IIRP-Net: Iterative Inference Residual Pyramid Network for Enhanced Image Registration

Tai Ma, Suwei Zhang, Jiafeng Li, Ying Wen*
School of Communication and Electronic Engineering, East China Normal University, Shanghai, China.
{51194506028,51255904033,51205904113}@stu.ecnu.edu.cn;ywen@cs.ecnu.edu.cn

Abstract

Deep learning-based image registration (DLIR) methods have achieved remarkable success in deformable image registration. We observe that iterative inference can exploit the well-trained registration network to the fullest extent. In this work, we propose a novel Iterative Inference Residual Pyramid Network (IIRP-Net) to enhance registration performance without any additional training costs. In IIRP-Net, we construct a streamlined pyramid registration network consisting of a feature extractor and residual flow estimators (RP-Net) to achieve generalized capabilities in feature extraction and registration. Then, in the inference phase, IIRP-Net employs an iterative inference strategy to enhance RP-Net by iteratively reutilizing residual flow estimators from coarse to fine. The number of iterations is adaptively determined by the proposed IterStop mechanism. We conduct extensive experiments on the FLARE and Mindboggle datasets and the results verify the effectiveness of the proposed method, outperforming state-of-the-art deformable image registration methods. Our code is available at https://github.com/Torbjorn1997/IIRP-Net.

1. Introduction

Deformable image registration is an important task in computer vision, focusing on establishing non-linear dense correspondences between two n-D images. It has widespread applications, particularly in medical image analysis [6, 29, 32] and remote sensing [9]. Traditional deformable registration methods usually formulate image registration as an optimization task and attempt to minimize the energy function in an iterative manner. Common intensity-based optimization methods [5, 33, 37] utilize the intensity differences between images as the energy function. However, in regions with weak texture, the gradient tends to be small, leading to optimization falling into local minima [20]. To address this issue, feature-based optimization methods are proposed, which utilize various feature representation functions such as graph spectral representation [21], structure tensor [35], and Gabor features [28, 36]. These methods extend the image intensity information into more comprehensive feature information, enhancing the capability to handle areas where intensity-based methods might struggle. Traditional optimization registration methods treat the registration task as an independent iterative optimization problem. When the target image pairs exhibit significant anatomical appearance variations, the registration time increases dramatically [25].

In recent years, some deep learning-based methods [3, 4, 13, 25] have been proposed for deformable image registration. These methods utilize Convolutional Neural Networks (CNNs) to directly estimate the displacement field for registering a pair of input images. Unlike the traditional optimization methods that optimize an independent registration function for each image pair, deep learning image registration (DLIR) methods optimize the parameters of a neural network to minimize the loss function across a dataset [3]. The DLIR methods can be divided into two phases: training and inference. In the training phase, the DLIR methods utilize a large number of image pairs from the dataset to learn a generalized registration network. In the inference phase, the trained registration network is used to predict the corresponding deformation field for a particular image pair. Current DLIR methods [3, 10, 13, 40] achieve fast and accurate registration during inference, but they show inferior performance in handling complicated large deformation problems and fine structure registration. One effective solution to deal with the limited alignment capability of a single registration network is to introduce a recursive cascading strategy [7, 26, 27, 31, 38]. The recursive cascading method utilizes multiple registration networks to break down the registration task into a series of recursive subtasks, which significantly enhances registration accuracy. Another solution employs the feature pyramid strategy [12, 14, 17, 24] to obtain multi-scale features and align feature pairs from coarse to fine, thus capturing inter-image correlations over various

*Corresponding author (e-mail: ywen@cs.ecnu.edu.cn) receptive fields. Although both registration strategies can effectively enhance registration accuracy, the potential of the networks for registration is not fully utilized.

In this paper, we propose an Iterative Inference Residual Pyramid Network (IIRP-Net) for unsupervised image registration which incorporates a novel iterative inference strategy. The inspiration behind the iterative inference strategy stems from an observation that once a registration network has been trained, it operates as a stationary function during the inference phase. Leveraging recursive iterations of this function has the potential to enhance registration accuracy. Specifically, we construct a streamlined pyramid registration network RP-Net, which uses a weight-sharing feature extractor to acquire multi-scale features and deploys several residual flow estimators to predict deformation fields from coarse to fine. During the training phase, RP-Net acquires a generalized registration pattern from an extensive dataset of image pairs, while in the inference phase, the trained residual flow estimators are iteratively applied from coarse to fine to fully leverage their registration capabilities. RP-Net enhanced with iterative inference strategy is referred to as IIRP-Net. In IIRP-Net, an IterStop mechanism is introduced to adaptively determine the number of iterative inferences for each residual flow estimator based on a specific image pair. This is similar to the optimizer in traditional optimization-based registration methods, which checks for stopping conditions. IIRP-Net efficiently combines the pyramid network structure and iterative inference strategy, providing a more conducive approach for deformable image registration.

In summary, the main contributions of this paper are as follows:

1. We construct RP-Net, a streamlined pyramid registration network, which enhances its flow estimators with ResBlocks to achieve high-accuracy registration.
2. We propose an iterative inference strategy that fully exploits the registration capability of the trained flow estimators during the inference phase. By adopting the proposed IterStop mechanism, IIRP-Net can dynamically determine the optimal number of iterations for each flow estimator, thus achieving superior alignment for specific image pairs.
3. The iterative inference strategy incurs no additional computational cost during the training phase with limited computational overhead during the inference phase. Furthermore, it has the potential to be extended to other pyramid-based registration networks.

To the best of our knowledge, IIRP-Net is the first pyramid registration network to be integrated with an iterative inference strategy. We validate and demonstrate the effectiveness of our method through the registration of the 3D abdomen CT dataset FLARE [22] and the brain MRI dataset Mindboggle [16], where IIRP-Net achieves state-of-the-art registration precision on both datasets.

2. Related Work

In recent years, DLIR methods gradually become the mainstream of registration tasks due to their advantages in registration speed and accuracy. According to different network structures, common DLIR methods can be categorized into three types: U-Net-based, pyramid-based, and cascade-based. The U-Net-based methods establish a straightforward and efficient framework for single-step registration. In response to the challenges of registering large deformations and fine structures, pyramid-based and cascade-based methods have been proposed. These methods employ multi-step recursive registration to effectively address complex registration scenarios.

