Bi-level Probabilistic Feature Learning for Deformable Image Registration

Risheng Liu[1,2] , **Zi Li**[1,2] , **Yuxi Zhang**[1,2] , **Xin Fan**[1,2,*] and **Zhongxuan Luo**[2,3,4]
[1]International School of Information Science & Engineering, Dalian University of Technology
[2]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province
[3]School of Software Technology, Dalian University of Technology
[4]Institute of Artificial Intelligence, Guilin University of Electronic Technology
{rsliu, xin.fan, zxluo}@dlut.edu.cn, alisonbrielee@gmail.com, yuxizhang@mail.dlut.edu.cn

Abstract

We address the challenging issue of deformable registration that robustly and efficiently builds dense correspondences between images. Traditional approaches upon iterative energy optimization typically invoke expensive computational load. Recent learning-based methods are able to efficiently predict deformation maps by incorporating learnable deep networks. Unfortunately, these deep networks are designated to learn deterministic features for classification tasks, which are not necessarily optimal for registration. In this paper, we propose a novel bi-level optimization model that enables jointly learning deformation maps and features for image registration. The bi-level model takes the energy for deformation computation as the upper-level optimization while formulates the maximum *a posterior* (MAP) for features as the lower-level optimization. Further, we design learnable deep networks to simultaneously optimize the cooperative bi-level model, yielding robust and efficient registration. These deep networks derived from our bi-level optimization constitute an unsupervised end-to-end framework for learning both features and deformations. Extensive experiments of image-to-atlas and image-to-image deformable registration on 3D brain MR datasets demonstrate that we achieve state-of-the-art performance in terms of accuracy, efficiency, and robustness.

1 Introduction

Image registration [Maintz and Viergever, 1998], transforming different sets of images into one common coordinate system, is one of the most fundamental tasks in computer vision. Especially, deformable image registration which builds a dense correspondence between image pairs has wide applications in medical image analysis such as multi-modality fusion, anatomical change diagnosis, and population modeling. 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 image registration approaches typically formulate the discrepancy between image pairs as well as the prior constraints on images and deformations into an energy function whose optimization finds the solution to registration [Ashburner, 2007]. These approaches are able to generate plausible deformation maps with desired mathematical properties, *e.g.*, invertibility, and topology-preserving, even when significant appearance variations exist. Unfortunately, the iterative optimization process demands calculating gradients over the high dimensional deformation and image/feature spaces at each step, resulting in extremely high computational expenses.

Recently, researchers bring deep learning networks that gain great success in classification/segmentation to image registration in order to achieve efficient prediction for deformations. Balakrishnan *et al.* develop the VoxelMorph (VM) network similar to UNet structure [Balakrishnan *et al.*, 2019], and the works of [Dalca *et al.*, 2019] and [Shen *et al.*, 2019] combine the learning with the diffeomorphic constraint for topology-preserving deformations. Hu *et al.* [Hu *et al.*, 2019] develop a multi-scale network to compute the registration field from convolutional feature pyramids. These works learn deep features in the image domain and then apply the learned parameters as the one-step prediction for deformation. They just directly learn parameters based on predesigned training loss/regularization to output the deformation filed. So it is hard to adaptively enforce registration information for the front-end feature learning phase. Consequently, images deviating from the training greatly affect the performance, yielding unstable registration.

In this work, deformation learning embraces feature learning in a bi-level optimization framework. We formulate the energy for deformation in the *feature* domain as one level of optimization embedded with the other level of maximum *a posterior* (MAP) optimization for *features*. Hence, two challenging issues in learning-based approaches, *i.e.,*, learning optimal networks for deformation and learning optimal features for registration, are cooperatively resolved in one model. We detail our contributions as:

- We establish a bi-level model to simultaneously address deformable registration and feature learning. The upper-level of optimization learns the deformation field in the feature domain while the lower-level learns the features required for image registration. We alternately solve the deformation filed based on the upper-level subproblem and adjust the features based on the lower-level subprob-

lem, giving robust and efficient registration.

- We pose the feature learning in the lower-level as probabilistic MAP estimation with Gaussian priors. This level of optimization adaptively enforces task-based constraints for the feature extraction phase, producing optimal features for the upper-level optimization.
- We develop an unsupervised end-to-end training scheme for the deep networks respective to the two levels of optimization. The scheme, training the networks using the loss function on the feature domain and image domain respectively, runs more efficiently than training a complicated image-to-deformation network.

Extensive experiments on 3D brain MRI registration tasks demonstrate that our approach achieves state-of-the-art performance in terms of accuracy, robustness, and efficiency.

