Learning Deformable Image Registration From Optimization: Perspective, Modules, Bilevel Training and Beyond

Risheng Liu, *Member, IEEE*, Zi Li, Xin Fan, *Senior Member, IEEE*, Chenying Zhao, Hao Huang, and Zhongxuan Luo

Abstract—Conventional deformable registration methods aim at solving an optimization model carefully designed on image pairs and their computational costs are exceptionally high. In contrast, recent deep learning-based approaches can provide fast deformation estimation. These heuristic network architectures are fully data-driven and thus lack explicit geometric constraints which are indispensable to generate plausible deformations, e.g., topology-preserving. Moreover, these learning-based approaches typically pose hyper-parameter learning as a black-box problem and require considerable computational and human effort to perform many training runs. To tackle the aforementioned problems, we propose a new learning-based framework to optimize a diffeomorphic model via multi-scale propagation. Specifically, we introduce a generic optimization model to formulate diffeomorphic registration and develop a series of learnable architectures to obtain propagative updating in the coarse-to-fine feature space. Further, we propose a new bilevel self-tuned training strategy, allowing efficient search of task-specific hyper-parameters. This training strategy increases the flexibility to various types of data while reduces computational and human burdens. We conduct two groups of image registration experiments on 3D volume datasets including image-to-atlas registration on brain MRI data and image-to-image registration on liver CT data. Extensive results demonstrate the state-of-the-art performance of the proposed method with diffeomorphic guarantee and extreme efficiency. We also apply our framework to challenging multi-modal image registration, and investigate how our registration to support the down-streaming tasks for medical image analysis including multi-modal fusion and image segmentation.

Index Terms—Medical image analysis, diffeomorphic deformable registration, deep propagative network, bilevel self-tuned training

1 Introduction

Registration plays a critical role in medical image analysis, which transforms different images into one common coordinate system with matched contents by finding the spatial correspondence between images [1]. It is fundamental to many clinical tasks such as image fusion of different modalities, anatomical change diagnosis, motion extraction, and population modeling. Traditional image registration is formulated as an optimization problem to minimize image mismatching between a target and a warped source image, subject to transformation constraints. Deformable registration method computes a dense correspondence between image pairs [2]. The high degrees of freedom for the solution space (deformation maps) and great variations on source/target image pairs are major challenges for this issue.

Conventional deformable registration techniques aim at solving the optimization problem and offer rigorous theoretical treatments. However, the optimization is typically computationally expensive and time-consuming as the iterative process involves gradient computations over the high dimensional parameter and image spaces [3], [4], [5], [6]. Recent learning-based methods replace the costly numerical optimization with one step of prediction by learned deep networks so that they can provide fast deformation estimation. Balakrishnan *et al.* propose a UNet network structure, called VoxelMorph, to address deformable image registration [7]. Later, some researchers further combine the

- *Risheng Liu, Zi Li, and Xin Fan are with the DUT-RU International School of Information Science and Engineering, Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian University of Technology, Dalian, Liaoning 116024, China. E-mail: rsliu@dlut.edu.cn, alisonbrielee@gmail.com, xin.fan@ieee.org.*
- *Chenying Zhao is with the Department of Radiology, Children's Hospital of Philadelphia, Philadelphia, PA 19104 USA, and also with the Department of Bioengineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA 19104 USA. E-mail: chenyzh@seas.upenn.edu.*
- *Hao Huang is with the Department of Radiology, Children's Hospital of Philadelphia, Philadelphia, PA 19104 USA, and also with the Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104 USA. E-mail: huangh6@email.chop.edu.*
- *Zhongxuan Luo is with the DUT-RU International School of Information Science and Engineering, Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian University of Technology, Dalian, Liaoning 116024, China, and also with the Institute of Artificial Intelligence, Guilin University of Electronic Technology, Guilin 541004, China. E-mail: zxluo@dlut.edu.cn.*

Manuscript received 16 Jan. 2021; revised 8 July 2021; accepted 15 Sept. 2021. Date of publication 28 Sept. 2021; date of current version 3 Oct. 2022.
This work was supported in part by the National Key R&D Program of China under Grant 2020YFB1313503, in part by the National Natural Science Foundation of China under Grants 61922019, 61733002, and 61672125, in part by the LiaoNing Revitalization Talents Program under Grant XLYC1807088, and in part by the Fundamental Research Funds for the Central Universities.
(Corresponding author: Xin Fan).
Recommended for acceptance by J. Wang.
Digital Object Identifier no. 10.1109/TPAMI.2021.3115825

learning with the diffeomorphic constraint to provide topology-preserving deformations [8], [9]. These methods mostly learn parameters upon a pre-defined training loss to output deformation fields. Hence, it is difficult to adaptively include registration information for the front-end feature learning phase. Some existing coarse-to-fine approaches [10], [11], predicting deformation fields at different scales, may lead to more accurate registration and a controllable training procedure compared to those single scale approaches.

Regularization on hyper-parameters, *e.g.,* trade-off parameters, weight decay, and dropout, are crucial to the generalization of registration networks [7], [8]. The quality of output deformation fields for different deep networks highly depends on the choice of hyper-parameters. However, hyper-parameter choosing typically involves training many separate models with various hyper-parameter configurations, posing a significant computational challenge and potentially leading to sub-optimal results. For example, the grid search and random search work well only when ample computational resources are available. Generally, previous learning-based registration approaches pose hyper-parameter optimization as a black-box optimization problem, ignoring information that is important for faster convergence, thus require many training runs.

1.1 Our Contributions

To address the limitations of both optimization-based and learning-based approaches, we design a new deep propagation framework to optimize a diffeomorphic model via multi-scale propagation for deformable registration. First, we introduce a generic optimization model to formulate the diffeomorphic deformation problem. Rather than performing the optimization over the image domain, we learn a more discriminative feature space that handles deformations more powerfully. Then we employ deep modules to propagate deformation fields on the learned multi-scale feature space, efficiently optimizing the diffeomorphic energy. This optimization perspective differentiates our scheme from naively cascading deep networks in most existing learning-based approaches, and provides a computational interpretation of network architectures that guarantees diffeomorphism. Moreover, to tackle the inefficiency of hyper-parameter tuning, we introduce a bilevel self-tuned training strategy for our registration model. This bilevel training takes the hyper-parameter learning as the upper-level objective while formulates learning for model parameters as the lower-level objective. With the help of upper-level and lower-level objectives, the model parameters and hyper-parameters can be obtained collaboratively. The main contributions of this work can be summarized in the following aspects:

1) We establish a deep propagation framework to optimize the diffeomorphic registration energy on the learned multi-scale feature space. Circumventing expensive computations of iterative gradients on the image domain and performing the optimization over the discriminative feature space, this framework renders fast and efficient registration.
2) We develop the error-based data matching, context-based regularization and constraint modules to yield the propagating process and to design the training loss. Each module has physical inspiration or geometrical priors so that helps to obtain the solutions stably and controllably. Moreover, we may interpret our learning-based registration as the optimization of an energy with explicit diffeomorphism constraints.
3) We devise a new self-tuned training strategy that simultaneously learns optimal hyper-parameters of the loss function and network parameters of deep modules. We pose this joint training as bilevel optimization and propose an approximation algorithm to tackle its computation difficulties. This strategy allows flexible and efficient training rather than intensive labors of manually tuning in order to accommodate multiple types of medical data presenting significant variations.
4) Comprehensive evaluations on challenging multiple registration tasks of image-to-atlas, uni-modal and cross-modal image-to-image demonstrate that our approach achieves state-of-the-art performance. We also investigate the effectiveness of our approach to support the down-streaming medical image analysis including fusion and segmentation.

The paper is organized as follows. Section 2 describes related work. Section 3 introduces our optimization-inspired propagation framework and Section 4 describes our training strategies. We demonstrate experimental results in Section 5 and conclude the paper in Section 6.

2 Related Work

In this section, we describe registration methods based on different mechanisms, e.g, simple low-dimensional parametric models, model-based registration, prior-based registration, deep learning-based methods and related optical flow task.

Low-Dimensional Parametric Registration. Simple, low-dimensional parametric models, e.g., rigid [12], affine [13], or homography transformations [14], try to find a matrix that achieves the best possible agreement between a transformed source and a target image, which consists of 6 or 8 degrees of freedom. These parametric models usually serve as an initial alignment followed by more advanced, high-dimensional parametric or non-parametric registration models, such as deformable transformations that are with more degrees of freedom and the ability to capture subtle, localized images deformations. In our paper, we concentrate on the latter step, in which we compute a dense, nonlinear correspondence for all voxels and things become more complex and challenging.

Model-Based Registration. Physical models can be typically separated into elastic body models [15], viscous fluid flow models [16], diffusion models [17], curvature registration [18], and flows of diffeomorphisms [5]. In these cases, the transformation is governed by different types of Partial Differential Equation (PDE). These physical principle inspired methods generally ensure desirable properties such as inverse consistency and topology preservation. Among them, diffeomorphic frameworks [3], [4], [5], [19], [20] use smooth velocity fields to represent the deformation, have shown remarkable success in various computational anatomy applications. However, the methods estimate transformation through the solution of the PDE that can be computationally demanding and too complicated to work with.

Prior-Based Registration. It is possible to introduce prior knowledge about deformations when registration involves image acquisitions of specific anatomical organs, such as the tumor, prostate, breast, brain, lung and cardiac. Prior-based approaches that exploit our knowledge regarding the problem through the use of more informed priors at the cost of being constrained to well-defined settings, include statistically-constrained methods and biophysical model inspired methods. Statistically-constrained registration frameworks [21], [22], [23] can capture statistical information about deformation fields across a population of subjects but are generally limited by previously-observed deformations. The biomechanical inspired methods [24], [25] highly rely on complicated priors. With the complexity of different clinical scenes, it is difficult to introduce priors flexibly that ideally regularize the transformation.

Learning-Based Registration. Deep learning based methods [7], [8], [26], [27], [28], [29] taking advantage of neuron networks have shown impressive results, especially in terms of runtime. Inspired by the work of spatial transformer [30], plenty of works [7], [8], [9], [31], [32] have focused on replacing costly numerical optimization with global function optimization over the training data in an unsupervised way. Recently, some researchers [8], [9], [10], [11], [33] propose to estimate the velocity fields or momentum fields, which can be used to obtain diffeomorphic transformations. Typically, hyper-parameters such as regularization parameters in loss function exist in registration networks, which are crucial to the registration performance and generalization capability. However, learning-based approaches pose hyper-parameter learning as a black-box problem and require considerable computational and human effort to perform many training runs, especially when switching to other registration tasks.

Related Tasks. Optical flow estimation is a related registration problem for 2D images which returns a dense displacement vector field depicting small displacements between image pairs. The interest of optical flow estimation is to recover the apparent motion of objects between sequences of successive images, where its spatial correspondences/displacement field are associated with different time points. Learning-based optical flow approaches take a pair of images as input and use a convolutional neural network to learn the optical flow from data. Most of these works [27], [34], [35], [36] require supervision in the form of ground truth flow fields while using an unsupervised objective [37], [38] has emerged as a trend in recent. PWC-Net [34] and IRR [36] apply iterative refinement using coarse-to-fine pyramids. Other related applications include from tracking to depth prediction, stereo reconstruction and so on. However, on memory-costly 3D image registration tasks, things become more complex compared to those of 2D tasks. That is, designing effective architectures that perform better, train more easily and generalize well to novel scenes is more difficult and challenging, which is exactly the key contributions of this work.

3 Learning Registration From Optimization

To capture large deformations, diffeomorphic registrations have been frequently employed. These diffeomorphic methods [4], [10], [26], [39] have many desirable mathematical properties, such as invertibility, one-to-one smoothness, and topology-preserving. However, these physical model inspired methods [3], [4], [5], [19], [20] generally solve the optimization problem on the image domain, while the high dimensionality of the registration field parameters as well as the non-linear relationship between the images and the parameters pose a significant computational challenge. Prior-based methods [24], [24] introduce prior knowledge regarding the physical properties of the underlying anatomical structure. The informed priors may help to render the registration method more robust and stable, but it is challenging to introduce priors flexibly under different clinical scenes. These limitations make purely optimization-based registration methods hard to obtain the solutions efficiently and flexibly. The goal of this work is to learn a powerful solver to aggregate a variety of mechanisms to address the deformable registration problem efficiently. We first introduce a general optimization model to formulate deformable image registration, inspired by which we present our optimization propagation framework with a series of modules on multi-scale feature space.

Fundamental Optimization Formulation of Registration. Given a source image $\mathbf{s}$ and a target image $\mathbf{t}$ with a spatial domain $\Omega \in \mathbb{R}^d$, specifically, we aim at minimizing the following constrained optimization model:

$$\begin{aligned} &\min_{\boldsymbol{v}} \ \mathtt{Mat}(\boldsymbol{\varphi} \circ \mathbf{s}, \mathbf{t}) + \mathtt{Reg}(\boldsymbol{v}), \\ &s.t. \ \frac{\partial \boldsymbol{\phi}(t)}{\partial t} = \boldsymbol{v}(\boldsymbol{\phi}(t)), \ \boldsymbol{\phi}(0) = Id, \ \boldsymbol{\varphi} = \boldsymbol{\phi}(1), \end{aligned} \tag{1}$$

where the $\circ$ represents warping operation, $\boldsymbol{\varphi} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is the final deformation field, $\boldsymbol{v}$ is the stationary velocity fields for unit time. $\mathtt{Mat}(\cdot,\cdot)$ is data matching term, forcing the similarity of image pairs. $\mathtt{Reg}(\cdot)$ imposes regularization on the deformation and guarantees its smoothness, by constraining on the velocity fields $\boldsymbol{v}$. Govern by the ordinary differential equation constraint, $\boldsymbol{\phi}(0) = Id$ is the identity transformation, $t \in [0,1]$ represents the time, such that generating final registration field involves starting with an identity transform $\boldsymbol{\phi}(0) = Id$ and integrating of a stationary velocity field over unit time to obtain $\boldsymbol{\varphi} = \boldsymbol{\phi}(1)$. The time t is the time step for the integration process, and the stationary velocity field at each time step can be conceptualized as an integration with a single time step.