2.1. U-Net-based Registration Method

VoxelMorph [3, 4] proposed by Dalca et al. is the most widely applied DLIR method, which obtains the corresponding deformation field by using a network similar to U-Net [30] to register a pair of images. Later, Dalca et al. [10] proposed the diffeomorphic VoxelMorph which employs the scaling and squaring method [1] to approximate the integration of the static velocity fields [2] and ensures the diffeomorphic properties of the deformation field. Chen et al. [8] introduced TransMorph which exploits the long-range dependency modeling capabilities of Swin Transformer [19] for more accurate registration results. Jia et al. [13] proposed LKU-Net which utilizes parallel large-kernel convolutions to significantly reduce parameters and capture long-range correlations. The above methods share a common U-Net architecture and directly estimate the deformation field. However, using these methods to predict a single deformation field tends to lead to local optima during the optimization process. U-Net-based registration methods often fail to achieve satisfactory registration accuracy, particularly when dealing with large deformations, due to the lack of a recursive updating mechanism.

2.2. Pyramid-based Registration Method

Pyramid-based registration methods leverage multi-scale information and use a single network to estimate and combine deformation fields from coarse to fine. Dual-PRNet++ [12, 14] achieves high-precision image registration by constructing feature pyramids for two images respectively and registering them layer by layer. NICE-Net [24] constructs a dual-stream encoder and introduces multi-resolution warped moving images in the prediction of the deformation field, integrating multi-stage registration tasks into one network. Im2grid [18] and ModeT [34] introduce matching scores based on cross-attention in the pyramid registration framework. PIViT [23] introduces recursive cascade Swin-Transformer-based decoders at the lowest scale of the pyramid structure, handling large deformations effectively. However, there is still room for improvement in the precision of pyramid-based registration methods.

2.3. Cascade-based Registration Method

Considering the limited registration capabilities of a single network, an effective improvement is to cascade multiple DLIR networks. Cascade-based registration method uses multiple networks to iteratively register image pairs, thereby decomposing the registration process into several recursive sub-processes. Zhao et al. [38] proposed a Recursive Cascade Network (RCN) to decompose a large deformation field into an affine matrix and several smaller deformable deformation fields. LapIRN [26] effectively solves large deformation registration by constructing an image Laplacian pyramid and learning the progressive deformation field from coarse to fine. SDH-Net [39] combines pyramid registration and cascade registration methods, achieving excellent registration effects but also incurring higher computational costs. These cascade-based registration methods yield fairly accurate registration results. However, the repetitive encoding-decoding process reduces computational efficiency, leading to increased resource consumption and extended processing time, especially during the training phase.

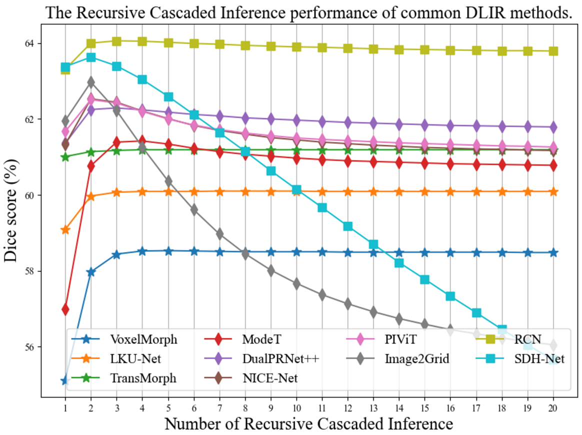

Figure 1. Line chart depicting the Dice scores achieved by 10 common DLIR methods on the Mindboggle dataset, following up to 20 inference iterations. ★, ♦, and ■ represent methods based on U-Net, pyramid, and cascade, respectively.

3. Method

Pyramid-based methods construct a high-performance non-iterative registration network, while cascade-based methods significantly enhance the performance of the base network at the cost of computational efficiency. Inspired by these methods, the IIRP-Net is proposed by introducing an iterative inference strategy into a pyramid-based network RP-Net enhanced with ResBlocks for high-accuracy registration.

3.1. Rethinking of Iterative Inference in DLIR

DLIR methods learn general image registration patterns during the training phase and use a generalized neural network in the inference phase to estimate the deformation field between specific image pairs. However, as a progressive task, image registration often yields suboptimal results by just using a generalized function for a single inference. Therefore, we can apply the deformation field to the moving image and use the resulting warped image along with the fixed image to predict a new deformation field through the same generalized registration function. Repeating this process multiple times constitutes an iterative inference process.

Figure 1 illustrates the results of applying the aforementioned prevalent DLIR methods with iterative inference on the Mindboggle dataset, where we perform recursive cascaded inference t times on the entire registration network. A single inference is conducted for $t = 1$, and iterative inferences are executed for $t = [2 \cdots 20]$. Evidently, iterative inference enhances registration accuracy, indicating that directly applying a generalized registration function or neural network to specific image pairs does not always yield optimal performance. The foundational performance of pyramid-based registration networks surpasses that of U-Net-based networks. As the number of iterations increases, the accuracy of U-Net-based methods continues to improve and eventually stabilizes. In contrast, the accuracy of pyramid-based methods initially increases but then gradually declines, especially in SDH-Net and Im2grid. The primary reason for the decline in accuracy may be that the recursive cascaded inference of the entire network overlooks the inherent multi-scale structure of pyramid networks, which operates from coarse to fine. Based on these observations, we explore how to effectively integrate iterative inference within a pyramid network. Subsequently, we introduce the construction of the pyramid network tailored for iterative inference and the design of the iterative inference strategy specifically devised for the pyramid network.