2 Related Work

Recently, feature learning has been employed in many image analysis tasks. Taking advantage of the network to learn the high-level features has shown impressive results [Liu *et al.*, 2018a] [Liu *et al.*, 2019a]. Research in [Banerjee *et al.*, 2019] uses a conditional encoder-decoder network for the segmentation task. The work of [Sun *et al.*, 2018] estimates the optical flow on the multi-scale feature space. Replacing handcrafted features with more discriminative learnable feature space has been extensively developed [Cheng *et al.*, 2018] [Liu *et al.*, 2019b] [Liu *et al.*, 2018b]. Researchers are tending towards superseding the feature engineering or original image with learned convolution filters, within end-to-end trainable architectures. The feature learning designed for the registration task has also been explored in [Hu *et al.*, 2019] [de Vos *et al.*, 2019] [Liu *et al.*, 2020]. However, these networks are designed to learn deterministic features.

Traditional deformable registration techniques include the b-splines model with control points [Sun *et al.*, 2014], elastic-type models [Bajcsy and Kovacic, 1989] and diffeomorphic registration. Diffeomorphic frameworks [Ashburner, 2007] use smooth flow fields to represent the deformation, regularization is typically introduced as part of the ordinary differential equation constraining on the vector fields. Due to a large number of parameters, numerical optimization becomes computationally costly, which is the main limitation of these traditional methods. Recently, learning-based methods [Dosovitskiy *et al.*, 2015] have been widely applied. The research in [Yang *et al.*, 2017] proposes a supervised learning approach to rapidly predict 3D deformable registrations, achieving an order of magnitude speed-up. But ground truth registration fields are hard and expensive to obtain. Inspired by the VoxelMorph [Balakrishnan *et al.*, 2019], researchers have focused on replacing costly numerical optimization with global function optimization over the training data in an unsupervised way. The research in [Hu *et al.*, 2019] employs an unsupervised coarse-to-fine dual-stream registration network, enabling the capability for handling significant deformations. Moreover, some researches[Shen *et al.*, 2019] propose to estimate the velocity fields or momentum fields, which can be used to obtain diffeomorphic transformations.

3 The Proposed Method

In this section, we first introduce a novel bi-level framework to incorporate feature optimizing strategy into deformable image registration. Bi-level optimization [Colson *et al.*, 2005] is a special optimization scheme, in which one optimization task is nested within another. The outer level is given by the constraint inner level problem, referred as upper-level tasks and lower-level tasks respectively. Then, we propose our deep architectures, the probabilistic feature learning module and the deformable registration module to solve the two nested tasks.

3.1 Bi-level Registration Framework

The proposed bi-level model jointly optimizes the features and registration fields. In this novel model, the upper-level problem denotes the deformable registration, while the lower-level problem is a maximum posterior probability problem. The outer deformable registration problem is constrained by the inner feature optimization. Specifically, given a source image $\mathbf{I_s}$ and a target image $\mathbf{I_t}$ with a spatial domain $\Omega \in \mathbb{R}^d$, we aim at minimizing:

$$\begin{aligned} &\min_{\boldsymbol{\varphi}} E_D(\boldsymbol{\varphi}; \mathbf{f_s}, \mathbf{f_t}) + E_R(\boldsymbol{\varphi}), \\ &s.t.\ \mathbf{f_s}, \mathbf{f_t} = \arg\max_{\mathbf{f_s}, \mathbf{f_t}} p(\mathbf{f_s}|\mathbf{I_s}, \mathbf{f_t}|\mathbf{I_t}, \boldsymbol{\varphi}), \end{aligned} \tag{1}$$

where $\mathbf{f_s}, \mathbf{f_t}$ are the feature representations of the source and target image, $\boldsymbol{\varphi} : \Omega \times \mathbb{R} \rightarrow \Omega$ is the final deformation field. E_D is a data matching term, forcing the image similarity. E_R is a regularization term, encouraging the smoothness of the deformation fields. $p(\mathbf{f_s}|\mathbf{I_s}, \mathbf{f_t}|\mathbf{I_t}, \boldsymbol{\varphi})$ denotes the posterior probability of features based on the observation input source, target image and the deformation field $\boldsymbol{\varphi}$. Note that, to better solve the deformable registration problem, we employ the stationary velocity fields. Such that the final registration field $\boldsymbol{\varphi}$ is obtained via integration of a series of velocity fields over time. We employ regularization into the registration fields by introducing the integration of the velocity fields.

The proposed bi-level model considers the feature optimization and deformable registration model simultaneously, which is different from other models in essence. Our formulation is much more robust than the typical registration model since it takes feature optimization into consideration.