We try to bridge the correspondences between well-established principles in conventional methods and registration networks. As the first line of Fig. 1 shows, we unroll the optimization process on the discriminative multi-scale feature space, then design the error-based data matching, context-based regularization, and constraint modules, corresponding to data matching, regularization, and constraint in Eq. (1). Note that although our framework uses insights from conventional registration methods, the proposed modules do not *exactly* solve the corresponding optimization/energy. Next, we will elaborate on our different modules.

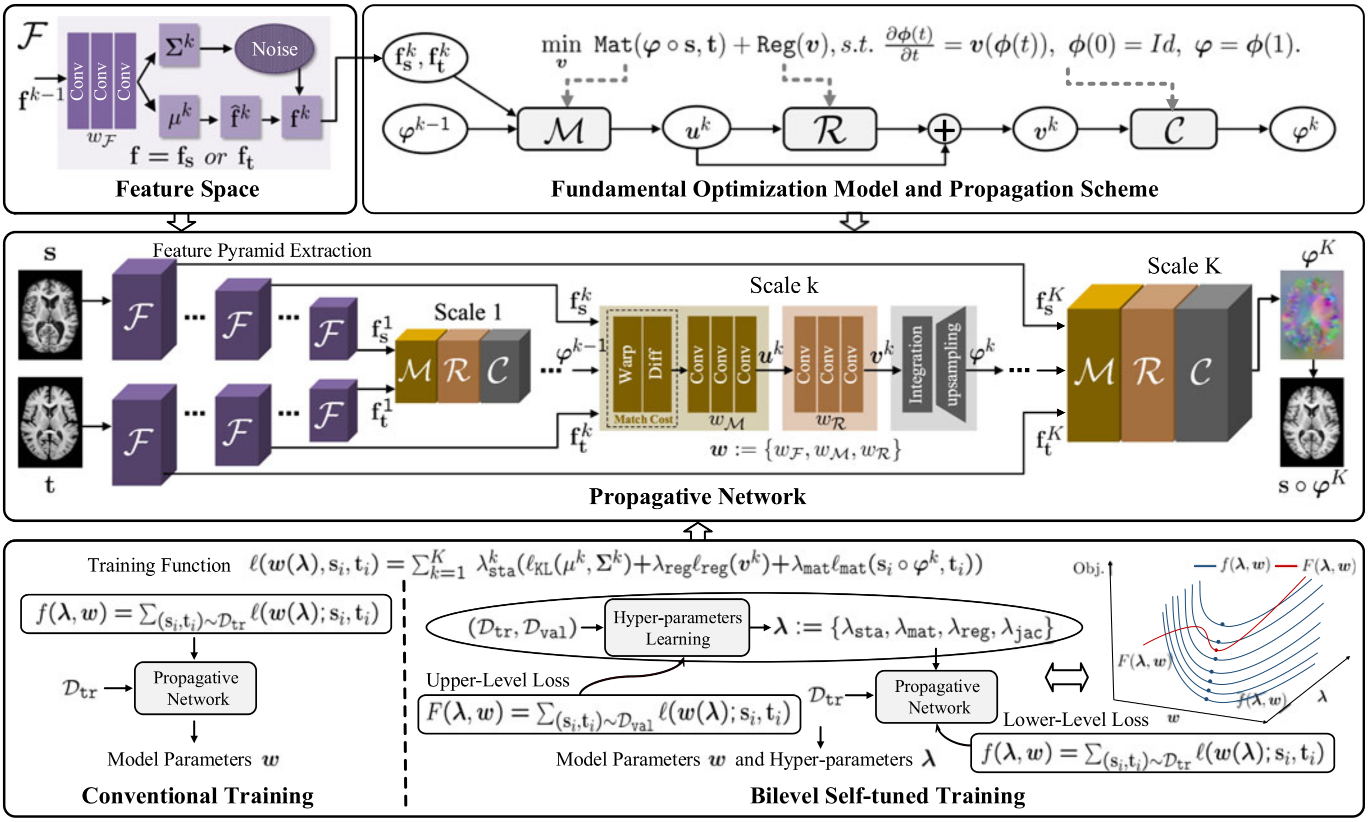

$\mathcal{F}$
$\mathbf{f}^{k-1}$
Conv
Conv
Conv
$w_{\mathcal{F}}$
Σ^k
Noise
μ^k
$\hat{\mathbf{f}}^k$
$\mathbf{f}^k$
$\mathbf{f} = \mathbf{f}_s$ or $\mathbf{f}_t$
Feature Space
$\min_{\boldsymbol{v}} \mathtt{Mat}(\boldsymbol{\varphi} \circ \mathbf{s}, \mathbf{t}) + \mathtt{Reg}(\boldsymbol{v}), s.t. \frac{\partial \phi(t)}{\partial t} = \boldsymbol{v}(\phi(t)),\ \phi(0) = Id,\ \boldsymbol{\varphi} = \phi(1).$
$\mathbf{f}_s^k, \mathbf{f}_t^k$
$\boldsymbol{\varphi}^{k-1}$
$\mathcal{M}$
$\boldsymbol{u}^k$
$\mathcal{R}$
$\boldsymbol{v}^k$
$\mathcal{C}$
$\boldsymbol{\varphi}^k$
Fundamental Optimization Model and Propagation Scheme
s
Feature Pyramid Extraction
t
Scale 1
$\mathbf{f}_s^1$
$\mathbf{f}_t^1$
Scale k
$\mathbf{f}_s^k$
$\mathbf{f}_t^k$
Warp
Diff
Match Cost
Conv
$w_{\mathcal{M}}$
$w_{\mathcal{R}}$
Integration
upsampling
$\boldsymbol{w} := \{w_{\mathcal{F}}, w_{\mathcal{M}}, w_{\mathcal{R}}\}$
Scale K
$\mathbf{f}_s^K$
$\mathbf{f}_t^K$
$\boldsymbol{\varphi}^K$
$\mathbf{s} \circ \boldsymbol{\varphi}^K$
Propagative Network
Training Function $\ell(\boldsymbol{w}(\boldsymbol{\lambda}), \mathbf{s}_i, \mathbf{t}_i) = \sum_{k=1}^K \lambda_{\mathtt{sta}}^k(\ell_{\mathrm{KL}}(\mu^k, \Sigma^k) + \lambda_{\mathtt{reg}}\ell_{\mathtt{reg}}(\boldsymbol{v}^k) + \lambda_{\mathtt{mat}}\ell_{\mathtt{mat}}(\mathbf{s}_i \circ \boldsymbol{\varphi}^k, \mathbf{t}_i))$
$f(\boldsymbol{\lambda}, \boldsymbol{w}) = \sum_{(\mathbf{s}_i,\mathbf{t}_i)\sim\mathcal{D}_{\mathtt{tr}}} \ell(\boldsymbol{w}(\boldsymbol{\lambda}); \mathbf{s}_i, \mathbf{t}_i)$
$\mathcal{D}_{\mathtt{tr}}$
Propagative Network
Model Parameters $\boldsymbol{w}$
Conventional Training
$(\mathcal{D}_{\mathtt{tr}}, \mathcal{D}_{\mathtt{val}})$
Hyper-parameters Learning
$\boldsymbol{\lambda} := \{\lambda_{\mathtt{sta}}, \lambda_{\mathtt{mat}}, \lambda_{\mathtt{reg}}, \lambda_{\mathtt{jac}}\}$
Upper-Level Loss
$F(\boldsymbol{\lambda}, \boldsymbol{w}) = \sum_{(\mathbf{s}_i,\mathbf{t}_i)\sim\mathcal{D}_{\mathtt{val}}} \ell(\boldsymbol{w}(\boldsymbol{\lambda}); \mathbf{s}_i, \mathbf{t}_i)$
$\mathcal{D}_{\mathtt{tr}}$
Propagative Network
Lower-Level Loss
$f(\boldsymbol{\lambda}, \boldsymbol{w}) = \sum_{(\mathbf{s}_i,\mathbf{t}_i)\sim\mathcal{D}_{\mathtt{tr}}} \ell(\boldsymbol{w}(\boldsymbol{\lambda}); \mathbf{s}_i, \mathbf{t}_i)$
Model Parameters $\boldsymbol{w}$ and Hyper-parameters $\boldsymbol{\lambda}$
Obj.
$f(\boldsymbol{\lambda}, \boldsymbol{w})$
$F(\boldsymbol{\lambda}, \boldsymbol{w})$
$\boldsymbol{\lambda}$
$\boldsymbol{w}$
Bilevel Self-tuned Training

Fig. 1. The first row shows our optimization learning perspective, which is to cascade three modules to propagate the optimization of registration fields on feature space. In the second row, the feature extraction $\mathcal{F}$, error-based matching $\mathcal{M}$, regularization $\mathcal{R}$, constraint $\mathcal{C}$ modules form one iteration of our framework. Different from conventional training, thanks to upper-level and lower-level objectives, our bilevel training in the third row could learn model parameters w and hyper-parameters λ collaboratively. We also give its illustration. Blue cures represent the lower-level objective and their minimal values are shown as blue dots. The red cure denotes the upper-level objective, whose minimal value is the red dot.

3.1 Feature Pyramid Extraction

To learn features that are invariant to noise and uninformative intensity-variations, we propose a generative feature module, which involves a latent variable model (as in VAE [40]). We assume a probabilistic distribution for feature representations $p(\mathbf{f}|\mathbf{I})$, where $\mathbf{I}:\{\mathbf{s},\mathbf{t}\}$, $\mathbf{f}:\{\mathbf{f}_s,\mathbf{f}_t\}$, the later means feature representations of the source and target image. The prior probability of the feature $\mathbf{f}$ can be modeled as the multivariate normal distribution with spherical covariance I:

$$p(\mathbf{f}) = \mathcal{N}(0, I). \tag{2}$$

We formulate the process of feature preparation as:

$$\mathbf{f}_s^{k+1}, \mathbf{f}_t^{k+1} = \begin{cases} \mathcal{F}(\mathbf{s}, \mathbf{t};\ w_{\mathcal{F}}^{k+1}), & \text{if } k = 0, \\ \mathcal{F}(\mathbf{f}_s^k, \mathbf{f}_t^k;\ w_{\mathcal{F}}^{k+1}) & \text{otherwise.} \end{cases} \tag{3}$$

where $w_{\mathcal{F}}^{k+1}$ denotes trainable parameters of the feature extraction network at the $(k+1)$-th scale. The calculation of the mapping $\mathcal{F}$ will be discussed in Section 4.1.

3.2 Deep Propagative Modules

Based on learned feature pyramids, we further unroll the optimization process and design a series of deep modules to propagate the optimization of registration fields. The second line of Fig. 1 demonstrates the cascade of all ingredients at each iteration of our propagative network.

Error-Based Data Matching Module. At each iteration, to establish accurate voxel-to-voxel correspondence and reduce the feature space distance between image pairs, we propose to generate the error-based reconstructed registration field $\boldsymbol{u}^{k+1}$as:

$$\boldsymbol{u}^{k+1} = \mathcal{M}(\boldsymbol{\varphi}^k, \mathbf{f}_s^{k+1}, \mathbf{f}_t^{k+1}, \mathbf{e}^{k+1});\ w_{\mathcal{M}}^{k+1}), \tag{4}$$

where $w_{\mathcal{M}}^{k+1}$ denotes the parameters of the matching network, $\mathbf{e}$ stores the matching error of the corresponding voxels of the two feature presentations. Inspired by image warping, we directly perform feature warping using the spatial transform function [7] [30], then construct this error/misalignment as:

$$\mathbf{e}^{k+1} = \|\mathbf{f}_t^{k+1} - \mathbf{f}_s^{k+1} \circ \boldsymbol{\varphi}^k\|_1. \tag{5}$$

Regularization Module. Using contextual information to regularize the flow/registration field has been widely used in traditional flow methods. At each iteration, we thus apply a context-based regularization network to produce refined registration field $\boldsymbol{v}^{k+1}$ as:

$$\boldsymbol{v}^{k+1} = \mathcal{R}(\boldsymbol{u}^{k+1};\ w_{\mathcal{R}}^{k+1}), \tag{6}$$

where w_{R}^{k+1} are the learnable parameters of the regularization network. To employ the contextual information, this sub-network applies dilated convolution, which effectively enlarges the receptive field size.

Constraint Module. At each iteration, to provide the diffeomorphism guarantee, we define the deformation field through the ordinary differential equation constraint. Thus, we append a numerical integration module to generate the deformation field $\boldsymbol{\varphi}^{k+1}$ as:

$$\boldsymbol{\varphi}^{k+1} = \mathcal{C}(\boldsymbol{v}^{k+1};\ w_{\mathcal{C}}), \tag{7}$$

where $w_{\mathcal{C}}$ are the parameters in this module. We compute this integration using scaling and squaring [3] [7] method. Specifically, it recursively computes the solution in successive small time-steps h as: $\boldsymbol{\phi}(t+h) = \boldsymbol{\phi}(t) + h\boldsymbol{v}(\boldsymbol{\phi}(t)) = (x + h\boldsymbol{v}) \circ \boldsymbol{\phi}(t)$. In our experiments, we use seven steps.

4 Bilevel Self-Tuned Training

Conventional training involves choosing and tuning hyper-parameters that significantly affect model performance, especially when switching to other data or applications. General training generally uses a grid-search algorithm or manually tuning to obtain task-specific hyper-parameters, requiring many training runs with various hyper-parameter configurations, potentially leading to sub-optimal results and requiring considerably computational and human effort. In the following, we first introduce our training objective, which regards model parameters and hyper-parameters. Then, to tackle the inefficiency of hyper-parameter tuning, we propose our bilevel self-tuned training and give the solution strategy.