3.2. The Network Structure of RP-Net

We first construct a streamlined 4-level pyramid network RP-Net. To minimize potential disturbances during subsequent iterative inferences on RP-Net, it is designed to be as simple as possible, comprising only a feature extractor and four flow estimators. The overall structure of RP-Net is illustrated in Figure 2. RP-Net employs a weight-sharing feature extractor to construct multi-scale feature pyramids $\{F_f^i\}$ for the fixed image I_f and $\{F_m^i\}$ for the moving im-

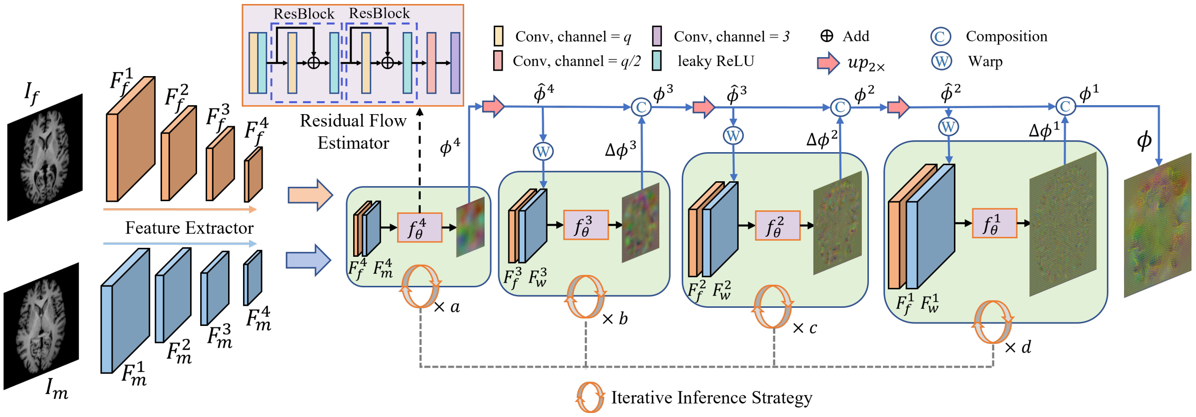

Figure 2. The overview of IIRP-Net. IIRP-Net integrates the iterative inference strategy within the pyramid network RP-Net during the inference phase. RP-Net consists of a feature extractor and four residual flow estimators f_θ^i. Within each f_θ^i, two consecutive ResBlocks are utilized. During the training phase, RP-Net employs each f_θ^i for a single prediction. In the inference phase, IIRP-Net iteratively infers f_θ^i from coarse to fine. The number of iterations (a, b, c, d) for each f_θ^i is adaptively determined by the IterStop mechanism.

age I_m, where i ranges from 1 to 4. The feature extractor consists of four 3D convolutions followed by leaky ReLU activations (with $\alpha = 0.1$). After each 3D convolution, average pooling is applied to downsample the feature maps by half. The output channels of the 3D convolutions are 8, 16, 32, and 32, respectively. Subsequently, $\{F_m^i\}$ and $\{F_f^i\}$ are processed through flow estimators in a pyramid-like manner.

The performance of the flow estimators f_θ^i in RP-Net which predict the deformation fields based on feature maps influences the effectiveness of iterative inference, where θ represents the parameters of f_θ^i. Consequently, we integrate two consecutive ResBlocks into the flow estimators to increase their network depth. Such flow estimators are referred to as residual flow estimators. Their structure is shown in the purple block of Figure 2, where the channel number q is set to 32 for $i = 4$, 3, and 2, and reduced to 16 for $i = 1$ to decrease computational costs. Incorporating ResBlocks to deepen the network expands the receptive field of f_θ^i, enabling it to capture long-range correlations effectively. Moreover, the deep network structure allows f_θ^i to fit complex functions. This not only enhances the network's adaptability to large deformation but also strengthens its ability to recognize fine structures.

At the coarsest scale, the concatenated features F_f^4 and F_m^4 are processed through f_θ^4 to generate the output flow field ϕ^4. This process can be formulated as:

$$\phi^i = f_\theta^i(F_f^i, F_m^i), i = 4. \tag{1}$$

At each subsequent scale i, the process initiates by upscaling the flow field ϕ^{i+1} to 2× resolution, producing $\hat{\phi}^{i+1}$. $\hat{\phi}^{i+1}$ is utilized to warp F_m^i, resulting in the warped moving image F_w^i. Subsequently, the concatenated features F_f^i and F_w^i are fed into the residual flow estimator to acquire the residual flow field $\Delta\phi^i$. The composition of $\hat{\phi}^{i+1}$ and $\Delta\phi^i$ constitutes the overall flow field ϕ^i. The process at subsequent scales can be formulated as:

$$\begin{cases} \hat{\phi}^{i+1} = up_{2\times}(\phi^{i+1}), \\ F_w^i = F_m^i \circ \hat{\phi}^{i+1}, \\ \Delta\phi^i = f_\theta^i(F_f^i, F_w^i), \\ \phi^i = \hat{\phi}^{i+1} \circ \Delta\phi^i, \end{cases} \quad i \in [3, 2, 1], \tag{2}$$

where $up_{2\times}$ denotes the operation of rescaling to 2× resolution and $\circ$ represents the composite transform. ϕ^1 is the final global deformation field ϕ.

3.3. Iterative Inference Strategy

After finalizing the training phase of RP-Net, each residual flow estimator f_θ^i possesses the generalized capability to predict $\Delta\phi^i$ based on the feature maps F_f^i and F_w^i. We propose a tailored iterative inference strategy for pyramid networks to further exploit the registration capabilities of f_θ^i. Unlike the image-level external iterative inference of the entire network performed in Figure 1, the iterative inference strategy iteratively applies each f_θ^i from coarse to fine, achieving an internal iteration at the feature map level. This design is consistent with the pyramid network's operation of warping feature maps and maintains the network's coarse to fine nature. RP-Net enhanced with the iterative inference strategy is referred to as IIRP-Net. When aligning a specific pair of images, we need to consider how many times each f_θ^i should be iteratively inferred. Therefore, mimicking the iterative process in optimization-based registration methods, we introduce an IterStop mechanism.