3.2 Probabilistic Feature Learning Module

To learn the most suitable feature representations that are invariant to noise and uninformative intensity-variations, we propose to learn the feature via MAP estimation, corresponding to the lower-level problem in Eq. (1). Specially, we use the posterior probability to learn the most meaningful nonlinear mapping from the input intensities to feature representation. We aim to estimate the most suitable feature for the registration task as follows:

$$\mathbf{f} = \arg\min_{\mathbf{f}} \ln p(\mathbf{f}|\mathbf{I}, \boldsymbol{\varphi}), \tag{2}$$

$$= \arg\min_{\mathbf{f}} \ln p(\mathbf{I}|\mathbf{f}, \boldsymbol{\varphi}) + \ln p(\mathbf{f}), \tag{3}$$

where $\boldsymbol{\varphi}$ is the corresponding registration field, and $\mathbf{I}$ is the input source or target image. The first term is the data likelihood term and the second one is the prior term.

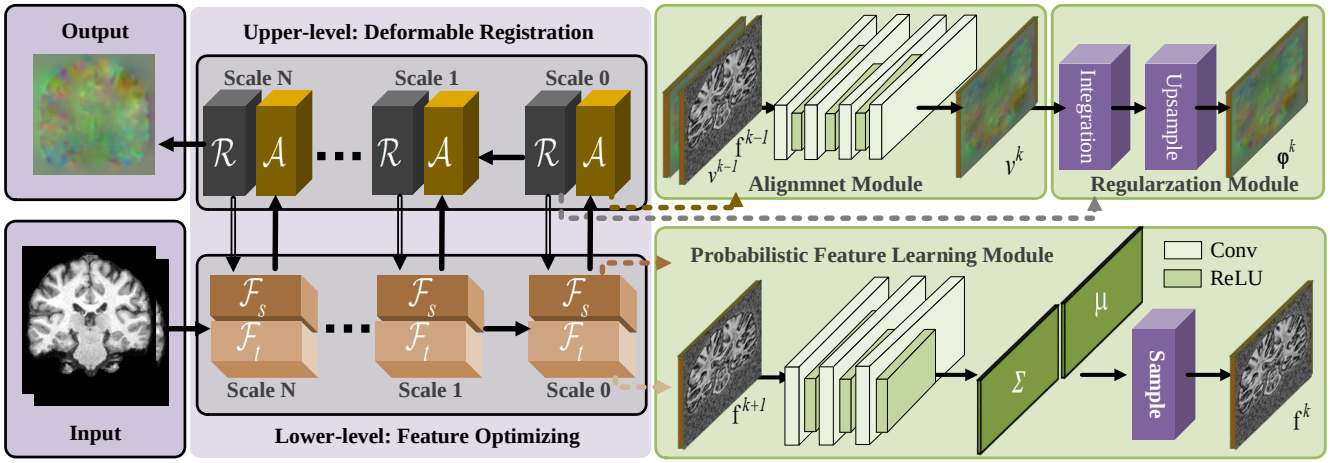

Figure 1: The pipeline of our bi-level framework. The lower-level task learns a reliable feature representation via MAP estimation. The learned features are employed to estimate the deformation field in the upper-level. The feature learning and deformation fields are updated in an alternative and collaborative manner. The single lines indicate the direction of the data propagation. The double lines indicate the backpropagation of the loss.

We assume the prior $p(\mathbf{f})$ to be a multivariate unit normal distribution with covariance I. However, the posterior probability is difficult to compute, we use an approximation posterior probability $q_\theta(\mathbf{f})$ parametrized by θ to speed it up. We try to minimize the Kullback-Leibler (KL) divergence between the really posterior distribution and approximation distribution, equivalent to maximizing the Evidence Lower Bound (ELBO) [Kingma and Welling, 2014] as follows:

$$\max_{\theta} E_q[\ln p(\mathbf{I}|\mathbf{f}, \boldsymbol{\varphi})] - KL[q_\theta(\mathbf{f})||p(\mathbf{f})], \tag{4}$$

where the first term means samples from the approximate distribution to perform the registration tasks. The second term is the KL term, also called Gauss prior term, which encourages the two probability distribution to be close. We model the approximate posterior $q_\theta(\mathbf{f})$ as a multivariate normal:

$$q_\theta(\mathbf{f}) = \mathcal{N}(\mathbf{f}; \mu_\mathbf{f}, \mathbf{\Sigma_f}), \tag{5}$$

where $\mathcal{N}(.; \mu, \mathbf{\Sigma})$ is the multivariate normal distribution with the mean μ and covariance $\mathbf{\Sigma}$. Then we optimize the above evidence lower bound.

We optimize the variational lower bound by learning the network parameters of the probabilistic feature learning networks using backpropagation methods. The feature learning networks take the input images/features and output the approximate posterior probability parameters of feature representations, representing mean $\mu_\mathbf{f}$ and covariance $\mathbf{\Sigma_f}$, then sample a new feature, which is processed to generate the deformation fields at different scale. Specifically, to generate feature representations/distribution at the kth scale, we use layers of convolutional filters to downsample the features at the previous $k+1$th pyramid level, by a factor of 2. We employ three-layers feature learning network, which includes convolutional layers with 16 filters or 32 filters. Each convolutional layer uses leaky ReLU function and $3 \times 3 \times 3$ kernel.