4.1 Training Objective

The objectives regarding model parameters $\boldsymbol{w}$ and hyper-parameters $\boldsymbol{\lambda}$ consist of three components: KL loss, matching loss, and regularization loss. Let $\boldsymbol{w}$ be the set of learnable parameters in the proposed framework, which includes the feature extraction network $w_{\mathcal{F}}$, error-based data matching network $w_{\mathcal{M}}$ and regularization network $w_{\mathcal{R}}$ at different scales (the integration module has no learnable parameters). And, the $\boldsymbol{\lambda}$ repersent the hyper-parameters to trade off the different loss terms, including $\lambda_{\mathtt{sta}}, \lambda_{\mathtt{mat}}, \lambda_{\mathtt{reg}}, \lambda_{\mathtt{jac}}$. Specifically, we compute the sum of the losses at different scales as:

$$\begin{aligned}&\ell(\boldsymbol{w}(\boldsymbol{\lambda}), \mathbf{s}_i, \mathbf{t}_i)\\ &= \sum_{k=0}^{K-1} \lambda_{\mathtt{sta}}^{k} (\ell_{\mathtt{KL}}(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k) + \lambda_{\mathtt{mat}} \ell_{\mathtt{mat}}(\mathbf{s}_i \circ \boldsymbol{\varphi}^k, \mathbf{t}_i) + \lambda_{\mathtt{reg}} \ell_{\mathtt{reg}}(\boldsymbol{v}^k)),\end{aligned} \tag{8}$$

where $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\varphi}, \boldsymbol{v}$ depends on $w_{\mathcal{F}}, w_{\mathcal{M}}, w_{\mathcal{R}}$ (cf Section 3.1 and Section 3.2). Then, we elaborate on detailed computations.

KL Loss. Computing the posterior probability $p(\mathbf{f}|\mathbf{I})$ of feature extraction module is intractable so that we utilize an approximation of the intractable true posterior probability, $q_\theta(\mathbf{f}|\mathbf{I})$ parametrized by θ [40]. We minimize the KL divergence:

$$\begin{aligned}&\min_{\theta}\ \mathrm{KL}[q_\theta(\mathbf{f}|\mathbf{I}) \| p(\mathbf{f}|\mathbf{I})],\\ &= \min_{\theta}\ \mathrm{KL}[q_\theta(\mathbf{f}|\mathbf{I}) || p(\mathbf{f})] - \mathbb{E}_{\mathbf{f}\sim q}[\ln p(\mathbf{I}|\mathbf{f})],\end{aligned} \tag{9}$$

where the first term acts as a regularizer, while the second term is an expected negative reconstruction error. In this work, the reconstruction error corresponds to the registration loss, including matching loss and regularization loss. Then, we model the approximate posterior $q_\theta(\mathbf{f}|\mathbf{I})$ as a multivariate normal:

$$q_\theta(\mathbf{f}|\mathbf{I}) = \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}(\mathbf{f}|\mathbf{I}), \boldsymbol{\Sigma}(\mathbf{f}|\mathbf{I})), \tag{10}$$

and apply feature extraction network to predict approximate posterior probability parameters, mean $\boldsymbol{\mu}(\mathbf{f}|\mathbf{I})$ and covariance $\boldsymbol{\Sigma}(\mathbf{f}|\mathbf{I})$, from which we then sample the feature to generate the deformation fields. The KL-term can be computed in closed form:

$$\ell_{\mathtt{KL}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 1/2(\mathrm{tr}(\boldsymbol{\Sigma}) + ||\boldsymbol{\mu}|| - \log\ \det(\boldsymbol{\Sigma}) - m), \tag{11}$$

where $\boldsymbol{\Sigma}_{\mathbf{f}|\mathbf{I}}$ need to be diagonal, and m is a const.

Matching Loss. We inherit similarity metric from energy-based approaches [4], [5], [6] to define the data matching loss. Specifically, we use local normalized cross correlation coefficient to penalize differences in appearance. Note that, when computing matching loss, we scale the image pairs and warp the downsampled images with the deformation field at each scale. At different scales, we use different window sizes to compute the local normalized correlation coefficient. From the zeroth to the third scale, the window sizes are set to $3, 5, 7, 9$.

Regularization Loss. We employ the diffusion regularizer on spatial gradients of the velocity fields and apply the Jacobian determinant loss to further constrain the smoothness of the deformation fields as:

$$\ell_{\mathtt{reg}}(\boldsymbol{v}) = \sum_{x\in\Omega} \parallel \bigtriangledown \boldsymbol{v}(x) \|_2^2 + \lambda_{\mathtt{jac}} \max(0, -|J_{\boldsymbol{v}}(x)|), \tag{12}$$

where the Jacobian matrix $J_\phi(x) = \nabla\phi(x)$ captures the local properties of ϕ around voxel x. We penalize these negative determinants to enforce topology-preservation [3].

4.2 Self-Tuned Bilevel Formulation

To tackle the inefficiency of hyper-parameter tuning, we propose our new bilevel self-tuned training, allowing the efficient search of the task-specific hyper-parameter, as shown in the right-bottom part of Fig. 1. We start by denoting $\mathcal{D}_{\mathtt{tr}}$ and $\mathcal{D}_{\mathtt{val}}$ as the training and validation sets, respectively. In our setting, we formulate the learning of the model parameter as f:

$$f(\boldsymbol{\lambda}, \boldsymbol{w}) = \sum_{(\mathbf{s}_i, \mathbf{t}_i)\sim\mathcal{D}_{\mathtt{tr}}} \ell(\boldsymbol{w}(\boldsymbol{\lambda}); \mathbf{s}_i, \mathbf{t}_i), \tag{13}$$

where the $\ell(\boldsymbol{w}(\boldsymbol{\lambda}); \mathbf{s}_i, \mathbf{t}_i)$ denotes the loss function with the model parameter $\boldsymbol{w}$ coupled with hyper-parameters $\boldsymbol{\lambda}$, the $(\mathbf{s}_i, \mathbf{t}_i)$ corresponding to source and target image.

To identify optimal task-specific hyper-parameters $\boldsymbol{\lambda}$, we minimize the empirical loss on a validation set, which represents a proxy for the generalization error of $\boldsymbol{w}$. We define the objective of the hyper-parameter learning problem as F:

$$F(\boldsymbol{\lambda}, \boldsymbol{w}) = \sum_{(\mathbf{s}_i, \mathbf{t}_i)\sim\mathcal{D}_{\mathtt{val}}} \ell(\boldsymbol{w}(\boldsymbol{\lambda}); \mathbf{s}_i, \mathbf{t}_i). \tag{14}$$

The F does not depend explicitly on the hyper-parameter $\boldsymbol{\lambda}$, since in our setting $\boldsymbol{\lambda}$ is instrumental in finding a good model $\boldsymbol{w}$, which is our final goal.

We then consider bilevel optimization problems [41], [42], [43], [44], [45], [46], [47] of the form to formulate our self-tuning learning. Specifically, we introduce a hierarchical optimization model governed by the constraint $\mathcal{C}(\boldsymbol{\lambda})$ as:

$$\min_{\boldsymbol{\lambda}} F(\boldsymbol{\lambda}, \boldsymbol{w}),\ s.t.\ \boldsymbol{w} \in \mathcal{C}(\boldsymbol{\lambda}), \tag{15}$$

where we take optimization of task-specific hyper-parameters $\boldsymbol{\lambda}$ as the upper-level subproblem, constraint $\mathcal{C}(\boldsymbol{\lambda})$ as the solution set of the lower-level subproblem, while lower-level subproblem denotes optimization of model parameters $\boldsymbol{w}$:

$$\mathcal{C}(\boldsymbol{\lambda}) := \{\arg\min_{\boldsymbol{w}} f(\boldsymbol{\lambda}, \boldsymbol{w})\}, \tag{16}$$

where we try to find optimal parameters $\boldsymbol{w}$ by minimizing the objective f on training data $\mathcal{D}_{\texttt{tr}}$. We follow methods [41], [42], [45], [47] to rely on the uniqueness of $C(\boldsymbol{\lambda})$, where $\mathcal{C}(\boldsymbol{\lambda}) = \arg\min f(\boldsymbol{\lambda}, \boldsymbol{w})$ and the lower-level subproblem only has one single optimal solution $\boldsymbol{w}^*$ for a given $\boldsymbol{\lambda}$.

4.3 First-Order Solution Strategy

Due to the nested structure of the bilevel problem in Eq. (15), evaluating exact gradients of $\boldsymbol{\lambda}$ for the upper-level problem is difficult and computationally challenging. Moreover, this hierarchical structure between two levels greatly challenge traditional optimization techniques especially when the dimension of either level is huge. It is the case of our training where the lower level involves hundreds of thousands of deep network parameters. Moreover, computing gradients w.r.t. 3D medical volumes demands additional efforts. Existing solutions [41], [42], [43], [45] are inapplicable to our self-tuned training. Next, we introduce an efficient solution to compute the gradient of $\boldsymbol{\lambda}$ via one-step first-order gradient approximation.

Derivation of Computing the Gradient of $\boldsymbol{\lambda}$. Our goal is to calculate the derivatives of $F(\boldsymbol{\lambda}, \boldsymbol{w}^*)$ with respect to $\boldsymbol{\lambda}$ as[1]:

$$\frac{\partial F(\boldsymbol{\lambda}, \boldsymbol{w}^*)}{\partial \boldsymbol{\lambda}} = \frac{\partial F(\boldsymbol{\lambda}, \boldsymbol{w}^*)}{\partial \boldsymbol{\lambda}} + \left(\frac{\partial \boldsymbol{w}^*}{\partial \boldsymbol{\lambda}'}\right)' \frac{\partial F(\boldsymbol{\lambda}, \boldsymbol{w}^*)}{\partial \boldsymbol{w}^*}, \tag{17}$$

where we denote the transpose operation as "′".

To solve the aforementioned problem, first, we approximate the solutions of lower-level optimization $\boldsymbol{w}^*$ in Eq. (16) by the T-step iterate of a dynamical system. Given an initialization $\boldsymbol{w}_0 = \mathcal{E}_0(\boldsymbol{\lambda})$ at $t = 0$, the iteration process can be written as:

$$\boldsymbol{w}_t = \mathcal{E}_t(\boldsymbol{w}_{t-1}; \boldsymbol{\lambda}),\ t = 1, \ldots, T \tag{18}$$

where $\mathcal{E}_t$ denotes the operation performed at the tth step and T is the number of iterations. For example, we formulate $\mathcal{E}_t$ based on the gradient descent rule as:

$$\mathcal{E}_t(\boldsymbol{w}_{t-1}; \boldsymbol{\lambda}) = \boldsymbol{w}_{t-1} - s_t \frac{\partial f(\boldsymbol{\lambda}, \boldsymbol{w}_{t-1})}{\partial \boldsymbol{w}_{t-1}}, \tag{19}$$

where s_t denotes the learning rate. To reduce the computational burden further, we set $T = 1$. Similar to the work of [48], we perform one-step iteration. So that we propose a simple approximation scheme as:

$$\frac{\partial F(\boldsymbol{\lambda}, \boldsymbol{w}^*)}{\partial \boldsymbol{\lambda}} = \frac{\partial F(\boldsymbol{\lambda}, \boldsymbol{w}_0 - s_1 \frac{\partial f(\boldsymbol{\lambda}, \boldsymbol{w}_0)}{\partial \boldsymbol{w}_0})}{\partial \boldsymbol{\lambda}}. \tag{20}$$

Now, by formulating the dynamical system as that in Eq. (19), we then write $\frac{\partial \boldsymbol{w}^*}{\partial \boldsymbol{\lambda}}$ as:

$$\frac{\partial \boldsymbol{w}^*}{\partial \boldsymbol{\lambda}} = \frac{\partial \left(\boldsymbol{w}_0 - s_1 \frac{\partial f(\boldsymbol{\lambda}, \boldsymbol{w}_0)}{\partial \boldsymbol{w}_0}\right)}{\partial \boldsymbol{\lambda}} = -s_1 \frac{\partial^2 f(\boldsymbol{\lambda}, \boldsymbol{w}_0)}{\partial \boldsymbol{\lambda} \partial \boldsymbol{w}}. \tag{21}$$

The expression above contains expensive matrix-vector product for Hessian calculation. We then approximate the Hessian calculation with the first order gradients. We introduce the following central difference to approximate it as:

$$\frac{\partial F(\boldsymbol{\lambda}, \boldsymbol{w}^*)}{\partial \boldsymbol{w}^*} \frac{\partial^2 f(\boldsymbol{\lambda}, \boldsymbol{w}_0)}{\partial \boldsymbol{\lambda} \partial \boldsymbol{w}_0} \approx \frac{\frac{\partial f(\boldsymbol{\lambda}, \boldsymbol{w}_0^+)}{\partial \boldsymbol{\lambda}} - \frac{\partial f(\boldsymbol{\lambda}, \boldsymbol{w}_0^-)}{\partial \boldsymbol{\lambda}}}{2\epsilon}, \tag{22}$$

where $\boldsymbol{w}_0^{\pm} = \boldsymbol{w}_0 \pm \epsilon \frac{\partial F(\boldsymbol{\lambda}, \boldsymbol{w}^*)}{\partial \boldsymbol{w}^*}$, the ϵ is set to be a small scalar equal to the learning rate. So that we may further reduce the computation complexity.

Algorithm 1. Bilevel Self-Tuned Training Algorithm

```
Input: The training and validation datasets $\mathcal{D}_{\texttt{tr}}$ and $\mathcal{D}_{\texttt{val}}$ and initialization parameters.
Output: The optimal hyper-parameters and model parameters.
while not converged do
    Calculate the practical Jacobian $\frac{\partial F(\boldsymbol{\lambda}, \boldsymbol{w}^*)}{\partial \boldsymbol{\lambda}}$ in Eq. (17) via one-step first-order gradient approximation.
    Perform gradient descent to update $\boldsymbol{\lambda}$ based on $\frac{\partial F(\boldsymbol{\lambda}, \boldsymbol{w}^*)}{\partial \boldsymbol{\lambda}}$.
    Calculate $w$ based on Eq. (16).
end while
return The optimal $(\boldsymbol{\lambda}, \boldsymbol{w}(\boldsymbol{\lambda}))$.
```

The overall iterative procedure is outlined in Algorithm 1. Overall, we introduce an efficient and feasible solution strategy to solve bilevel optimization of hyper-parameters for registration networks.