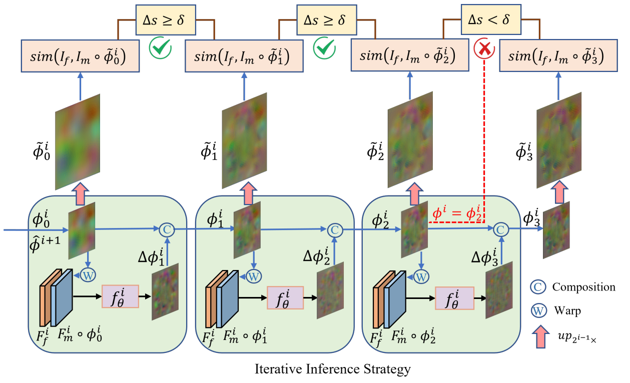

Figure 3. The process of applying the iterative inference strategy to f_θ^i in the inference phase of IIRP-Net. For clarity, we illustrate the process with 2 iterations as an example. During the 3rd iteration, $\Delta s < \delta$ meets the stopping condition of the iteration. Then, the iterative inference of f_θ^i is halted. ϕ_2^i serves as the output deformation field at scale i, denoted as ϕ^i.

Figure 3 illustrates how IIRP-Net incorporates the IterStop mechanism during the inference phase. At the i_{th} scale, IIRP-Net iteratively applies the residual flow estimator f_θ^i for multiple inferences, where ϕ_{t-1}^i and ϕ_t^i respectively represent the overall deformation fields after the $(t-1)_{th}$ and t_{th} inferences. The IterStop mechanism is a stopping criterion in the recursive iteration process of f_θ^i. It evaluates the necessity of further iterations by examining the difference in similarity between successive registration outputs. Standard similarity metrics $sim(,)$, including Mean Squared Error and Normalized Cross-Correlation, can be used to calculate the similarity of registration outputs. In the process of evaluating difference, we rescale ϕ_{t-1}^i and ϕ_t^i to $2^{i-1}\times$ resolution to obtain $\widetilde{\phi}_{t-1}^i$ and $\widetilde{\phi}_t^i$ to match the original images I_f and I_m. The iterative process is terminated when the difference Δs between $sim(I_m \circ \widetilde{\phi}_{t-1}^i, I_f)$ and $sim(I_m \circ \widetilde{\phi}_t^i, I_f)$ falls below a given threshold δ. Algorithm 1 provides a pseudo-code for this procedure. When iteration stops, the deformation field ϕ^i is rescaled to $2\times$ resolution to obtain $\hat{\phi}^i$, serving as the initial flow for the iterative inference of f_θ^{i-1}.

The IterStop mechanism introduced in IIRP-Net enables it to adaptively determine the number of iterations for each f_θ^i based on the registration results during the inference phase. This process can be seen as a combination of optimization-based registration methods and deep learning. During the training phase, the feature extractor and residual flow estimators are trained by using a large dataset. In the inference phase, the fixed residual flow estimators serve as the driving force for the iterative update of the deformation field. We employ a predefined similarity metric as an optimizer to determine when to halt the iterations, thereby replacing the time-consuming calculations of driving forces in optimization-based registration methods with fast neural network inference. Since each iteration utilizes only one f_θ^i, IIRP-Net maintains a rapid inference speed. The iterative inference strategy introduces only a limited additional computational cost during the inference phase without any extra cost in the training phase.

Algorithm 1 Iterative Inference Strategy

```
function IIS($f_\theta^i$, $F_f^i$, $F_m^i$, $\hat{\phi}^{i+1}$, $I_f$, $I_m$)
    $\phi_{t-1}^i \leftarrow \hat{\phi}^{i+1}, t \leftarrow 1$                ▷ get initial flow
    while True do
        $F_w^i \leftarrow F_m^i \circ \phi_{t-1}^i$
        $\Delta\phi_t^i \leftarrow f_\theta^i(F_f^i, F_w^i)$                ▷ iterative inference
        $\phi_t^i \leftarrow \phi_{t-1}^i \circ \Delta\phi_t^i$                ▷ flow update
        $\widetilde{\phi}_{t-1}^i \leftarrow up_{2^{i-1}\times}(\phi_{t-1}^i)$
        $\widetilde{\phi}_t^i \leftarrow up_{2^{i-1}\times}(\phi_t^i)$
        $\Delta s \leftarrow sim(I_m \circ \widetilde{\phi}_t^i, I_f) - sim(I_m \circ \widetilde{\phi}_{t-1}^i, I_f)$
        if $\Delta s < \delta$ then
            return $\phi^i \leftarrow \phi_{t-1}^i$                ▷ stop iteration
        end if
        $t \leftarrow t + 1$
    end while
end function
```

3.4. Loss Function

IIRP-Net is an unsupervised registration network. In this section, we present the loss function for training RP-Net. RP-Net uses NCC as the similarity measure. The image similarity loss function $\mathcal{L}_{ncc}$ can be defined as:

$$\mathcal{L}_{ncc}(I_f, I_m \circ \phi) = - \sum_{p \in \Omega} \frac{\sum_{p_i}(I_f(p_i) - \overline{I_f}(p))(I_m \circ \phi(p_i) - \overline{I_m \circ \phi}(p))}{\sqrt{\sum_{p_i}(I_f(p_i) - \overline{I_f}(p))^2 \sum_{p_i}(I_m \circ \phi(p_i) - \overline{I_m \circ \phi}(p))^2}}, \tag{3}$$

where $\overline{I_f}(p)$ and $\overline{I_m \circ \phi}(p)$ represent local mean intensity images, p_i represents the position within w^3 local window centered at p. The local mean is computed over a local w^3 window and w is set to 9 in the experiment.

Considering that a non-smooth flow field ϕ can generate discontinuities, we use a diffusion regularization on the spatial gradient of ϕ to ensure its smoothness:

$$\mathcal{L}_{smooth}(\phi) = \sum_{p \in \Omega} \|\nabla \phi(p)\|^2. \tag{4}$$

The complete loss function $\mathcal{L}_{RP}(I_f, I_m)$ of the proposed RP-Net can be formulated as:

$$\mathcal{L}_{RP}(I_f, I_m) = \mathcal{L}_{ncc}(I_f, I_m \circ \phi) + \lambda \mathcal{L}_{smooth}(\phi), \tag{5}$$

where λ is the hyperparameters used to balance the contribution of loss functions.