3.3 Deformable Registration Module

Diffeomorphic deformable registration [Ashburner, 2007] can capture large deformations and provide many desirable mathematical properties, such as invertibility, globally one-to-one smooth, and topology-preserving. In this work, we develop an optimization scheme to solve the diffeomorphic deformable registration, corresponding to the upper-level problem in Eq. (1), as follows:

$$\mathcal{V}^{k+1} = \arg\min_{\mathcal{V}} E_D(\boldsymbol{\varphi}^k; \mathbf{f}_\mathbf{s}^k, \mathbf{f}_\mathbf{t}^k) + E_R(\mathcal{V}), \tag{6}$$

where each deformation field $\boldsymbol{\varphi}$ is obtained via the integration of current $\mathcal{V}$, governed by the ordinary differential equation $\boldsymbol{\phi}_t(x,t) = \mathcal{V}(\boldsymbol{\phi}(x,t),t)$ over time t, with identity map $\boldsymbol{\phi}(0) = Id$. And $\boldsymbol{\phi}_1$ is the sought-for registration field, with k standing for scale.

To generate the velocity field sequence, we employ an efficient deep residual architecture, named by the alignment network, to solve the Eq. (6) as follows:

$$\mathcal{V}^{k+1} = \mathcal{V}^k - \mathcal{A}(\boldsymbol{\varphi}^k, \mathbf{f}_\mathbf{s}^k, \mathbf{f}_\mathbf{t}^k;\ \mathcal{W}^k), \tag{7}$$

where $\mathcal{W}^k$ are the learnable parameters of the alignment network, and the deformation field $\boldsymbol{\varphi}$ is defined through the ordinary differential equation of $\mathcal{V}$, corresponding to regularization module. We implement the regularization module using the efficient scaling and squaring [Dalca *et al.*, 2019] method. As for the alignment network, it processes the features and the deformation field from the previous scale to generate the current deformation field. The alignment networks at different blocks share the same network structure, including three convolutional layers with $\{48, 32, 16\}$ filters respectively. All convolutional layers are followed by a leaky ReLU activation function except the one that outputs the registration field. And we use the $3 \times 3 \times 3$ kernel for all the convolutional layers.

Dice score	Base	Elastix	NiftyReg	ANTs (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 [1]	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)**
PPMI [2]	0.613 (0.057)	0.724 (0.034)	0.777 (0.030)	0.773 (0.026)	0.757 (0.035)	0.765 (0.023)	**0.778 (0.023)**

Table 1: Qualitative comparison between our framework and other methods. The higher Dice score indicates the more accurate alignment. The first column shows the affine results. The last two rows give the Dice scores on the unseen PPMI dataset for image-to-atlas [1] and image-to-image [2] registration. Standard deviations are in bracket.

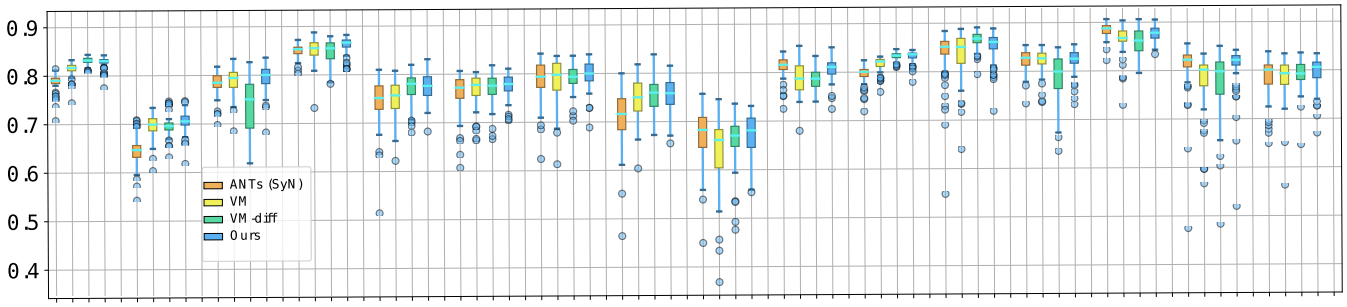

Figure 2: Boxplot indicates the Dice scores for ANTs (SyN), VM, VM-diff and our algorithm over sixteen anatomical structures including Cerebral White Matter (CblmWM), Cerebral Cortex (CblmC), Lateral Ventricle (LV), Inferior Lateral Ventricle (ILV), Cerebellum White Matter (CeblWM), Cerebellum Cortex (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).