5 Experiments

In this section, we first introduce our experimental setup. Then we explore the impact of each component of our paradigm and the benefits of the proposed bilevel self-tuned training strategy. Next, to demonstrate the superiority of the proposed algorithm, we compare it with state-of-the-art deformable registration techniques on accuracy, robustness, diffeomorphism preservation as well as efficiency. We evaluate the registration algorithms mainly on 3D brain MRI scans. Evaluations are conducted in the way of one aligning all the source data to a common atlas, called image-to-atlas registration. Then we extend our paradigm to address

[1] Please notice that we actually do not distinguish between the operation of the derivatives and partial derivatives to simplify our presentation.

general registration between two arbitrary volumes on liver CT scans, called image-to-image registration.

We also explore the utility of our registration framework to support the down-streaming medical image analysis tasks. Multi-modal image registration and fusion are two important research issues in medical image processing. To optimally fuse two medical images, one needs to first accurately align them by minimizing non-linear differences between them using registration techniques. In the following experiments, we demonstrated how to apply our paradigm to solve the challenging multi-modal registration tasks and powerfully support the following medical multi-modal fusion. Medical image segmentation is also crucial and highly relevant in medical image analysis. Manual segmentation of brain MR images requires expertise and is time-consuming, and thus efficient data augmentation methods should be explored. The proposed framework can be used to register labeled atlas images to produce more labels images for segmentation. Moreover, the latent feature space in Section 3.1 helps to encode similar deformations close to each other and allows the generation of synthetic deformations for a single image, which is beneficial for the data augmentation. We also demonstrated how to apply our paradigm to strongly support the medical image segmentation.

5.1 Experimental Setup

Dataset and Pre-Processing. We performed image-to-atlas registration on brain MR datasets, including 551 T1 weighted MR volumes from seven publicly available datasets: ADNI [49], ABIDE [50], PPMI [51], OASIS [52], and HCP [53]. These scans were splitted into 370, 40, and 141 for training, validation, and testing, respectively. We used the publicly available atlas from [7] as the target. Considering the large disparity among different datasets, all scans were preprocessed with motion correction, NU intensity correction, normalization, skull stripping, and affine registration. We used FreeSurfer [54] software to perform skull stripping and used FSL [55] software for affine registration. The images were cropped to $160 \times 192 \times 224$ with 1 mm isotropic resolution after cropping unnecessary areas. For evaluation, all test MRI scans were anatomically segmented with Freesurfer to extract 30 anatomical structures. As for image-to-image registration on liver CT scans, we included four publicly available datasets: MSD [56], BFH [57], SLIVER [58], and LSPIG [31]. We used the MSD and BFH with 1025 scans in total, and splitter into 900 and 125 as training and validation data. We randomly selected 380 image pairs with segmentation ground truth in SLIVER [58] for the evaluation. We also evaluated with 34 intrasubject image pairs with segmentation ground truth in LSPIG [31]. For liver CT scans, we carried out normalization preprocessing steps and resample to a size of $128 \times 128 \times 128$. We embedded the affine network as an integrated part. These experiments enable not only assessment of performance on multi-site datasets but also the evaluation of scans that were not observed by the deep networks during training.

In addition, we performed multi-modal registration on BraTS18 and ISeg19 datasets, which are obtained from the Brain Tumour Segmentation challenge 2018 and Infant Brain MRI Segmentation challenge 2019. Overall, the available training set consists of 135 cases, and for each case, two image modalities were standardized into a 3D volume in size of $160 \times 160 \times 160$ with 1 mm isotropic resolution. Among them, 10 cases have segmentation ground truth. The set was splitted into 115, 10 and 10 for train, validation and test. We use segmentation accuracy as a proxy for evaluating image registration accuracy. As most of the provided T1 and T2 weighted images were already aligned, we randomly chose one of the T2 scans as our atlas and modeled our task as trying to register T1 scans to this T2 atlas.

Implementation. The affine network progressively downsamples the input with 9 convolutions and employs a fully-connected layer to produce 12 numeric parameters, composed of a 3×3 transform matrix A and a 3D vector b. As for the propagation network, we use filters of size $3 \times 3 \times 3$ for all the convolutional layers. All convolutions are followed by a leaky ReLU function except the one that outputs the registration field. To generate feature representations, we use layers of convolutional filters to downsample the features at the previous pyramid level by a factor of 2. Computation on the full resolution may easily exhaust the memory, thus we choose to output a half-resolution smooth enough deformation field and up-sample it via interpolation to obtain the full-resolution deformation field. The proposed method[2] is implemented with Pytorch [59] package.

Evaluation Metrics. To achieve comprehensive evaluation, both the average Dice score [60] over registered testing pairs and the Jacobian matrix over the computed deformation are considered as evaluation metrics, evaluating the anatomical overlap correspondences of the registered volume pairs and the smoothness of the deformation fields, respectively. The Dice score of two regions A, B is formulated as:

$$Dice(A, B) = 2 \cdot \frac{|A \cap B|}{|A| + |B|}, \tag{23}$$

where a Dice score of 1 means the most perfectly overlap. The Jacobian matrix $\mathbf{J}_\phi(\mathbf{x}) = \nabla\phi(\mathbf{x})$ captures the local properties of ϕ around voxel $\mathbf{x}$. According to [3], the deformation is diffeomorphic at the locations where $\mathbf{J}_\phi(\mathbf{x}) > \mathbf{0}$. We count all the folds, where $\mathbf{J}_\phi(\mathbf{x}) \leq \mathbf{0}$. In addition, we compute the average Normalized Correlation Coefficient (NCC) between image pairs as an auxiliary evaluation metric.

State-of-the-Art Methods. We compare the proposed method with state-of-the-art registration techniques, including three optimization-based tools: Elastix [61], Symmetric Normalization (SyN) [62], NiftyReg [6], and two learning-based methods: VoxelMorph [7] and its diffeomorphic variant [8] (referred as VM and VM-diff, respectively). The parameter settings of the conventional methods are as follows. For Elastix, we run B-spline registration with Mattes Mutual Information as a cost function and set the control point spacing to 16 voxels. Four scales are used with 500 iterations per scale. For the SyN algorithm, we use the version implemented in the ANTs [63] package and take Cross-Correlation as the similarity measure metric and use the

[2] Our codes are available at https://github.com/dut-media-lab/MultiPropReg

TABLE 1
Ablation Analysis of the Feature Pyramid Extraction Module on Five Brain MRI Datasets in Terms of Dice Score and NCC

Model		Affine only	2-scale		3-scale		4-scale	
			W/O FEN	W/ FEN	W/O FEN	W/ FEN	W/O FEN	W/ FEN
OASIS	Dice score	0.580 ± 0.028	0.724 ± 0.019	0.744 ± 0.016	0.764 ± 0.011	$\mathbf{0.777 \pm 0.006}$	0.770 ± 0.008	0.773 ± 0.007
	NCC	0.088 ± 0.004	0.217 ± 0.004	0.229 ± 0.003	0.233 ± 0.003	$\mathbf{0.245 \pm 0.002}$	0.236 ± 0.003	0.240 ± 0.003
ABIDE	Dice score	0.624 ± 0.024	0.736 ± 0.016	0.740 ± 0.018	0.754 ± 0.016	$\mathbf{0.764 \pm 0.015}$	0.763 ± 0.014	0.761 ± 0.017
	NCC	0.094 ± 0.005	0.214 ± 0.004	0.227 ± 0.004	0.229 ± 0.004	$\mathbf{0.241 \pm 0.004}$	0.231 ± 0.004	0.237 ± 0.004
ADNI	Dice score	0.571 ± 0.049	0.702 ± 0.038	0.730 ± 0.033	0.752 ± 0.024	$\mathbf{0.773 \pm 0.017}$	0.763 ± 0.020	0.769 ± 0.018
	NCC	0.086 ± 0.006	0.213 ± 0.007	0.227 ± 0.006	0.231 ± 0.005	$\mathbf{0.244 \pm 0.005}$	0.233 ± 0.005	0.239 ± 0.005
PPMI	Dice score	0.610 ± 0.033	0.740 ± 0.023	0.758 ± 0.019	0.773 ± 0.013	0.785 ± 0.011	0.779 ± 0.011	$\mathbf{0.789 \pm 0.011}$
	NCC	0.088 ± 0.004	0.213 ± 0.005	0.225 ± 0.004	0.229 ± 0.004	$\mathbf{0.240 \pm 0.004}$	0.230 ± 0.004	0.235 ± 0.004
HCP	Dice score	0.666 ± 0.027	0.698 ± 0.027	0.759 ± 0.014	0.738 ± 0.018	0.776 ± 0.010	0.745 ± 0.021	$\mathbf{0.777 \pm 0.009}$
	NCC	0.098 ± 0.004	0.183 ± 0.011	0.220 ± 0.005	0.204 ± 0.010	$\mathbf{0.240 \pm 0.004}$	0.204 ± 0.011	0.234 ± 0.004

Larger values indicate better performance. The Affine only model represents the results for affine alignment.

SyN step size of 0.25, Gaussian parameters (9, 0.2), at three scales with 201 iterations each. As for NiftyReg, we use the Normalized Mutual Information cost function. We run it with 12 threads using 1500 iterations. We run Elastix, SyN, and NiftyReg on a PC with i7-8700 (@3.20GHz, 32G RAM), while learning-based methods on NVIDIA TITAN XP.

5.2 Ablation Analysis

We investigate the role of different propagation components in our model, including feature extraction, context information, multi-scale degree as well as the architecture of the matching network and regularization network. We also discuss the benefits of the bilevel self-tuned training.

Feature Extraction Network Evaluation. We substitute our Feature Extraction Network (FEN) with handcrafted image pyramids as the case of without FEN and compare the performance gap between these two cases. Table 1 lists the registration accuracy in terms of both the Dice score and NCC at the 2, 3, and 4-scale. As shown, the importance of this module is clear given the inferior quality of the models without FEN under different scales. Moreover, as the table shows, the FEN may evidently increase the accuracy in terms of both evaluation metrics and helps to achieve a more robust alignment quality (lower standard deviation). In principle, the number of the scale of the model shows the trade-off between accuracy and model size. On the whole, the 3-scale model outputs superior performance over the other scale.

Model Configurations Evaluation. First, we illustrate the effect of network architecture by changing the number of convolutional layers of either Matching Network (MN) or Regularization Network (RN). Our deep framework uses a three-layer matching network and three-layer regularization network at each level. Table 2 shows the results by four variants that use one layer ($-$) and six layers ($+$) and keeping the rest the same. As shown, the larger-capacity matching network leads to better results, resembling the critical role of data matching term for accurate spatial correspondence. Further, we consider the case where the regularization network is eliminated and only the matching network is engaged for the registration process to figure out the significance of regularization on the deformation field. From Table 3 we can see that the participation of regularization can ideally exploit the context information to refine the predicted field and obviously promote the registration performance. Besides, we observe in experiments that a deeper network architecture could more easily get stuck at overfitting, which can be solved by employing more training data. Therefore, we adopt the 3-scale model to perform the following experiments.

Constraint Module Evaluation. Except for the propagation networks, the advantage of constraint on the registration field is also explored. Fig. 2 illustrates the warped images and the corresponding flow grids generated by the model with and without integration operation. Apparently, the constraint can properly reduce unreasonable overlaps of the deformation field and preserve the topology of the warped volume, promising a smoother and more reliable registration. To provide an intuitive comprehension of the effect of each component, Fig. 3 visualizes the deformation fields generated in each module during the propagation process, and the enlargement of corresponding local detail is attached on the bottom right. As the visualization shows, the data matching network first provides a primary estimation of field $\boldsymbol{u}^1$, whereafter the regularization network refines the field to make $\boldsymbol{v}^1$ smoother. Further, the integration operation ideally reduces the unreasonable overlaps

TABLE 2
Ablation Analysis of Model Configurations on Five Brain MRI Datasets in Terms of Dice Score and NCC

Methods		RN-	MN-	RN+	MN+
OASIS	Dice score	0.771	0.751	0.774	**0.781**
	NCC	0.235	0.202	0.241	**0.250**
ABIDE	Dice score	0.759	0.731	0.765	**0.771**
	NCC	0.230	0.197	0.236	**0.246**
ADNI	Dice score	0.762	0.737	0.771	**0.775**
	NCC	0.236	0.197	0.239	**0.249**
PPMI	Dice score	0.780	0.757	0.784	**0.788**
	NCC	0.229	0.196	0.236	**0.245**
HCP	Dice score	0.763	0.741	0.771	**0.777**
	NCC	0.234	0.199	0.239	**0.249**

TABLE 3
Ablation Analysis of Model Components on Five Brain MRI Datasets in Terms of Dice Score and NCC

Model		Affine only	2-scale		3-scale		4-scale	
			W/O RN	W/ RN	W/O RN	W/ RN	W/O RN	W/ RN
OASIS	Dice score	0.580 ± 0.028	0.746 ± 0.015	0.754 ± 0.011	0.771 ± 0.009	**0.777** ± **0.006**	0.773 ± 0.007	0.774 ± 0.007
	NCC	0.088 ± 0.004	0.227 ± 0.003	0.228 ± 0.003	0.239 ± 0.003	**0.245** ± **0.002**	0.236 ± 0.003	0.245 ± 0.003
ABIDE	Dice score	0.624 ± 0.024	0.746 ± 0.016	0.745 ± 0.016	0.759 ± 0.014	0.764 ± 0.015	0.762 ± 0.014	**0.768** ± **0.015**
	NCC	0.094 ± 0.005	0.223 ± 0.003	0.225 ± 0.004	0.235 ± 0.004	0.241 ± 0.004	0.232 ± 0.004	**0.242** ± **0.004**
ADNI	Dice score	0.571 ± 0.049	0.731 ± 0.033	0.739 ± 0.029	0.765 ± 0.020	**0.773** ± **0.017**	0.767 ± 0.019	0.768 ± 0.020
	NCC	0.086 ± 0.006	0.224 ± 0.006	0.225 ± 0.006	0.237 ± 0.005	**0.244** ± **0.005**	0.234 ± 0.005	0.244 ± 0.005
PPMI	Dice score	0.610 ± 0.033	0.758 ± 0.018	0.762 ± 0.016	0.779 ± 0.013	**0.785** ± **0.011**	0.780 ± 0.012	0.782 ± 0.012
	NCC	0.088 ± 0.004	0.222 ± 0.004	0.224 ± 0.004	0.234 ± 0.004	0.240 ± 0.004	0.231 ± 0.004	**0.241** ± **0.004**
HCP	Dice score	0.666 ± 0.027	0.738 ± 0.028	0.745 ± 0.024	0.767 ± 0.017	**0.776** ±**0.010**	0.769 ± 0.015	0.770 ± 0.016
	NCC	0.098 ± 0.004	0.225 ± 0.005	0.226 ± 0.005	0.238 ± 0.004	0.240 ± 0.004	0.235 ± 0.005	**0.244** ± **0.004**

RN represents the regularization network.