4. Experiment

To validate the performance of the proposed RP-Net and IIRP-Net, we conduct experiments on two 3D medical image datasets. In these experiments, we compare the proposed method with widely used 3D registration methods. These include U-Net-based methods such as VoxelMorph [3], LKU-Net [13], and TransMorph [8]; pyramid-based methods like ModeT [34], DualPRNet++ [14], NICE-Net [24], PIViT [23], and Image2Grid [18]; as well as cascade-based methods RCN [38] and SDH-Net [39]. RCN recursively cascades 3 VoxelMorphs, while SDH-Net performs 6 iterations. The loss functions for all methods are consistent with those used in their respective papers. We use Dice score [11], Jacobian determinant ($|J_s|_{\leq 0}$), 95% maximum Hausdorff distance (HD95), MSE, training time per step (t_{train}), inference time (t_{infer}) and network parameters (Params) as evaluation metrics in these experiments.

We implement the models using Pytorch backend and the ADAM[15] optimizer with a learning rate of 10^{-4}. The models are trained on an NVIDIA GeForce RTX 3090 GPU and an Intel(R) Xeon(R) Silver 4210R CPU with a batch size of 1. The training phase is conducted for 100,000 steps. In the experiment, the images are normalized to the range of [0,1]. The hyperparameter λ of $\mathcal{L}_{RP}$ in (5) is set to 1.

4.1. Datasets

We conduct experiments on two types of 3D medical images: abdomen CT scans and brain MRI scans. Before conducting the experiments, we perform preprocessing on these medical images. We crop out the unlabeled areas in the abdomen CT scans and the background areas in the brain MRI scans. The 3D volume for each abdomen scan is 128 × 128 × 96, with an isotropic voxel size of 2.5 mm. The 3D volume for each brain scan is 160 × 192 × 160, with an isotropic voxel size of 1 mm.

For abdomen CT scans, we use the dataset from the MICCAI 2021 FLARE [22] (Fast and Low GPU memory Abdominal oRgan sEgmentation) Challenge for training, validation, and testing. The FLARE dataset consists of 361 scans, from which we select 301 for training, 20 for validation, and 40 for testing. The scans in the FLARE dataset include four different organ labels: liver, kidney, spleen, and pancreas. The main challenge in abdomen CT registration tasks stems from the substantial spatial distribution differences in organs between different scans, which results in difficulties in handling large deformations during the registration process.

In the task of brain image registration, we choose the Mindboggle-101 [16] dataset for training, testing, and validation. Specifically, we select the NKI-RS-22 and NKI-TRT-20 datasets from Mindboggle-101 as the training set, OASIS-TRT-20 as the validation set, and the MMRR-21 subset as the testing set. All images are pre-aligned to the MNI152 template space. Compared to abdomen scans, the challenge in brain image registration lies in aligning dense, fine structures.

4.2. Comparison with Baseline Methods

To validate the performance of the proposed registration network, we quantitatively compare the proposed RP-Net and IIRP-Net with ten popular DLIR methods, as shown in Table 1.

We first analyze the Dice score and the number of voxels with non-positive Jacobian determinants. Compared to U-Net-based methods such as VoxelMorph and TransMorph, IIRP-Net outperforms them on the FLARE dataset with Dice score improvements of 32.5% and 22.7%, respectively. On the Mindboggle dataset, IIRP-Net's Dice scores are higher than VoxelMorph and TransMorph by 10.8% and 4.6%, respectively. Compared to the optimal pyramid-based method Image2Grid, RP-Net, which also employs a pyramid structure, achieves Dice score improvements of 0.8% and 2.7% on the FLARE and Mindboggle datasets, respectively. Furthermore, with the introduction of the iterative inference strategy, the improvement increases to 5.4% and 3.8%. RP-Net surpasses other comparison methods on the Mindboggle dataset and is only slightly behind SDH-Net on the FLARE dataset, demonstrating the effectiveness of the residual flow estimator. IIRP-Net outperforms the state-of-the-art results by 1.1% and 2.2% on each dataset, respectively. Additionally, the iterative inference strategy enhances the diffeomorphic properties of registration. On the Mindboggle dataset, the number of folding points in IIRP-Net is reduced to 20% of those in RP-Net. RP-Net also performs well on the other two similarity metrics, HD95 and MSE. Building upon this, IIRP-Net achieves further enhancements in these aspects.

Another advantage of RP-Net is its computational efficiency. Compared with NICE-Net and PIViT based on pyramid structures, RP-Net has similar time costs for training and inference, yet it stands out with the least parameters while delivering the highest accuracy. IIRP-Net does not increase training time or parameters, only incurring a limited additional inference time. Compared to the state-of-the-art recursive cascaded pyramid network method SDH-Net, the time advantage of IIRP-Net is particularly evident. On the two datasets, the training time of IIRP-Net is only 31.9% and 31.3% of that of SDH-Net, respectively, while the inference time is just 19.4% and 14.6% of SDH-Net's.

Figure 4 illustrates the registration results of RP-Net and IIRP-Net compared to other methods on the Mindboggle dataset. The differences in registration primarily focus on the detailed structural areas of the brain images, with red

Table 1. Comparison among different registration methods on the FLARE and Mindboggle datasets.