4 End-to-end Training Strategy

All the modules are differentiable, thus we learn the network parameters with the end-to-end training. The proposed propagation network is jointly trained using stochastic gradient descent-based methods in an unsupervised way. The loss functions are as follows:

Loss on feature space. The training loss on feature space consists of the task-specific reconstruction loss and the KL loss, forcing the encoded distribution to be close to the prior probability distribution. The hypothesis is that to better optimize the reconstruction loss, the feature learning networks at each scale make use of the provided posterior probability information of features, resulting in the feature optimizing. For the task-specific reconstruction loss, we scale the image pairs and warp [Jaderberg *et al.*, 2015] the down-sampled images with the deformation fields at different scales, then compute the similarity loss using the Local Normalized Cross Correlation. Therefore, the deformation fields at each scale can influence the previous featuring learning. Specifically, we use different window sizes to compute the local normalized correlation coefficient at each scale, with a smaller window size for the lower resolution. From the zeroth to the second scale, the window sizes are set to $\{5, 5, 7\}$, respectively. For the KL term, we obtain $\mathbf{\Sigma_f}$ and $\mu_\mathbf{f}$ from the feature learning module, then compute it in the closed-form as:

$$KL[q_\theta(\mathbf{f})||p(\mathbf{f})] = 1/2(tr(\mathbf{\Sigma_f}) + ||\mu_\mathbf{f}|| - \ln det(\mathbf{\Sigma_f}) - k), \tag{8}$$

with $\mathbf{\Sigma_f}$ to be diagonal, and k is const. We use the multi-scale training loss, computing the sum of the losses at different scales.

Loss on final deformation field. We also consider the information on the image domain, using the final deformation field as follows:

$$L(\mathbf{I_s}, \mathbf{I_t}; \boldsymbol{\varphi}) = L_{NCC}(\mathbf{I_s} \circ \boldsymbol{\varphi}, \mathbf{I_t}) + L_{Reg}(\boldsymbol{\varphi}), \tag{9}$$

we take the Local Normalized Cross Correlation as image similarity loss, with the window size as 9. For the regularization term, we also employ a diffusion regularizer on the final deformation field.

5 Experiments

In this section, we first introduce our experiment setup, including datasets, evaluation metrics, baseline methods, and our implementation. Next, to demonstrate the superiority of our method, we compare it with the state-of-the-art deformable registration techniques on the accuracy, efficiency as well as diffeomorphism preservation of the deformation fields. Then we explore the impact of the inner task of our paradigm.

Data Preparation. Evaluations are conducted in two different sights, one aligning all the source data to a common atlas, called image-to-atlas registration, and the other addressing general registration between two arbitrary volumes, called images-to-image registration. For both of these two cases, 368 T1 weighted MR volumes from three publicly available datasets: ADNI [Mueller *et al.*, 2005], ABIDE [Di Martino *et al.*, 2014] and OASIS [Marcus *et al.*, 2007] are selected and split into 281, 17, and 70 for training, validation, and

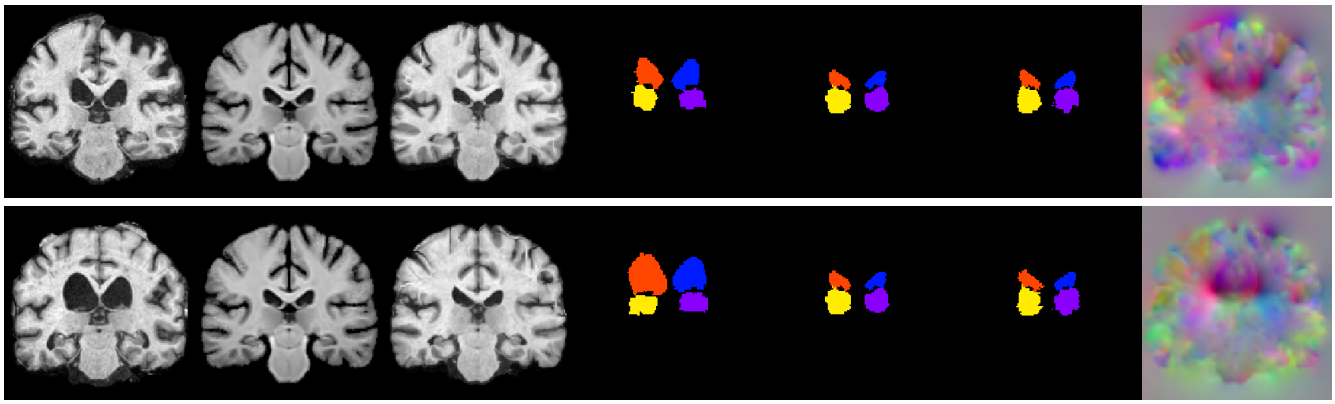

Figure 3: Example MR coronal slices of input source, target, and warped image for our method with corresponding labels of ventricles, thalami, and hippocampi. The last column shows the RGB image of the registration field. Each row refers to an example registration case. The source images are well aligned to the target, demonstrating the good registration performance of our approach.