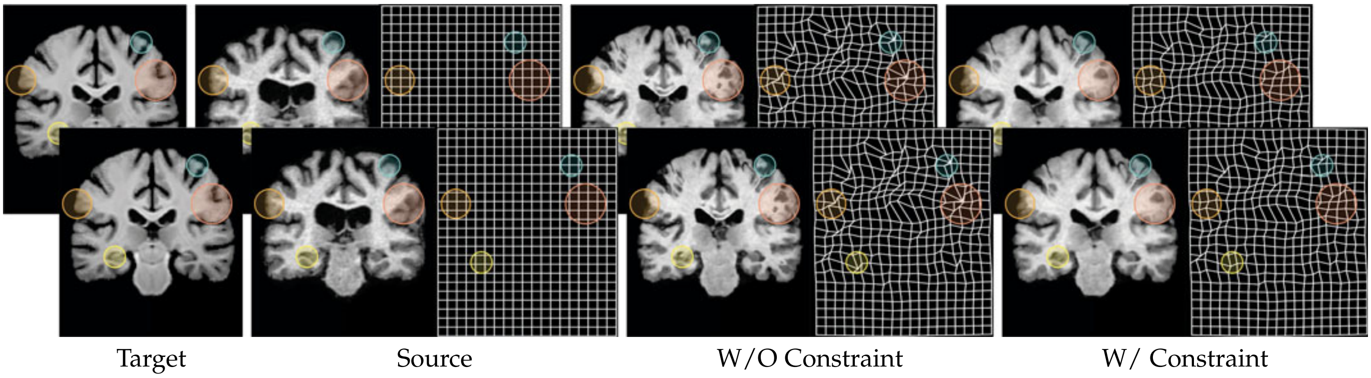


Fig. 2. Comparisons on deformation fields by warping the source image to target without and with the constraint module. Singularities emerge in the circled fields when applying no constraint.

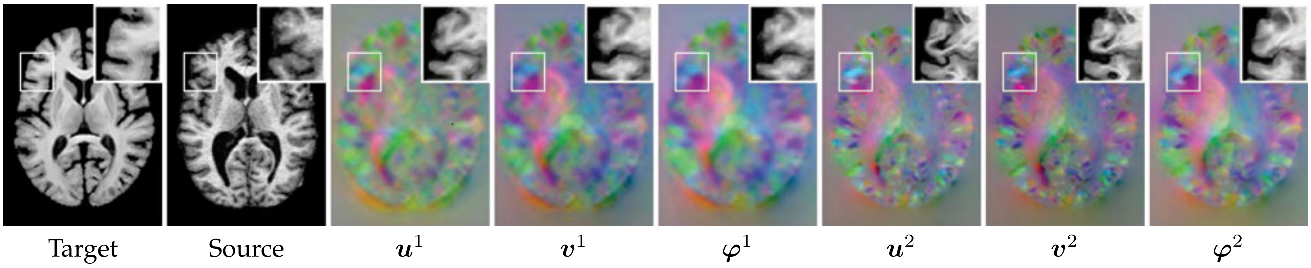


Fig. 3. The evolution of deformation color maps and registered images with the propagation of registration fields in the first two scales.

within the field and guarantees a diffeomorphic deformation $\boldsymbol{\varphi}^1$. Repeating this way, the propagated field can be guided towards the desirable results and achieves satisfying accuracy.

Training Strategies Evaluation. Given the training and validation data, thanks to our joint learning of the task-specific hyper-parameters and model parameters, our network does not have to manually select hyper-parameters during training. Instead, we design the bilevel training strategy to automatically discover the optimal value for task-specific hyper-parameters. As Table 4 shows, on the brain MR registration task the model taking bilevel self-tuned training can achieve much more accurate performance than the default training setting, where the hyper-parameters are manually tuned. When switching to liver CT data, the default hyper-parameters could not generalize well. Generally, a grid search algorithm or manually tuning will be applied to obtain these task-specific parameters, which require many training runs. In contrast, the proposed bilevel self-tuned training could auto-adapt to liver CT data and achieve satisfying performance. These results indicate that our bilevel training strategy may offer a compelling path towards automated hyper-parameter tuning for registration networks.

5.3 Image-to-Atlas Registration

First, Fig. 4 illustrates our representative registration results, two example registration cases including adult brain data and teen brain data. The large deformations that exist in the adult scans make registration challenging. As for teen MR scans, due to still in inherent myelination and maturation

TABLE 4
Ablation Analysis of the Bilevel Self-Tuned Training Strategy on Five Brain MRI Datasets and Two Liver CT Datasets in Terms of Dice Score and NCC

Methods		Default	Bilevel
OASIS	Dice score	**0.777 ± 0.006**	0.776 ± 0.007
	NCC	0.245 ± 0.002	**0.257 ± 0.002**
ABIDE	Dice score	0.764 ± 0.015	**0.770 ± 0.013**
	NCC	0.241 ± 0.004	**0.251 ± 0.003**
ADNI	Dice score	0.773 ± 0.017	**0.775 ± 0.016**
	NCC	0.244 ± 0.005	**0.255 ± 0.004**
PPMI	Dice score	0.785 ± 0.011	**0.785 ± 0.011**
	NCC	0.240 ± 0.004	**0.254 ± 0.003**
HCP	Dice score	0.776 ± 0.010	**0.776 ± 0.010**
	NCC	0.240 ± 0.004	**0.250 ± 0.003**
SLIVER	Dice score	0.883 ± 0.042	**0.910 ± 0.027**
	NCC	0.228 ± 0.153	**0.374 ± 0.035**
LSPIG	Dice score	0.822 ± 0.061	**0.855 ± 0.045**
	NCC	0.144 ± 0.124	**0.348 ± 0.056**

The default model represents the model that uses traditional training with manually selected hyper-parameters.

process, white matter and gray matter exhibit obvious differences in contrast to the fixed image, also making registration difficult. As result, all the source images are well aligned to the target, demonstrating our excellent registration performance.

Then, we quantitatively evaluate the accuracy and rationality of all the registration techniques. Table 5 depicts the accuracy and stability of the methods in terms of the Dice score on the five different datasets, where higher values and lower variance indicate a more accurate and stable registration. Our method gives an obvious lower variance and a comparable mean of Dice score on most of the datasets. As shown in Table 6, only Elastix and our method can decrease the number of folds to zero on specific datasets. However, the registration accuracy of Elastix on these datasets is far from satisfactory. Only our method achieves both high accuracy and strong stability while also having nearly zero non-negative Jacobian locations, benefiting from well-designed network architectures, regularization loss functions, and introduction of ordinary differential equation constraint.

To take a deeper perspective of the alignment of anatomical segmentation, we illustrate the Dice score of 30 anatomical structures in Fig. 5. Limited by space, besides our method, we take SyN and VM as the representatives for the optimization-based and learning-based techniques. We can see that compared with the top-performing conventional method SyN, the popular deep methods VM gives evenly accuracy but perform much less stable among different anatomical segmentations. In contrast, our deep model could achieve a good balance between accuracy and stability in virtue of a proper trade-off between the model-based domain knowledge and data-based deep representation. Fig. 6 visualizes one slice of the registered segmentations generated by different methods. The target is set as semi-transparent on the upper layer to present an intuitive discrepancy between the results and target. Compared with other approaches, ours has higher consistency with the target for both interior and outline.

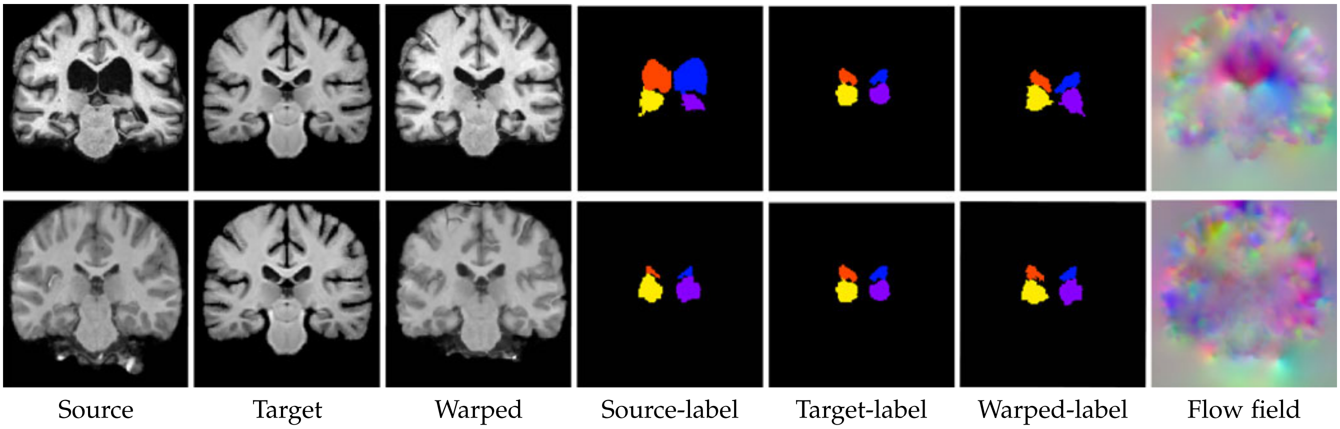


Fig. 4. Example MR coronal slices of input target, source and warped image for our method with corresponding labels of ventricles, thalami, and hippocampi. The last column shows the RGB image of the registration field. Each row refers to an example registration case of brain MR data.

TABLE 5
Qualitative Comparison Results on Brain MR Registration Tasks

Methods	Affine only	Elastix	NiftyReg	SyN	VM	VM-diff	Ours
OASIS	0.580 ± 0.028	0.709 ± 0.023	0.748 ± 0.017	0.765 ± 0.010	0.765 ± 0.010	0.757 ± 0.011	**0.777 ± 0.006**
ABIDE	0.624 ± 0.024	0.699 ± 0.025	0.747 ± 0.026	0.728 ± 0.029	0.754 ± 0.016	**0.773 ± 0.009**	0.764 ± 0.016
ADNI	0.571 ± 0.049	0.697 ± 0.039	0.737 ± 0.035	0.761 ± 0.021	0.761 ± 0.024	0.768 ± 0.020	**0.773 ± 0.017**
PPMI	0.610 ± 0.033	0.730 ± 0.021	0.765 ± 0.015	0.778 ± 0.013	0.775 ± 0.013	0.781 ± 0.011	**0.787 ± 0.010**
HCP	0.666 ± 0.027	0.729 ± 0.017	0.768 ± 0.013	0.767 ± 0.016	0.768 ± 0.013	0.413 ± 0.111	**0.776 ± 0.010**

The mean and standard deviations of the Dice score on five different datasets are listed. The average Dice score is computed over all the structures and subjects.

TABLE 6
Qualitative Comparison Results on Brain MR Registration Tasks

Methods	Elastix	NiftyReg	SyN	VM	VM-diff	Ours
OASIS	**0**	416.2 ± 416.0	29094 ± 8772	32029 ± 3498	35.7 ± 13.3	6.2 ± 4.6
ABIDE	**0**	11.4 ± 13.2	27288 ± 3411	28861 ± 1616	25.4 ± 13.1	1.0 ± 0.9
ADNI	307.5 ± 1068	572.2 ± 878.9	30737 ± 9537	33047 ± 4667	43.4 ± 33.1	**5.3 ± 6.0**
PPMI	2.0 ± 15.6	314.3 ± 353.6	25452 ± 6490	30192 ± 3375	29.7 ± 24.8	**0.1 ± 0.7**
HCP	**0**	9576 ± 2287	28379 ± 4411	30716 ± 2086	3945 ± 3854	**0**

The means and standard deviations of the number of occurred folds are listed.