Method	FLARE						Mindboggle						Params
	Dice(%) ↑	$\|J_s\|_{\leq 0}$ ↓	**HD95(mm)** ↓	**MSE**(10^{-3}) ↓	t_{train} ↓	t_{infer} ↓	**Dice**(%) ↑	$\|J_s\|_{\leq 0}$ ↓	**HD95(mm)** ↓	**MSE**(10^{-3}) ↓	t_{train} ↓	t_{infer} ↓	
VoxelMorph [3]	40.8±14.2	<4.5%	38.271	15.214	**0.20**s	**0.02**s	55.0±3.7	<0.7%	5.802	4.441	**0.60**s	0.09s	**319.4K**
LKU-Net [13]	42.7±14.3	<4.4%	37.089	13.995	0.28s	**0.02**s	59.3±3.0	<1.2%	5.565	3.015	0.81s	0.08s	2038.7K
TransMorph [8]	50.6±15.1	<3.2%	34.121	12.894	0.44s	0.09s	61.2±3.1	<1.0%	5.457	3.141	1.48s	0.15s	45650.3K
ModeT [34]	60.5±11.7	<**0.02**%	26.841	12.429	1.78s	0.32s	57.1±2.1	<**0.0004**%	5.286	5.197	5.41s	1.08s	1005.6K
DualPRNet++ [14]	58.8±13.8	<2.6%	30.929	10.048	0.66s	0.13s	61.5±3.0	<0.5%	5.449	2.763	2.05s	0.46s	1208.0K
NICE-Net [24]	64.1±13.1	<2.9%	26.270	5.347	0.23s	0.03s	61.4±1.5	<0.9%	5.177	2.430	0.67s	0.10s	1068.2K
PIViT [23]	67.3±12.8	<0.4%	24.045	8.989	0.23s	**0.02**s	61.8±1.2	<0.1%	4.938	3.168	0.62s	0.06s	649.3K
Image2Grid [18]	67.9±10.8	<0.2%	24.179	9.060	0.37s	0.11s	62.0±1.4	<0.03%	5.031	3.374	1.05s	0.66s	865.2K
RCN [38]	64.9±13.2	<2.2%	25.984	5.262	0.32s	0.06s	63.4±1.4	<0.8%	5.096	2.128	1.02s	0.19s	958.1K
SDH-Net [39]	72.2±12.8	<0.5%	**22.857**	6.784	0.69s	0.31s	63.6±1.7	<0.2%	5.048	2.421	2.01s	0.82s	17862.0K
RP-Net	68.7±12.8	<1.9%	24.018	5.124	0.22s	**0.02**s	64.7±1.2	<0.4%	4.889	1.910	0.63s	**0.05**s	410.1K
IIRP-Net	**73.3±11.5**	<1.5%	23.072	**3.591**	0.22s	0.06s	**65.8±1.2**	<0.08%	**4.840**	**1.436**	0.63s	0.12s	410.1K

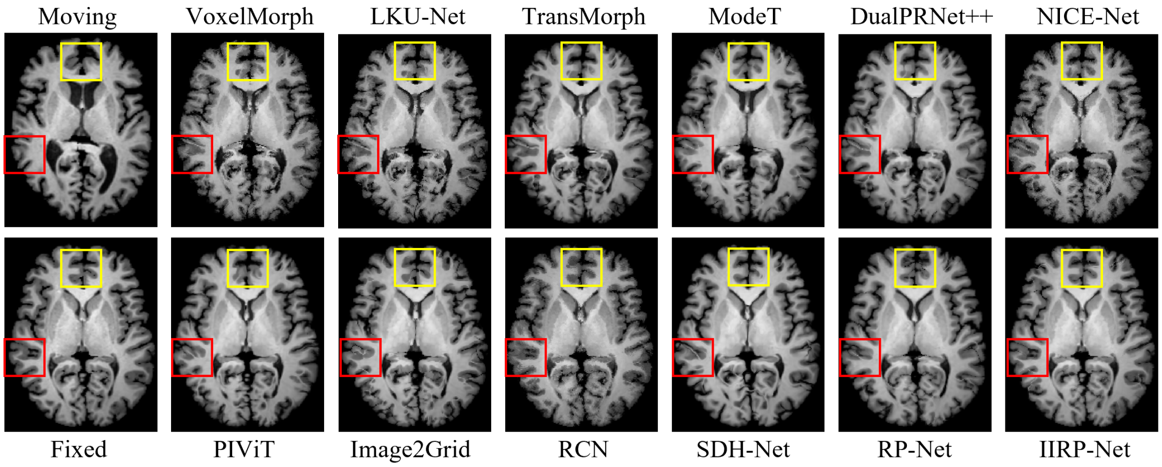

Figure 4. Example slices from the fixed images, moving images, and warped images by VoxelMorph, LKU-Net, TransMorph, ModeT, DualPRNet++, NICE-Net, PIViT, Image2Grid, RCN, SDH-Net, RP-Net and IIRP-Net on Mindboggle datasets. Red and yellow boxes highlight regions where IIRP-Net evidently outperforms other methods.

and yellow boxes highlighting two regions where the differences are particularly noticeable. In these areas, IIRP-Net achieves results that are nearly identical to the fixed image, demonstrating its superior performance in aligning fine structures.

4.3. Ablation Study on Residual Flow Estimator

Table 2. The effects of applying different numbers of ResBlocks in the residual flow estimators of RP-Net and IIRP-Net. The values in parentheses indicate the degree of improvement in the Dice score of IIRP-Net compared to RP-Net.

	FLARE			Mindboggle		
	RP Dice (%)	**IIRP Dice** (%)	t_{train}	**RP Dice** (%)	**IIRP Dice** (%)	t_{train}
Res×0	65.72	70.01 (+4.29)	**0.20s**	63.24	64.75 (+1.51)	**0.58s**
Res×1	67.58	71.61 (+4.03)	0.21s	64.05	65.33 (+1.28)	0.61s
Res×2	68.68	73.34 (+4.66)	0.22s	64.67	65.84 (+1.17)	0.63s
Res×3	**69.14**	**73.98 (+4.84)**	0.25s	64.89	66.14 (+1.25)	0.79s
Res×4	69.04	73.45 (+4.41)	0.27s	**64.97**	**66.38 (+1.41)**	0.88s

As the main improvement in RP-Net is the residual flow estimator, we analyze how the introduction of continuous ResBlocks impacts the performance of both RP-Net and IIRP-Net. Table 2 displays the registration results of RP-Net using residual flow estimators with different numbers of ResBlocks on the FLARE and Mindboggle datasets. Res×n indicates the use of n ResBlocks, and we provide the Dice scores for RP-Net and IIRP-Net under the corresponding settings, along with their training times. According to the results in Table 2, introducing ResBlocks improves registration accuracy, with the best performance on FLARE achieved at $n = 3$ and on Mindboggle at $n = 4$. The accuracy results of IIRP-Net demonstrate that the iterative inference strategy enhances registration across various network structures. An RP-Net with higher registration accuracy also achieves better results when the iterative inference strategy is introduced. However, increasing the number of ResBlocks also leads to higher computational costs. To balance accuracy with computational expense, we set n to 2 in the implementation of IIRP-Net.

4.4. Ablation Study on Iterative Inference

As the core mechanism of IIRP-Net, the iterative inference strategy plays an important role in enhancing registration. In this section, we analyze the effects of iterative inference from two perspectives: without and with the IterStop mechanism.