Folds	Elastix	NiftyReg	ANTs (SyN)	VM	VM-diff	Ours
img-to-atlas	91 (0.001)	6105 (0.088)	393 (0.005)	40674 (0.591)	**0**	**0**
img-to-img	642.2 (0.009)	10065 (0.146)	39699 (0.576)	52798 (0.767)	**0**	**0**

Table 2: The number of folds occurred in deformation fields with the percentages parenthesized under two experiment setting.

testing, respectively. Also, the unseen PPMI [Marek *et al.*, 2011] dataset containing 59 scans is employed for testing. Specifically, for the image-to-atlas registration, we use the atlas provided by [Balakrishnan *et al.*, 2019] as target, and for image-to-image case, the target images are randomly selected from datasets. Considering the large disparity among different datasets, standard pre-processing operations were conducted, including motion correction, NU intensity correction, normalization, skull stripping, and affine registration, as described in the research [Balakrishnan *et al.*, 2019]. To reduce memory usage, the images are cropped to $160 \times 192 \times 224$.

Evaluation Metrics. To achieve a more comprehensive evaluation, both the average Dice score [Dice, 1945] over registered testing pairs and the Jacobian matrix over the computed deformation are considered as evaluation metrics, to evaluate the anatomical overlap correspondences of the registered volume pairs and the smoothness of the deformation fields. To calculate the Dice score, segmentation is performed with FreeSurfer on each of the testing volumes to extract 30 anatomical structures, on which the average Dice is calculated. The Jacobian matrix $\mathbf{J}_\phi(\mathbf{x}) = \nabla\phi(\mathbf{x})$ captures the local properties of ϕ around voxel $\mathbf{x}$, such as stretching and rotating, negative determinants mean the loss of the one-to-one mapping [Ashburner, 2007]. So the local deformation is diffeomorphic at the locations where $|\mathbf{x} : \mathbf{J}_\phi(\mathbf{x}) > \mathbf{0}|$. We count all the folds, the negative Jacobian locations, and use Folds to represent the number of folds.

Baseline Methods. We compare our method with state-of-the-art registration techniques, including three optimization-based tools: Elastix [Klein *et al.*, 2010], Symmetric Normalization (SyN) [Avants *et al.*, 2008], NiftyReg [Sun *et al.*, 2014], and two learning-based methods: VoxelMorph [Balakrishnan *et al.*, 2019] and its diffeomorphic variant [Dalca *et al.*, 2019] (referred as VM and VM-diff, respectively). For the SyN algorithm, we use the version implemented in the ANTs [Avants *et al.*, 2011] package. The parameter settings of the conventional methods are as follows. For Elastix, we run B-spline registration with Mattes Mutual Information as 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 take Cross Correlation as the similarity measure metric and use the 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, ANTs (SyN), and NiftyReg on a PC with i7-8700 (@3.20GHz, 32G RAM), while learning-based methods on NVIDIA TITAN XP.

Implementation. We implement our networks with TensorFlow [Abadi *et al.*, 2016] package. It takes about 12 hours to train our model from scratch with 28100 iterations. During training, we use Adam optimizer [Kingma and Ba, 2015] with a learning rate of $1e^{-4}$. We set the batch size as 1. 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 to full-resolution.

5.1 Comparison Results

First, we quantitatively evaluate the accuracy, rationality and time consumption of all these techniques for both cases of image-to-atlas and image-to-image registration. Tab. 1 depicts the stability of the methods in terms of the Dice score on the different datasets, where higher values and lower variance

Runtime (s)	Elastix	NiftyReg	ANTs (SyN)	VM	VM-diff	Ours
img-to-atlas	90 (10)	486 (40)	4529 (1010)	0.615 (0.010)	0.512 (0.010)	**0.351 (0.007)**
img-to-img	67 (10)	323 (39)	4799 (1030)	0.697 (0.025)	0.586 (0.011)	**0.360 (0.006)**

Table 3: Comparison of time-consuming under two experiment settings.