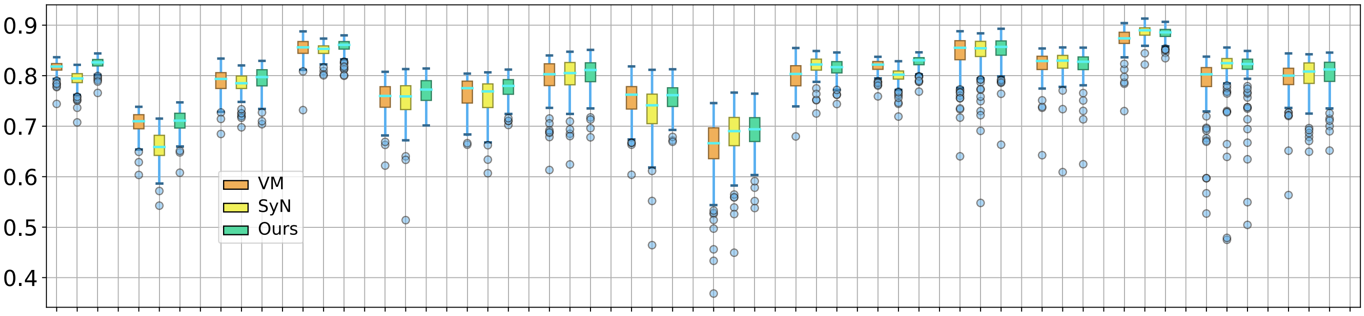


Fig. 5. Boxplot indicates the Dice scores for SyN, VM and our algorithm over sixteen anatomical structures including Cerebral White Matter (CblmWM), Cerebral Cortex (CblmC), Lateral Ventricle (LV), Inferior Lateral Ventricle (ILV), Cerebellum White Matter (CeblWM), Cerebellum Cortex (CereC), Thalamus (Tha), Caudate (Cau), Putamen (Pu), Pallidum (Pa), Hippocampus (Hi), Accumbens area (Am), Vessel, Third Ventricle (3V), Fourth Ventricle (4V), and Brain Stem (BS).

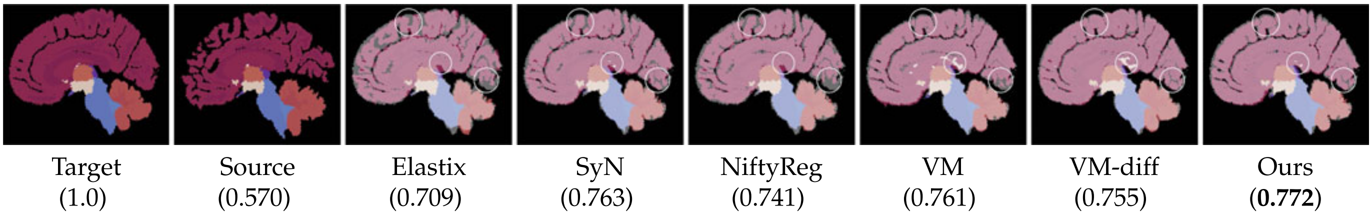


Fig. 6. Registered MR slices overlaid with atlas using different methods. The Dice scores are given in the bottom parentheses. Circles indicate several evident inconsistencies.

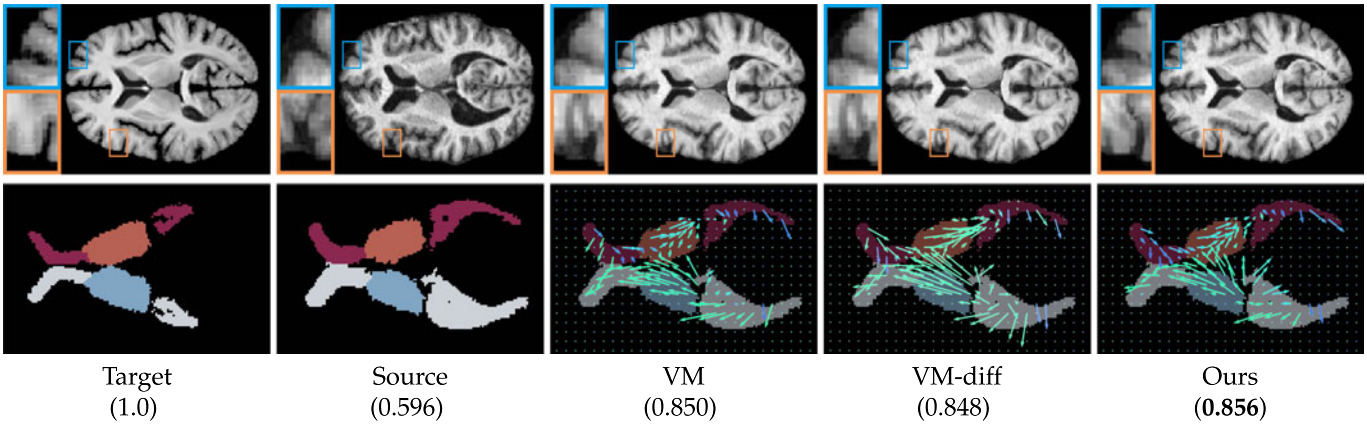


Fig. 7. Registered MR slices and segmented anatomical structures using VM, VM-diff, and our method. The Dice scores are parenthesized.

Except for the boost of accuracy, we can also ideally promote efficiency by virtue of a well-designed network and regularization on estimated flow, such as context information and integration operation. Fig. 7 visualizes both the corresponding 2D slice with a blowup of details on the right and the direction and magnitude of the generated flow with a brain volume. As the pictures on the below row show, the flow generated by our method contains fewer displacements, and the magnified details on the lower pictures demonstrate qualitatively the superiority of our method over the competitors, indicating our simpler and efficient transformation. Fig. 8 depicts the 3D view of cortical modeling and 2D slices of the registration results, from which we can see that our method can ideally guarantee the topology of the

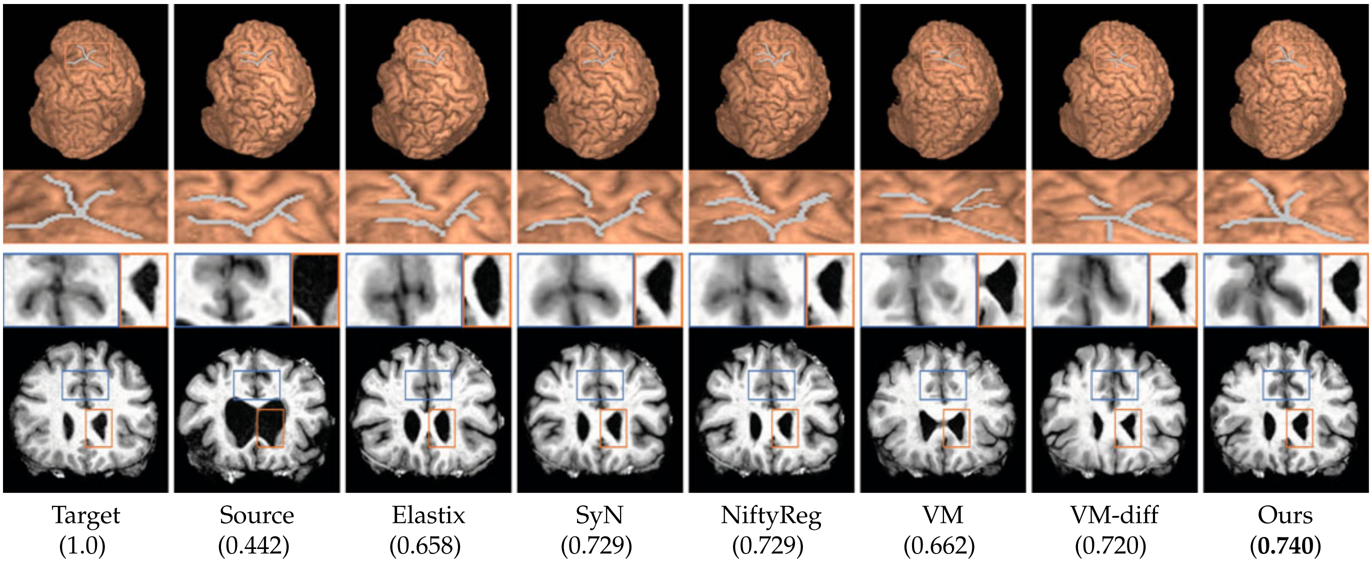


Fig. 8. The first row demonstrates cortex visualization and zoomed-in overlaid sulci registered using different methods. The second row gives the MR slices and zoomed-in warped patches. Parenthesized values in the bottom give the Dice score.

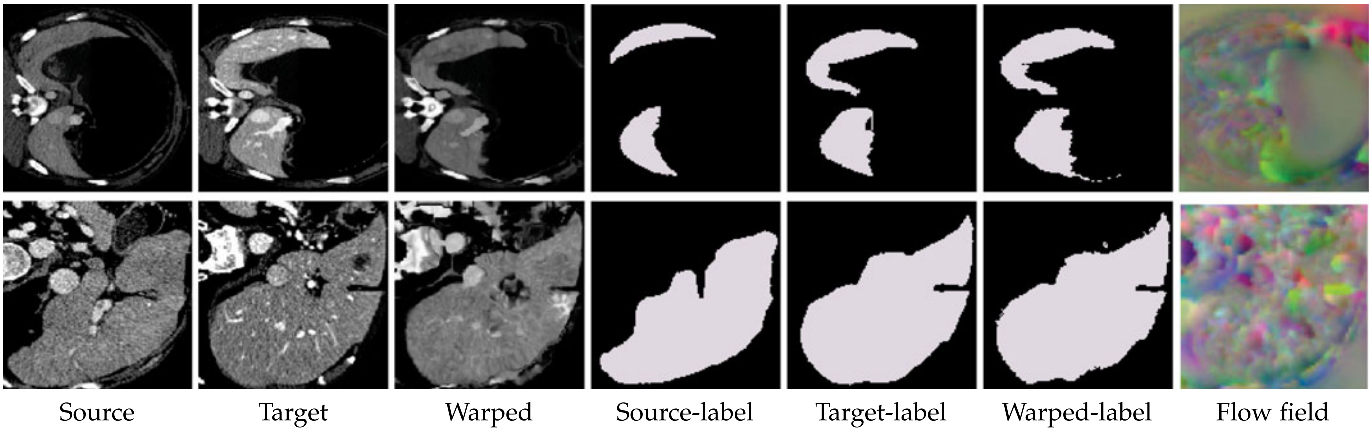


Fig. 9. Two example registration cases of liver CT data, including target, source, registered images, and the corresponding labels and flow fields.

TABLE 7
Qualitative Comparison Results on Liver CT Registration Tasks

Methods	Affine only	Elastix	NiftyReg	SyN	VM	VM-diff	Ours
SLIVER	0.794 ± 0.042	0.910 ± 0.038	**0.931 ± 0.031**	0.895 ± 0.037	0.883 ± 0.034	0.878 ± 0.042	0.910 ± 0.027
LSPIG	0.727 ± 0.054	0.825 ± 0.059	0.821 ± 0.122	0.825 ± 0.059	0.715 ± 0.009	0.788 ± 0.099	**0.855 ± 0.045**

The mean and standard deviations of the Dice score are listed.

registered volumes and preserve the contour of anatomical structure like cerebral cortex and ventricles.

5.4 Image-to-Image Registration

To show the generalizability, we extended our paradigm to the case of challenging image-to-image Liver CT registration tasks. We evaluated the accuracy, rationality of all the registration methods. Additionally, we propose an initial affine network to perform affine transform before predicting deformation fields, substituting the traditional affine preprocessing. We also demonstrated the benefits of the affine network.

Visual registration examples of liver CT scans are shown in Fig. 9, with the first raw data from LSPIG [31], covering intrasubject registration with the image pair which comes from the same pig (preoperative to perioperative), and others from SLIVER [58], covering intersubject registration for different persons. Although large deformations exist, source images are well aligned to the target, demonstrating the good registration performance of our approach. Table 7 depicts the stability of the methods in terms of the Dice score on the different datasets, where higher values and lower variance indicate a more accurate and stable registration. Our method gives an obvious lower variance with a comparable mean of Dice score, demonstrating superiority over the competitors.

Table 8 depicts the qualitative comparison between the network trained without and with the affine network under

TABLE 8
Ablation Analysis of the Affine Network on Two Liver CT Datasets in Terms of Dice Score and NCC

Methods		W/O Affine	W/ Affine
SLIVER	Dice score	0.858 ± 0.053	**0.910 ± 0.027**
	NCC	0.338 ± 0.032	**0.374 ± 0.035**
LSPIG	Dice score	0.814 ± 0.057	**0.855 ± 0.045**
	NCC	0.286 ± 0.032	**0.348 ± 0.056**

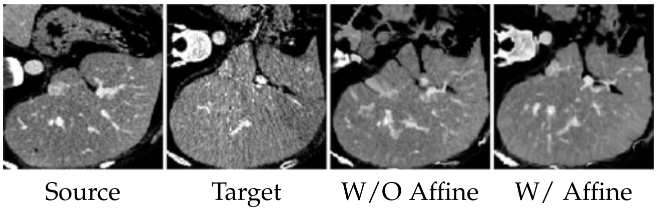


Fig. 10. Illustration of the importance of affine pre-processing for registration result. The source, target, warped image of our model trained without and with the affine network.

the experiment setting of liver CT registration. Fig. 10 gives the illustration of the impact of affine transformation on the final registration result. As shown, the registered image is more rational after introducing the affine network. Moreover, the affine network helps to achieve a much more accurate and robust alignment quality.

We reported the elapsed time for computations on brain MR and liver CT registration tasks. Cause of the data size, the runtime of registring the brain MR test pair is a little more than the liver CT scans. As shown in Table 9, on brain MRI registration task, conventional optimization-based SyN require two or more hours and NiftyReg requires roughly ten minutes, while the proposed method completes the registration for 3D medical image pairs in under half-second and runs orders of magnitude faster than conventional techniques. Owing to the well-designed lightweight network and dealing with half-resolution, we require less runtime even compared with learning-based methods.

TABLE 9
Average Registration Time in Second for Brain MR and Liver CT Test Pairs

Methods	Elastix	NiftyReg	SyN	VM	VM-diff	Ours
Brain MR	167	323	4799	0.69	0.58	**0.36**
Liver CT	115	53	748	0.20	0.19	**0.13**

The mean and standard deviations of the registration time are listed.