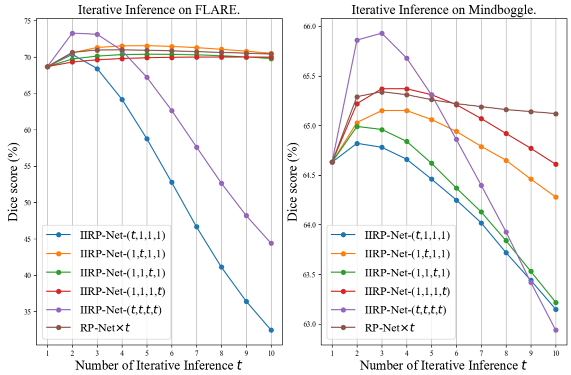

Figure 5. The line chart depicts the Dice scores achieved by RP-Net and IIRP-Net with different iteration ways on the FLARE and Mindboggle datasets, with inference iterations going up to 10 times.

Iterative inference without IterStop. Without the IterStop mechanism, it is impossible to determine the exact number of iterations needed for each f_θ^i. Hence, we adopt five different ways: iterating a particular f_θ^i t times and iterating all f_θ^i t times, where t is a predetermined fixed number of iterations. These methods are denoted as IIRP-Net-(a, b, c, d), where a, b, c, d represent the iteration counts for each f_θ^i from coarse to fine. Additionally, we compare the practice of externally cascading RP-Net for t iterations, denoted as RP-Net$\times t$, which is the same as the recursive cascaded inference conducted in Figure 1. The experimental results on the FLARE and Mindboggle datasets are presented in Figure 5, with the iteration limit set to 10 times in the experiments.

Comparing the peak values of the line graphs for IIRP-Net-(t, t, t, t) and RP-Net$\times t$, it's clear that IIRP-Net-(t, t, t, t) achieves higher registration accuracy. When t increases from 1 to 2 or 3, iterative methods demonstrate an improvement in accuracy, with IIRP-Net-(t, t, t, t) showing the greatest enhancement. However, as t further increases, the registration accuracy of IIRP-Net-(t, t, t, t) rapidly declines. This is primarily due to the iterative inference strategy continuously warping the feature maps. Without sufficient supervisory information, excessive iterations can lead to misalignment, especially at coarser scales. Therefore, introducing a conditional mechanism to terminate the iteration is important.

Table 3. The impact of selecting different thresholds (δ) in the IterStop mechanism on the iterative inference strategy. The numbers in parentheses indicate the differences compared to RP-Net.

Method		FLARE		Mindboggle	
		Dice (%)	(a, b, c, d)	**Dice** (%)	(a, b, c, d)
RP-Net		68.68 (-)	(1.0, 1.0, 1.0, 1.0)	64.67 (-)	(1.0, 1.0, 1.0, 1.0)
IIRP-Net	δ=0.01	73.05 (+4.37)	(1.2, 2.9, 2.5, 2.0)	65.61 (+0.94)	(1.1, 3.0, 2.0, 1.4)
	δ=0.005	73.34 (+4.66)	(1.3, 3.4, 3.0, 2.6)	**65.84 (+1.17)**)	(1.2, 3.5, 2.0, 2.0)
	δ=0.001	**73.47** (**+4.79**)	(1.4, 4.2, 4.0, 4.6)	65.74 (+1.07)	(2.1, 5.7, 3.0, 3.0)
	δ=0.0005	73.44 (+4.76)	(1.4, 4.4, 4.4, 5.7)	65.58 (+0.91)	(2.3, 6.3, 4.0, 3.2)

Iterative inference with IterStop. After introducing the IterStop mechanism, IIRP-Net can adaptively determine the iteration counts a, b, c, d for each f_θ^i to specific image pair. IIRP-Net employs NCC as the similarity metric for the IterStop mechanism. Table 3 shows the Dice scores of IIRP-Net on the FLARE and Mindboggle datasets when using different thresholds δ. Since each image pair in the dataset undergoes a different number of iterations, (a, b, c, d) represents the average number of iterations for each f_θ^i. Based on the results in Table 3, we select $\delta = 0.005$ as the threshold for IIRP-Net. As an auxiliary automatic decision-making tool, the IterStop mechanism enables IIRP-Net to achieve robust iterative inference results, avoiding the rapid decline in accuracy shown in Figure 5.

5. Conclusion

We present an Iterative Inference Residual Pyramid Network (IIRP-Net) for deformable image registration. Building upon the high-performance RP-Net, IIRP-Net enhances registration through iterative inference. The iterative inference strategy equipped with the IterStop mechanism in IIRP-Net further exploits the registration capabilities of the residual flow estimators with limited additional inference consumption. Experimental results demonstrate that IIRP-Net combines the advantages of multi-scale processing and iterative inference, achieving state-of-the-art registration accuracy on both FLARE and Mindboggle datasets while maintaining computational efficiency. The RP-Net, designed with simplicity as its foundation, facilitates the extension of the iterative inference strategy to various other pyramid-based registration networks.

Limitations. The IterStop mechanism in the iterative inference strategy, which uses intensity differences (such as NCC) to determine the need for further iteration, still has room for improvement. This is because intensity information may not always correspond completely with anatomical structures, potentially limiting the accuracy of IterStop's decisions.

Acknowledgments. This work was supported in part by National Nature Science Foundation of China (62273150, 62171323), Shanghai Natural Science Foundation (22ZR1421000), Shanghai Outstanding Academic Leaders Plan (21XD1430600), Science and Technology Commission of Shanghai Municipality (22DZ2229004).