Model		2-scales		3-scales		4-scales	
		w/o PFL	w/ PFL	w/o PFL	w/ PFL	w/o PFL	w/ PFL
Test	Dice	0.713 (0.034)	0.736 (0.028)	0.756 (0.020)	**0.773 (0.014)**	0.765 (0.017)	0.770 (0.016)
	LNCC	0.214 (0.006)	0.228 (0.005)	0.232 (0.005)	**0.244 (0.004)**	0.234 (0.005)	0.239 (0.004)
PPMI	Dice	0.740 (0.023)	0.758 (0.019)	0.773 (0.013)	0.785 (0.011)	0.779 (0.011)	**0.789 (0.011)**
	LNCC	0.213 (0.005)	0.225 (0.004)	0.229 (0.004)	**0.240 (0.004)**	0.230 (0.004)	0.235 (0.004)

Table 4: Ablation analysis of the lower-level task and the number of scales. Standard deviations are in bracket.

indicate a more accurate and stable registration. Our method gives an obvious lower variance with a comparable mean of Dice for both these two cases, showing stronger stability. As shown in Tab. 2, only VM-diff and our method can decrease the number of folds to zero, preserving the diffeomorphism. Besides, as Tab. 3 shows, our approach requires less running time, benefiting of the well-designed network architectures. Dealing with half-resolution rather than the original scale further accelerates the registration process.

To have a better understanding of the alignment results, we illustrate the Dice score of 30 anatomical structures in Fig. 2. Limited by space, besides our method, we only present ANTs (SyN), VM and VM-diff as the representatives for the optimization-base and learning-based techniques. We can see that compared with the conventional method ANTs (SyN), the deep methods VM, VM-diff give evenly accuracy but perform much less stable among different anatomical segmentations. While our deep model achieves a good balance between accuracy and stability. In summary, the robustness, accuracy, and speed are all the key performance indexes for the registration method. And extensive validation experiments indicate that our method performs favorably against state-of-the-art methods in practice.

Fig. 3 shows our representative registration results, from which we can see that our approach can ideally handle large changes in shapes. As shown, our method can ideally preserve the contour of anatomical structures, guaranteeing the topology of the registered volumes.

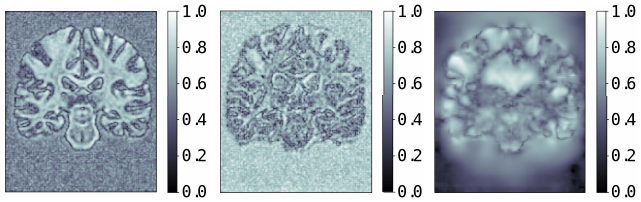

Figure 4: Example uncertainty of target feature, source feature and final registration field. Higher $H(.)$ value means lower uncertainty.

5.2 Ablation Studies

We investigate the roles of probabilistic feature learning, the lower-level optimization in our bi-level model, and the scales of learned features. Experiments were conducted on the testing data from multiple datasets the mixed datasets (referred to as Test) and the unseen PPMI dataset (referred to as PPMI). We substitute our probabilistic learning feature pyramids (PFL) with handcrafted image pyramids as the case of 'w/o PFL' and compare the performance gap between these two cases. Tab. 4 lists the registration accuracy in terms of both the Dice score and Local Normalized Correlation Coefficient (LNCC) at 2, 3 and 4 scales of features. We can see that our probabilistic feature learning evidently increases the accuracy in terms of both metrics. As the table shows, the features with three scales output the superior accuracy over the other scales. we observe in our experiments that more scales have more risks of overfitting when a limited number of training examples are available. Therefore, we adopt the probabilistic feature learning with three scales in all our experiments.

Fig. 4 shows the confidence maps of learned features for one slice of the target and source images, and the deformation fields in order to give a more intuitive illustration of the probabilistic features and uncertainty estimation, we show uncertainty maps. The features with higher confidences are coincident with prominent image structures. Therefore, the registration upon aligning these structures is able to produce stabler outputs than deterministic learning.

6 Conclusions

We introduce a bi-level optimization model to simultaneously address the deformable registration and feature optimization. It takes the energy for deformation computation as the upper-level optimization while formulates the maximum a posterior for features as the lower-level optimization. Such deformation learning policy with adaptive registration feature extraction is completely different from existing straightforward learning-based image registration methods. Then we employ efficient deep architectures to simultaneously propagate deformation fields and perform feature optimization. The losses on the feature space and image domain are utilized in our unsupervised end-to-end learning framework. We conduct two

groups of image registration experiments on 3D brain MRI datasets including image-to-atlas and image-to-image registrations. Extensive results show that our method achieves state-of-the-art performance with extreme efficiency. We have demonstrated our performance on the mon-modal registration task, and future validation remains on multi-modal registration.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Nos. 61922019, 61672125, 61733002 and 61772105), and LiaoNing Revitalization Talents Program (XLYC1807088). We thank Adrian V. Dalca for sharing code to simulate the image for the Boxplots indicating Dice scores for anatomical structures.

References

[Abadi *et al.*, 2016] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467, 2016.

[Ashburner, 2007] John Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95–113, 2007.

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

[Avants *et al.*, 2011] Brian B Avants, Nicholas J Tustison, Gang Song, Philip A Cook, Arno Klein, and James C Gee. A reproducible evaluation of ants similarity metric performance in brain image registration. *NeuroImage*, 54(3):2033–2044, 2011.