5.5 Applications on Multi-Modal Data

Multi-Modal Registration Experiments. We evaluate our method on datasets of brain scans acquired with T1 weighted and T2 weighted MRI, where T1 images do well to distinguish between different healthy tissues, whereas T2 images are best for highlighting abnormal structures in the brain such as tumors. Because of the complex intensity relationship between different modalities, in addition to default local normalized cross-correlation coefficient, we also use the multi-dimensional Modality Independent Neighborhood Descriptor (MIND) [64] as the loss function for the multi-modal registration. These self-similarity context-based descriptors may transform images into representations independent of the underlying image acquisition and have been frequently used in multi-modal registration.

Fig. 11 first illustrates the examples of the multi-modal registration problem and the result alignments. As shown, the source images and labels are well aligned to the targets, although large deformations and complex intensity relationship between the two modalities exist. Table 10 depicts the accuracy and stability of the methods in terms of the Dice score on the different multi-modal setting. Our method gives an obvious higher mean and a lower variance of Dice score on all the datasets, which indicates a more accurate and stable registration. These experiments demonstrate the capability of our framework to implicitly learn the complex similarity measure between different modalities.

Multi-Modal Fusion Experiments. We investigate the utility of our framework for assisting downstream image fusion tasks. Specifically, the registered multi-modal images with different registration algorithms are set as the input of the same fusion algorithm. We introduce an image fusion algorithm [65], to fuse properties of medical images of MRI T1

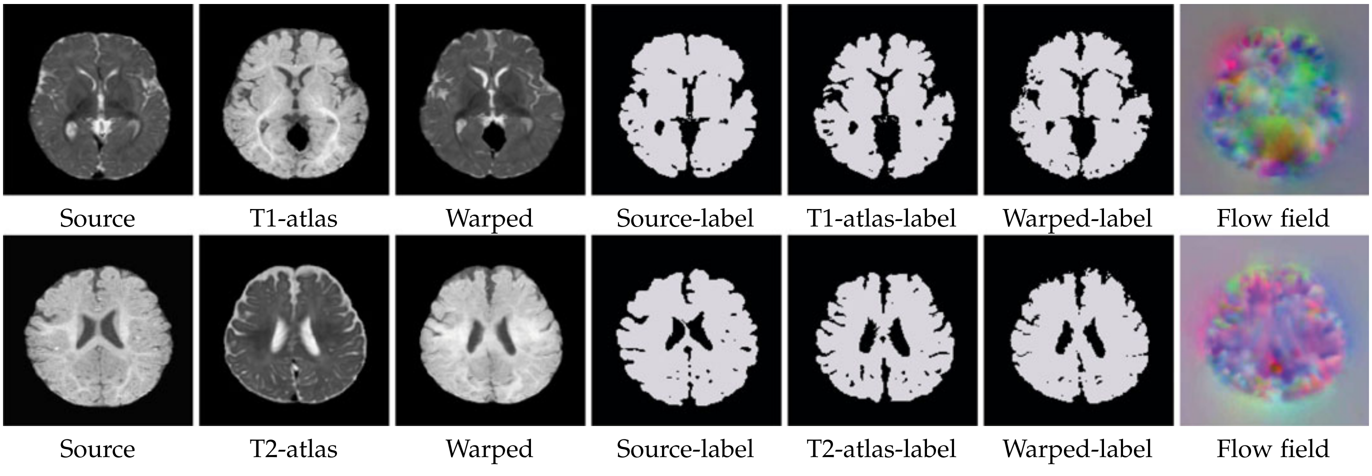


Fig. 11. Two example registration cases of multi-modal data. A clearly improved alignment of the ventricles is visible after registration.

TABLE 10
Qualitative Comparison Results for Multi-Modal Registration Tests

Methods	Affine only	Elastix	NiftyReg	SyN	VM	Ours	Ours-MIND
T1-T2atlas	0.539 ± 0.007	0.532 ± 0.014	0.619 ± 0.007	0.528 ± 0.011	0.579 ± 0.013	0.586 ± 0.009	**0.625 ± 0.009**
T2-T1atlas	0.539 ± 0.007	0.541 ± 0.016	0.639 ± 0.011	0.610 ± 0.010	0.579 ± 0.013	0.600 ± 0.014	**0.644 ± 0.007**

The mean and standard deviations of the Dice score are listed.

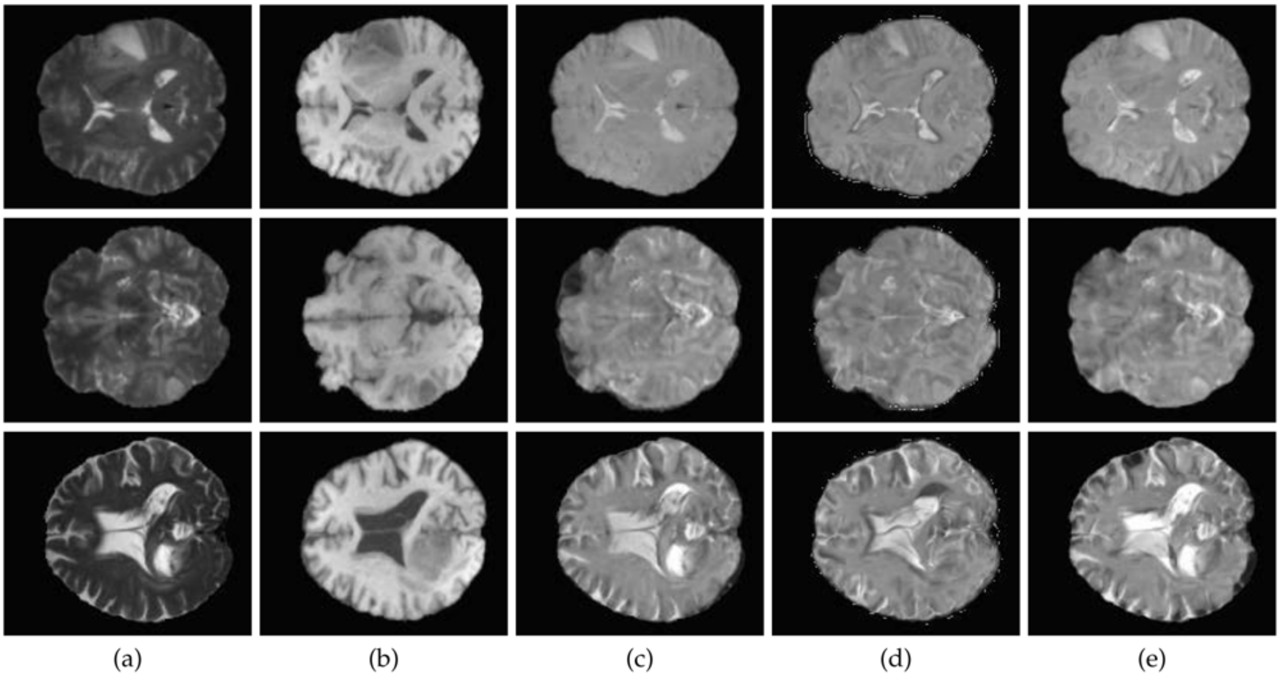


Fig. 12. Comparisons on multi-modal fusion results by fusing the source and target without and with registration pre-processing. The (a)-(e) columns refer to target, source, fusion result without registration, fusion result with NiftyReg registration method and fusion result with our registration method. Limited by space, we take the second-best NiftyReg as the representative comparison method (cf Table 10).

and T2 brain images. We trained this fusion algorithm on the MR T1 and T2 multi-modal data in BraTS18.

Fig. 12 shows three groups of medical images with T1 and T2 MR. T1 images contain anatomical structure details, while T2 images provide normal and pathological content. In column (c), the fused images without pre-registration processing suffer from low contrast and lose some weak details. In column (d), the fused images obtained with pre-registration using NiftyReg may improve structure details and contrast but can not decrease the artifacts. In column (e), the fused images obtained with pre-registration using our method can ideally improve structure details and contrast and greatly decrease the artifacts.

5.6 Applications on Medical Image Segmentation

We apply our framework to data augmentation of segmentation. To be specific, we simulate deformations by sampling from the feature distribution, these deformations can be used to register labeled atlas images to produce more labels images for segmentation tasks. We use NLU-Net [66] as the baseline method. We conduct experiments on the dataset that comes from the ISeg19 and contains images and segmentation labels from 10 healthy 6-month-old infants. By using the model without and with KL loss/feature distribution design, we may change the number of training data to get two experimental setups that use one training data (A) and six training data (B). Table 11 provides the segmentation accuracy under two experiment settings in terms of Dice score. Fig. 13 shows the example segmentation results of different variants. These experimental results indicate that thanks to the data augmentation using our registration model, the segmentation models outperform previous baseline models significantly.

TABLE 11
Ablation Analysis of Our Efficient Data Augmentation Strategy Under Two Segmentation Experiment Settings in Terms of Dice Score

Methods		NLU-Net	NLU-Net + Ours
(A)	CSF	0.902 ± 0.021	**0.919 ± 0.012**
	GM	0.885 ± 0.013	**0.900 ± 0.009**
	WM	0.851 ± 0.005	**0.877 ± 0.008**
	Average	0.879 ± 0.009	**0.899 ± 0.007**
(B)	CSF	0.934 ± 0.017	**0.939 ± 0.011**
	GM	0.903 ± 0.008	**0.913 ± 0.007**
	WM	0.888 ± 0.021	**0.894 ± 0.013**
	Average	0.908 ± 0.009	**0.915 ± 0.007**

The segmentation labels include cerebrospinal fluid (CSF), gray matter (GM), and white matter (WM).

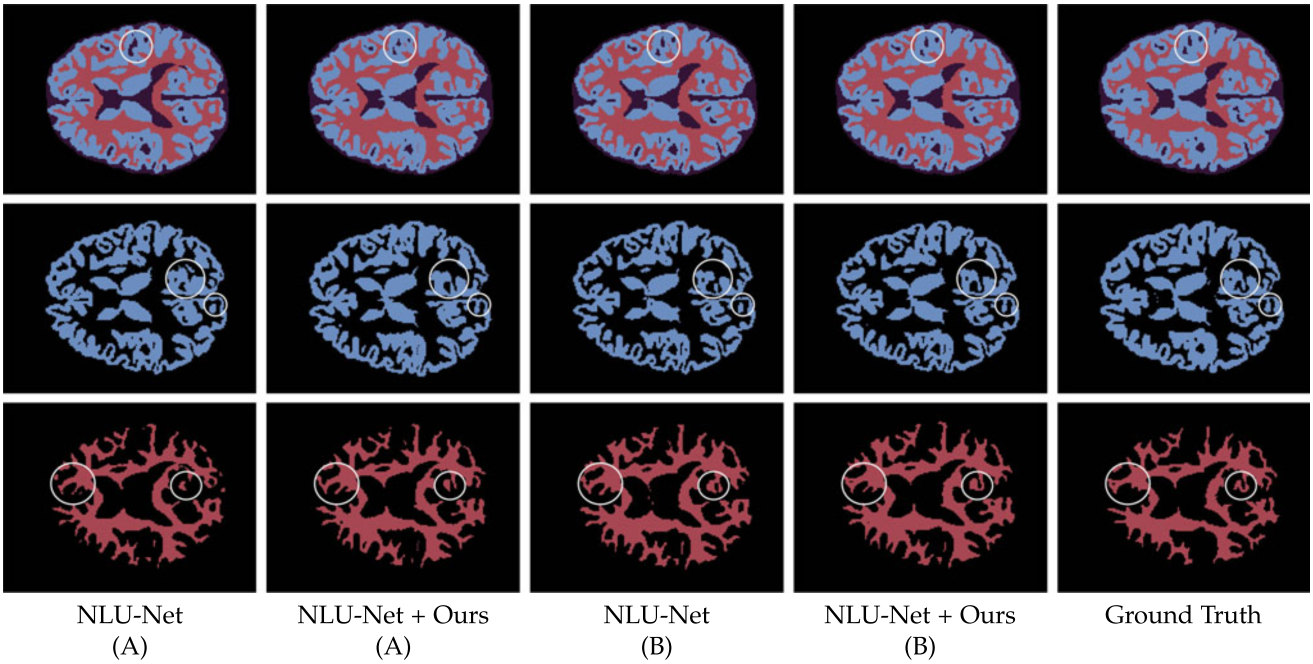


Fig. 13. Visualization of the segmentation results without and with data augmentation using our registration framework. The first row shows the original segmentation maps. The second and third rows show the segmentation maps for GM and WM. The (A) refers to the experiment setup using one training data, and (B) represents the case of six training data.

6 Conclusions and Future Work

We introduced a fundamental optimization model to formulate diffeomorphic registration and then established a learning framework to optimize it on a multi-scale feature space. This framework may propagate learned multi-scale features and deep parameters for optimization, and thus could render fast optimization without needing iteratively computing gradients on the image domain. We developed a series of deep modules to yield the multi-scale propagating process and to design the training objective. This optimization perspective could differentiate our framework from naively cascading registration networks, and provide a computational interpretation of network architectures that guarantees diffeomorphism. Moreover, we proposed our new bilevel self-tuned training, which allows the efficient search of the task-specific hyper-parameters and leads to the increased model flexibility and reduced computational burden. Extensive experiments on image-to-atlas and image-to-image registration tasks showed that our method achieved state-of-the-art performance with diffeomorphic guarantee and extreme efficiency. We also employed our framework to ideally solve challenging multi-modal registration tasks and investigated the utility of our framework to support the down-streaming image fusion and segmentation.

We demonstrate the performance on uni-modal and multi-modal registration tasks, and future validation remains on more challenging cross-modal registration and other scenarios where the moving and target exhibit more significant appearance differences.

Acknowledgments

The authors would like to thank Dr. Adrian V. Dalca at Massachusetts Institute of Technology for sharing the codes to generate the boxplots indicating Dice scores for anatomical structures.

References

[1] J. B. A. Maintz and M. A. Viergever, “A survey of medical image registration,” *Med. Image Anal.*, vol. 2, no. 1, pp. 1–36, 1998.

[2] A. Sotiras, C. Davatzikos, and N. Paragios, “Deformable medical image registration: A survey,” *IEEE Trans. Med. Imag.*, vol. 32, no. 7, pp. 1153–1190, Jul. 2013.