References

[1] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 924–931. Springer, 2006. 2

[2] John Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007. 2

[3] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 6, 7

[4] 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. 1, 2

[5] M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61(2):139–157, 2005. 1

[6] Naxin Cai, Houjin Chen, Yanfeng Li, Yahui Peng, and Jiaxin Li. Adaptive weighting landmark-based group-wise registration on lung dce-mri images. *IEEE Transactions on Medical Imaging*, 40(2):673–687, 2021. 1

[7] Tongtong Che, Xiuying Wang, Kun Zhao, Yan Zhao, Debin Zeng, Qiongling Li, Yuanjie Zheng, Ning Yang, Jian Wang, and Shuyu Li. Amnet: Adaptive multi-level network for deformable registration of 3d brain mr images. *Medical Image Analysis*, 85:102740, 2023. 1

[8] Junyu Chen, Eric C Frey, Yufan He, William P Segars, Ye Li, and Yong Du. Transmorph: Transformer for unsupervised medical image registration. *Medical image analysis*, 82:102615, 2022. 2, 6, 7

[9] Shuhan Chen, Shengwei Zhong, Bai Xue, Xiaorun Li, Liaoying Zhao, and Chein-I Chang. Iterative scale-invariant feature transform for remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing*, 59(4): 3244–3265, 2021. 1

[10] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 729–738. Springer, 2018. 1, 2

[11] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 6

[12] Xiaojun Hu, Miao Kang, Weilin Huang, Matthew R Scott, Roland Wiest, and Mauricio Reyes. Dual-stream pyramid registration network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 382–390. Springer, 2019. 1, 2

[13] Xi Jia, Joseph Bartlett, Tianyang Zhang, Wenqi Lu, Zhaowen Qiu, and Jinming Duan. U-net vs transformer: Is u-net outdated in medical image registration? In *International Workshop on Machine Learning in Medical Imaging*, pages 151–160. Springer, 2022. 1, 2, 6, 7

[14] Miao Kang, Xiaojun Hu, Weilin Huang, Matthew R Scott, and Mauricio Reyes. Dual-stream pyramid registration network. *Medical Image Analysis*, 78:102379, 2022. 1, 2, 6, 7

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[16] Arno Klein and Jason Tourville. 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in neuroscience*, 6:171, 2012. 2, 6

[17] Risheng Liu, Zi Li, Xin Fan, Chenying Zhao, Hao Huang, and Zhongxuan Luo. Learning deformable image registration from optimization: perspective, modules, bilevel training and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7688–7704, 2021. 1

[18] Yihao Liu, Lianrui Zuo, Shuo Han, Yuan Xue, Jerry L Prince, and Aaron Carass. Coordinate translator for learning deformable medical image registration. In *International Workshop on Multiscale Multimodal Medical Imaging*, pages 98–109. Springer, 2022. 2, 6, 7

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[20] Herve Lombaert, Leo Grady, Xavier Pennec, Nicholas Ayache, and Farida Cheriet. Spectral demons–image registration via global spectral correspondence. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II 12*, pages 30–44. Springer, 2012. 1

[21] Herve Lombaert, Leo Grady, Xavier Pennec, Nicholas Ayache, and Farida Cheriet. Spectral log-demons: diffeomorphic image registration with very large deformations. *International journal of computer vision*, 107:254–271, 2014. 1

[22] Jun Ma, Yao Zhang, Song Gu, Xingle An, Zhihe Wang, Cheng Ge, Congcong Wang, Fan Zhang, Yu Wang, Yinan Xu, et al. Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. *Medical Image Analysis*, 82: 102616, 2022. 2, 6

[23] Tai Ma, Xinru Dai, Suwei Zhang, and Ying Wen. Pivit: Large deformation image registration with pyramid-iterative vision transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 602–612. Springer, 2023. 2, 6, 7

[24] Mingyuan Meng, Lei Bi, Dagan Feng, and Jinman Kim. Non-iterative coarse-to-fine registration based on single-pass deep cumulative learning. *arXiv preprint arXiv:2206.12596*, 2022. 1, 2, 6, 7

[25] Tony CW Mok and Albert Chung. Fast symmetric diffeomorphic image registration with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4644–4653, 2020. 1

[26] Tony CW Mok and Albert CS Chung. Large deformation diffeomorphic image registration with laplacian pyramid networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 211–221. Springer, 2020. 1, 3

[27] Tony CW Mok and Albert CS Chung. Conditional deformable image registration with convolutional neural network. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*, pages 35–45. Springer, 2021. 1

[28] Yangming Ou, Aristeidis Sotiras, Nikos Paragios, and Christos Davatzikos. Dramms: Deformable registration via attribute matching and mutual-saliency weighting. *Medical image analysis*, 15(4):622–639, 2011. 1

[29] Shuvendu Rana, Rory Hampson, and Gordon Dobie. Breast cancer: Model reconstruction and image registration from segmented deformed image using visual and force based analysis. *IEEE Transactions on Medical Imaging*, 39(5): 1295–1305, 2020. 1

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[31] Yucheng Shu, Hao Wang, Bin Xiao, Xiuli Bi, and Weisheng Li. Medical image registration based on uncoupled learning and accumulative enhancement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–13. Springer, 2021. 1

[32] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. Deformable medical image registration: A survey. *IEEE Transactions on Medical Imaging*, 32(7):1153–1190, 2013. 1

[33] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Symmetric log-domain diffeomorphic registration: A demons-based approach. In *International conference on medical image computing and computer-assisted intervention*, pages 754–761. Springer, 2008. 1

[34] Haiqiao Wang, Dong Ni, and Yi Wang. Modet: Learning deformable image registration via motion decomposition transformer. *arXiv preprint arXiv:2306.05688*, 2023. 2, 6, 7

[35] Ying Wen, Le Zhang, LiangHua He, and MengChu Zhou. Incorporation of structural tensor and driving force into log-demons for large-deformation image registration. *IEEE Transactions on Image Processing*, 28(12):6091–6102, 2019. 1

[36] Ying Wen, Cheng Xu, Yue Lu, Qingli Li, Haibin Cai, and Lianghua He. Gabor feature-based logdemons with inertial constraint for nonrigid image registration. *IEEE Transactions on Image Processing*, 29:8238–8250, 2020. 1

[37] Miaomiao Zhang and P Thomas Fletcher. Fast diffeomorphic image registration via fourier-approximated lie algebras. *International Journal of Computer Vision*, 127:61–73, 2019. 1

[38] Shengyu Zhao, Yue Dong, Eric I Chang, Yan Xu, et al. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10600–10610, 2019. 1, 3, 6, 7

[39] Shenglong Zhou, Bo Hu, Zhiwei Xiong, and Feng Wu. Self-distilled hierarchical network for unsupervised deformable image registration. *IEEE Transactions on Medical Imaging*, 2023. 3, 6, 7

[40] Yongpei Zhu and Shi Lu. Swin-voxelmorph: A symmetric unsupervised learning model for deformable medical image registration using swin transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 78–87. Springer, 2022. 1