[Bajcsy and Kovacic, 1989] Ruzena Bajcsy and Stanislav Kovacic. Multiresolution elastic matching. *Computer Vision, Graphics, and Image Processing*, 46(1):1–21, 1989.

[Balakrishnan *et al.*, 2019] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019.

[Banerjee *et al.*, 2019] Sayan Banerjee, Avik Hati, Subhasis Chaudhuri, and Rajbabu Velmurugan. Cosegnet: Image co-segmentation using a conditional siamese convolutional network. In *IJCAI*, pages 673–679, 2019.

[Cheng *et al.*, 2018] Gong Cheng, Decheng Gao, Yang Liu, and Junwei Han. Multi-scale and discriminative part detectors based features for multi-label image classification. In *IJCAI*, pages 649–655, 2018.

[Colson *et al.*, 2005] Benoît Colson, Patrice Marcotte, and Gilles Savard. Bilevel programming: A survey. *A Quarterly Journal of Operations Research*, 3(2):87–107, 2005.

[Dalca *et al.*, 2019] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical Image Analysis*, 57:226–236, 2019.

[de Vos *et al.*, 2019] Bob D. de Vos, Floris F. Berendsen, Max A. Viergever, Hessam Sokooti, Marius Staring, and Ivana Isgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical Image Analysis*, 52:128–143, 2019.

[Di Martino *et al.*, 2014] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The Autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in Autism. *Molecular psychiatry*, 19(6):659–667, 2014.

[Dice, 1945] Lee Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[Dosovitskiy *et al.*, 2015] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision*, pages 2758–2766, 2015.

[Hu *et al.*, 2019] Xiaojun Hu, Miao Kang, Weilin Huang, Matthew R Scott, Roland Wiest, and Mauricio Reyes. Dual-stream pyramid registration network. In *MICCAI*, pages 382–390, 2019.

[Jaderberg *et al.*, 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NeurIPS*, pages 2017–2025, 2015.

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. In *ICLR*, 2015.

[Kingma and Welling, 2014] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[Klein *et al.*, 2010] Stefan Klein, Marius Staring, Keelin Murphy, Max A Viergever, and Josien PW Pluim. elastix: A toolbox for intensity-based medical image registration. *IEEE Transactions on Medical Imaging*, 29(1):196–205, 2010.

[Liu *et al.*, 2018a] Risheng Liu, Shichao Cheng, Yi He, Xin Fan, Zhouchen Lin, and Zhongxuan Luo. On the convergence of learning-based iterative methods for nonconvex inverse problems. *CoRR*, abs/1808.05331, 2018.

[Liu *et al.*, 2018b] Risheng Liu, Shichao Cheng, Xiaokun Liu, Long Ma, Xin Fan, and Zhongxuan Luo. A bridging framework for model optimization and deep propagation. In *NeurIPS*, pages 4323–4332, 2018.

[Liu *et al.*, 2019a] Risheng Liu, Long Ma, Yiyang Wang, and Lei Zhang. Learning converged propagations with deep prior ensemble for image enhancement. *IEEE Transactions on Image Processing*, 28(3):1528–1543, 2019.

[Liu *et al.*, 2019b] Risheng Liu, Yuxi Zhang, Shichao Cheng, Xin Fan, and Zhongxuan Luo. A theoretically guaranteed deep optimization framework for robust compressive sensing MRI. In *AAAI*, pages 4368–4375, 2019.

[Liu *et al.*, 2020] Risheng Liu, Zi Li, Yuxi Zhang, Chenying Zhao, Hao Huang, Zhongxuan Luo, and Xin Fan. A multi-scale optimization learning framework for diffeomorphic deformable registration. *CoRR*, abs/2004.14557, 2020.

[Maintz and Viergever, 1998] J. B. Antoine Maintz and Max A. Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1):1–36, 1998.

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

[Marek *et al.*, 2011] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative. *Progress in Neurobiology*, 95(4):629–635, 2011.

[Mueller *et al.*, 2005] Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford R Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's disease neuroimaging initiative. *Alzheimer's Dementia*, 1(1):55–66, 2005.

[Shen *et al.*, 2019] Zhengyang Shen, Xu Han, Zhenlin Xu, and Marc Niethammer. Networks for joint affine and non-parametric image registration. In *IEEE CVPR*, pages 4224–4233, 2019.

[Sun *et al.*, 2014] Wei Sun, Wiro J Niessen, and Stefan Klein. Free-form deformation using lower-order B-spline for nonrigid image registration. In *MICCAI*, pages 194–201, 2014.

[Sun *et al.*, 2018] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *IEEE CVPR*, pages 8934–8943, 2018.

[Yang *et al.*, 2017] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. Quicksilver: Fast predictive image registration - A deep learning approach. *NeuroImage*, 158:378–396, 2017.