[3] J. Ashburner, “A fast diffeomorphic image registration algorithm,” *NeuroImage*, vol. 38, no. 1, pp. 95–113, 2007.

[4] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes, “Computing large deformation metric mappings via geodesic flows of diffeomorphisms,” *Int. J. Comput. Vis.*, vol. 61, no. 2, pp. 139–157, 2005.

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

[6] W. Sun, W. J. Niessen, and S. Klein, “Free-form deformation using lower-order B-spline for nonrigid image registration,” in *Proc. Med. Image Comput. Comput. Assist. Interv.*, 2014, pp. 194–201.

[7] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, “VoxelMorph: A learning framework for deformable medical image registration,” *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1788–1800, Aug. 2019.

[8] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, “Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces,” *Med. Image. Anal.*, vol. 57, pp. 226–236, 2019.

[9] Z. Shen, X. Han, Z. Xu, and M. Niethammer, “Networks for joint affine and non-parametric image registration,” in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2019, pp. 4224–4233.

[10] R. Liu, Z. Li, Y. Zhang, X. Fan, and Z. Luo, “Bi-level probabilistic feature learning for deformable image registration,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2020, pp. 723–730.

[11] J. Wang and M. Zhang, “DeepFLASH: An efficient network for learning-based medical image registration,” in *Proc. Comput. Vis. Pattern Recognit.*, 2020, pp. 4443–4451.

[12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[13] C. Buerger, T. Schaeffter, and A. P. King, “Hierarchical adaptive local affine registration for fast and robust respiratory motion estimation,” *Med. Image Anal.*, vol. 15, no. 4, pp. 551–564, 2011.

[14] J. Zhang *et al.*, "Content-aware unsupervised deep homography estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 653–669.

[15] X. Pennec, R. Stefanescu, V. Arsigny, P. Fillard, and N. Ayache, "Riemannian elasticity: A statistical regularization framework for non-linear registration," in *Proc. Med. Image Comput. Comput.-Assist. Interv.*, 2005, pp. 943–950.

[16] M. Chiang *et al.*, "Fluid registration of diffusion tensor images using information theory," *IEEE Trans. Med. Imag.*, vol. 27, no. 4, pp. 442–456, Apr. 2008.

[17] T. Mansi, X. Pennec, M. Sermesant, H. Delingette, and N. Ayache, "iLogDemons: A demons-based registration algorithm for tracking incompressible elastic biological tissues," *Int. J. Comput. Vis.*, vol. 92, no. 1, pp. 92–111, 2011.

[18] B. Beuthien, A. Kamen, and B. Fischer, "Recursive green's function registration," in *Proc. Med. Image Comput. Comput.-Assist. Interv.*, 2010, pp. 546–553.

[19] X. Yang, Y. Li, D. Reutens, and T. Jiang, "Diffeomorphic metric landmark mapping using stationary velocity field parameterization," *Int. J. Comput. Vis.*, vol. 115, no. 2, pp. 69–86, 2015.

[20] A. Pai, S. Sommer, L. Sørensen, S. Darkner, J. Sporring, and M. Nielsen, "Kernel bundle diffeomorphic image registration using stationary velocity fields and Wendland basis functions," *IEEE Trans. Med. Imag.*, vol. 35, no. 6, pp. 1369–1380, Jun. 2016.

[21] S. Tang, Y. Fan, G. Wu, M. Kim, and D. Shen, "RABBIT: Rapid alignment of brains by building intermediate templates," *NeuroImage*, vol. 47, no. 4, pp. 1277–1287, 2009.

[22] M. Kim, G. Wu, P. Yap, and D. Shen, "A general fast registration framework by learning deformation-appearance correlation," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1823–1833, 2012.

[23] C. C. Brun *et al.*, "A nonconservative Lagrangian framework for statistical fluid registration - SAFIRA," *IEEE Trans. Med. Imag.*, vol. 30, no. 2, pp. 184–202, Feb. 2011.

[24] E. Konukoglu *et al.*, "Image guided personalization of reaction-diffusion type tumor growth models using modified anisotropic Eikonal equations," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 77–95, Jan. 2010.

[25] N. S. Phatak *et al.*, "Strain measurement in the left ventricle during systole with deformable image registration," *Med. Image Anal.*, vol. 13, no. 2, pp. 354–361, 2009.

[26] X. Yang, R. Kwitt, M. Styner, and M. Niethammer, "Quicksilver: Fast predictive image registration - A deep learning approach," *NeuroImage*, vol. 158, pp. 378–396, 2017.

[27] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 1647–1655.

[28] R. Liu, Y. Zhang, S. Cheng, X. Fan, and Z. Luo, "A theoretically guaranteed deep optimization framework for robust compressive sensing MRI," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, pp. 4368–4375, 2019.

[29] Y. Zhang, R. Liu, Z. Li, Z. Liu, X. Fan, and Z. Luo, "Coupling principled refinement with bi-directional deep estimation for robust deformable 3D medical image registration," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2020, pp. 86–90.

[30] M. Jaderberg *et al.*, "Spatial transformer networks," in *Proc. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[31] S. Zhao, Y. Dong, E. I. Chang, and Y. Xu, "Recursive cascaded networks for unsupervised medical image registration," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 10599–10609.

[32] A. Hering, B. van Ginneken, and S. Heldmann, "mlVIRNET: Multilevel variational image registration network," in *Proc. Med. Image Comput. Comput. Assist. Interv.*, 2019, pp. 257–265.

[33] T. C. W. Mok and A. C. S. Chung, "Fast symmetric diffeomorphic image registration with convolutional neural networks," in *Proc. Comput. Vis. Pattern Recognit.*, 2020, pp. 4643–4652.

[34] D. Sun, X. Yang, M. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. Comput. Vis. Pattern Recognit.*, 2018, pp. 8934–8943.

[35] T.-W. Hui, X. Tang, and C. C. Loy, "LiteFlowNet: A lightweight convolutional neural network for optical flow estimation," in *Proc. Comput. Vis. Pattern Recognit.*, 2018, pp. 8981–8989.

[36] J. Hur and S. Roth, "Iterative residual refinement for joint optical flow and occlusion estimation," in *Proc. Comput. Vis. Pattern Recognit.*, 2019, pp. 5754–5763.

[37] S. Meister, J. Hur, and S. Roth, "UnFlow: Unsupervised learning of optical flow with a bidirectional census loss," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7251–7259.

[38] K. Luo, C. Wang, S. Liu, H. Fan, J. Wang, and J. Sun, "UPFlow: Upsampling pyramid for unsupervised optical flow learning," 2020, *arXiv:2012.00212*.

[39] M. Niethammer, R. Kwitt, and F.-X. Vialard, "Metric learning for image registration," in *Proc. Comput. Vis. Pattern Recognit.*, 2019, pp. 8463–8472.

[40] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Representations*, 2014.

[41] L. Franceschi, M. Donini, P. Frasconi, and M. Pontil, "Forward and reverse gradient-based hyperparameter optimization," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1165–1173.

[42] M. MacKay, P. Vicol, J. Lorraine, D. Duvenaud, and R. B. Grosse, "Self-tuning networks: Bilevel optimization of hyperparameters using structured best-response functions," in *Proc. Int. Conf. Learn. Representations*, 2019.

[43] R. Liu, P. Mu, X. Yuan, S. Zeng, and J. Zhang, "A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 6305–6315.

[44] R. Liu, S. Cheng, Y. He, X. Fan, Z. Lin, and Z. Luo, "On the convergence of learning-based iterative methods for nonconvex inverse problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 12, pp. 3027–3039, Dec. 2020.

[45] R. Liu, J. Liu, Z. Jiang, X. Fan, and Z. Luo, "A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion," *IEEE Trans. Image Process.*, vol. 30, pp. 1261–1274, Dec. 2021.

[46] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, "Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement," 2020, *arXiv:2012.05609*.

[47] R. Liu, J. Gao, J. Zhang, D. Meng, and Z. Lin, "Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond," 2021, *arXiv:2101.11517*.

[48] H. Liu, K. Simonyan, and Y. Yang, "DARTS: Differentiable architecture search," in *Proc. Int. Conf. Learn. Representations*, 2019.

[49] S. G. Mueller *et al.*, "Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's disease neuroimaging initiative (ADNI)," *Alzheimer's Dement.*, vol. 1, no. 1, pp. 55–66, 2005.

[50] A. Di Martino *et al.*, "The Autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in Autism," *Mol. Psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.

[51] K. Marek *et al.*, "The Parkinson progression marker initiative (PPMI)," *Prog. Neurobiol.*, vol. 95, no. 4, pp. 629–635, 2011.

[52] D. S. Marcus, A. F. Fotenos, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies: Longitudinal MRI data in nondemented and demented older adults," *J. Cogn. Neurosci.*, vol. 22, no. 12, pp. 2677–2684, 2010.

[53] D. C. V. Essen, S. M. Smith, D. M. Barch, T. E. J. Behrens, E. Yacoub, and K. Ugurbil, "The WU-Minn human connectome project: An overview," *NeuroImage*, vol. 80, pp. 62–79, 2013.

[54] B. Fischl, "Freesurfer," *NeuroImage*, vol. 62, no. 2, pp. 774–781, 2012.

[55] M. W. Woolrich *et al.*, "Bayesian analysis of neuroimaging data in FSL," *NeuroImage*, vol. 45, no. 1, pp. S173–S186, 2009.

[56] Medical Segmentation Decathlon Datasets. Accessed: Oct. 20, 2021. [Online]. Available: http://medicaldecathlon.com/

[57] S. Zhao, T. F. Lau, J. Luo, E. I. Chang, and Y. Xu, "Unsupervised 3D end-to-end medical image registration with volume tweening network," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 5, pp. 1394–1404, May 2020.

[58] T. Heimann *et al.*, "Comparison and evaluation of methods for liver segmentation from CT datasets," *IEEE Trans. Med. Imag.*, vol. 28, no. 8, pp. 1251–1265, Aug. 2009.

[59] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[60] L. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[61] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: A toolbox for intensity-based medical image registration," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 196–205, Jan. 2010.

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

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

[64] M. P. Heinrich *et al.*, "MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration," *Med. Image Anal.*, vol. 16, no. 7, pp. 1423–1435, 2012.

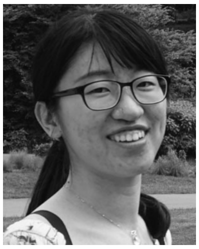
[65] H. Li and X. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.

[66] Z. Wang, N. Zou, D. Shen, and S. Ji, "Non-local U-Nets for biomedical image segmentation," in *Proc. Conf. Artif. Intell.*, 2020, pp. 6315–6322.

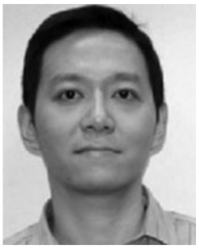
Risheng Liu (Member, IEEE) received the BSc and PhD degrees from the Dalian University of Technology, China, in 2007 and 2012, respectively. From 2010 to 2012, he was doing research as joint PhD with Robotics Institute, Carnegie Mellon University. From 2016 to 2018, he was doing research as Hong Kong Scholar with the Hong Kong Polytechnic University. He is currently a full professor with the Digital Media Department, International School of Information Science & Engineering, Dalian University of Technology. His research interests include optimization, computer vision, and multimedia. He was the recipient of the Outstanding Youth Science Foundation of the National Natural Science Foundation of China. He was the editor of the *Journal of Electronic Imaging* (Senior Editor), *The Visual Computer*, and the *IET Image Processing*.

Zi Li received the BE degree in software engineering in 2019 from the Dalian University of Technology, Dalian, China, where she is currently working toward the master's degree in software engineering. Her research interests include medical image analysis, computer vision, and deep learning.

Xin Fan (Senior Member, IEEE) received the BE and PhD degrees in information and communication engineering from Xian Jiaotong University, Xian, China, in 1998 and 2004, respectively. From 2006 to 2007, he was a postdoctoral research fellow with Oklahoma State University, Stillwater. In 2009, he was with the School of Software, Dalian University of Technology, Dalian, China. His research interests include computational geometry and machine learning, and their applications to low-level image processing and DTI-MR image analysis.



Chenying Zhao received the BEng degree in biomedical engineering from Tsinghua University in 2017. She is currently working toward the PhD degree in bioengineering with the University of Pennsylvania. Her research interests include diffusion MRI, brain connectome, medical image analysis, and applications in pediatric population. She was the recipient of the International Society for Magnetic Resonance in Medicine Summa Cum Laude Merit Award in 2020.



Hao Huang received the BE degree from Tsinghua University, China, in 1996, the MS degree from Peking University, China, in 1999, and the MSE degree in computer and electrical engineering and the PhD degree in biomedical engineering from the Johns Hopkins University in 2004 and 2005, respectively. From 2005 to 2007, he was a research associate faculty with the Johns Hopkins University School of Medicine. In 2007, he was an assisant professor with the University of Texas Southwestern Medical Center. Since 2014, he has been a faculty member with the University of Pennsylvania, where he was an associate professor in 2014 and a professor in 2021. His research interests include pushing the technical boundaries of neural magnetic resonance imaging (MRI) in health and disease, including advanced MR acquisition and analysis techniques in diffusion MRI, functional MRI, and perfusion MRI. He was in a number of leadership positions in international committees. He has been recognized as the distinguished investigator of the Academy for Radiology and Biomedical Imaging Research in 2019. He has been elected as the fellow of American Institute of Medical and Biological Engineering (AIMBE) in 2021.

Zhongxuan Luo received the BS and MS degrees in computational mathematics from Jilin University, China, in 1985 and 1988, respectively, and the PhD degree in computational mathematics from the Dalian University of Technology, China, in 1991. Since 1997, he has been a full professor with the School of Mathematical Sciences, Dalian University of Technology. His research interests include computational geometry and computer vision.