Youyong Kong, Jean-Louis Coatrieux, Huazhong Shu, Guanyu Yang, and Shuo Li. Xmorpher: Full transformer for deformable medical image registration via cross attention. *arXiv preprint arXiv:2206.07349*, 2022.

[41] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.

[42] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. Deformable medical image registration: A survey. *IEEE transactions on medical imaging*, 32(7):1153–1190, 2013.

[43] Shanlin Sun, Kun Han, Deying Kong, Hao Tang, Xiangyi Yan, and Xiaohui Xie. Topology-preserving shape reconstruction and registration via neural diffeomorphic flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20845–20855, 2022.

[44] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020.

[45] Hao Tang, Xingwei Liu, Kun Han, Xiaohui Xie, Xuming Chen, Huang Qian, Yong Liu, Shanlin Sun, and Narisu Bai. Spatial context-aware self-attention model for multi-organ segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 939–949, 2021.

[46] Hao Tang, Xingwei Liu, Shanlin Sun, Xiangyi Yan, and Xiaohui Xie. Recurrent mask refinement for few-shot medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3918–3928, 2021.

[47] J-P Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical image analysis*, 2(3):243–260, 1998.

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[49] Jelmer M Wolterink, Jesse C Zwienenberg, and Christoph Brune. Implicit neural representations for deformable image registration. In *Medical Imaging with Deep Learning*, 2021.

[50] Qing Wu, Yuwei Li, Lan Xu, Ruiming Feng, Hongjiang Wei, Qing Yang, Boliang Yu, Xiaozhao Liu, Jingyi Yu, and Yuyao Zhang. Irem: High-resolution magnetic resonance image reconstruction via implicit neural representation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 65–74. Springer, 2021.

[51] Yifan Wu, Tom Z Jiahao, Jiancong Wang, Paul A Yushkevich, M Ani Hsieh, and James C Gee. Nodeo: A neural ordinary differential equation based optimization framework for deformable image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20804–20813, 2022.

[52] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, volume 41, pages 641–676. Wiley Online Library, 2022.

[53] Junshen Xu, Eric Z Chen, Xiao Chen, Terrence Chen, and Shanhui Sun. Multi-scale neural odes for 3d medical image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 213–223. Springer, 2021.

[54] Zhenlin Xu and Marc Niethammer. Deepatlas: Joint semi-supervised learning of image registration and segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 420–429. Springer, 2019.

[55] Xiangyi Yan, Hao Tang, Shanlin Sun, Haoyu Ma, Deying Kong, and Xiaohui Xie. After-unet: Axial fusion transformer unet for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3971–3981, 2022.

[56] Chenyu You, Weicheng Dai, Fenglin Liu, Haoran Su, Xiaoran Zhang, Lawrence Staib, and James S Duncan. Mine your own anatomy: Revisiting medical image segmentation with extremely limited labels. *arXiv preprint arXiv:2209.13476*, 2022.

[57] Chenyu You, Weicheng Dai, Lawrence Staib, and James S Duncan. Bootstrapping semi-supervised medical image segmentation with anatomical-aware contrastive distillation. *arXiv preprint arXiv:2206.02307*, 2022.

[58] Chenyu You, Jinlin Xiang, Kun Su, Xiaoran Zhang, Siyuan Dong, John Onofrey, Lawrence Staib, and James S Duncan. Incremental learning meets transfer learning: Application to multi-site prostate mri segmentation. *arXiv preprint arXiv:2206.01369*, 2022.

[59] Chenyu You, Ruihan Zhao, Fenglin Liu, Sandeep Chinchali, Ufuk Topcu, Lawrence Staib, and James S Duncan. Class-aware generative adversarial transformers for medical image segmentation. *arXiv preprint arXiv:2201.10737*, 2022.

[60] Miaomiao Zhang, Ruizhi Liao, Adrian V Dalca, Esra A Turk, Jie Luo, P Ellen Grant, and Polina Golland. Frequency diffeomorphisms for efficient image registration. In *International conference on information processing in medical imaging*, pages 559–570. Springer, 2017.

[61] Xiaoran Zhang, Chenyu You, Shawn Ahn, Juntang Zhuang, Lawrence Staib, and James Duncan. Learning correspondences of cardiac motion from images using biomechanics-informed modeling. *arXiv preprint arXiv:2209.00726*, 2022.

[62] Yungeng Zhang, Yuru Pei, and Hongbin Zha. Learning dual transformer network for diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 129–138. Springer, 2021.

[63] Shengyu Zhao, Yue Dong, Eric I Chang, Yan Xu, et al. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10600–10610, 2019.