Self-distilled Evolutionary Network for Deformable Medical Image Registration via Pyramid and Cascaded Progressive Learning

Yibo Wang[a,1], Mingwei Wen[a,1], Xuming Zhang[a,*]

[a]*Department of Biomedical Engineering, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, 430074, Hubei, China*

Abstract

Deformable image registration serves as an important prerequisite for multi-modal image fusion, disease diagnosis, and surgical navigation. Although existing learning-based registration methods are very popular, they mainly rely on the pyramid and cascaded structure based networks trained in an end-to-end manner, leading to their insufficient parameter optimization and unsatisfactory accuracy. In this paper, we have proposed a novel self-distilled evolutionary network (SEN) based on the idea of knowledge distillation. In the proposed SEN, the optimization order of each component is reasonably arranged based on the idea of evolution and the corresponding learning gaps are filled by using the self-distillation scheme. The layers of different levels in SEN are progressively updated based on the depth-wise pyramid weight-inherited evolution, thereby enhancing the sensitivity of SEN to multi-scale deformations. Meanwhile, the feature maps of trained layers are treated as teachers and shifted to distill the learning process of new generated layers. Besides, more sub-networks in SEN are progressively appended based on the broad-wise cascaded evolution, thereby making it easier for SEN to learn the decomposed deformations. Distinctively, a distribution-consistent distillation is proposed to fill the learning gaps between the new appended sub-networks and old ones. Due to the introduction of two types of evolution and distillation, it is easier for SEN to handle complex deformations. Experiments on both mono-modal datasets (LPBA40, OASIS, SLIVER, LSPIG) and multi-modal dataset (MMWHS2017) demonstrate the superior performance of SEN to several baseline methods with higher accuracy in terms of the Dice score and faster inference speed. Our code will be released at `https://github.com/YiboWang3813/SEN`.

Keywords: Deformable image registration, Self-distilled evolutionary network, Knowledge distillation, Progressive learning

1. Introduction

Deformable image registration is an essential task in the field of medical image analysis, and it aims to determine the transformation (known as deformation field) to warp the moving image so that its pixels or voxels can be aligned to those of the fixed image [1]. Applications of deformable image registration include matching multi-modal images from different imaging devices for the later fusion [2], aligning scans of the same patient over a long temporal span for monitoring disease development [3], and assisting in the image-guided surgical navigation during interventions [4].

The traditional methods treat the registration task as an optimization problem and estimate the deformation field by iteratively warping the moving image based on an energy function [5, 6, 7, 8, 9, 10, 11]. Obviously, the involved iterative optimization is computation-intensive and time-consuming, thereby prohibiting these methods from being utilized in the practical scenarios where the real-time registration is required.

Recently, with the development of deep learning, learning-based methods have been widely utilized in the field of deformable image registration due to their high efficiency. According to whether the ground-truth deformation field is needed, learning-based methods can be further categorized into supervised and unsupervised ones. Supervised learning-based approaches require to use the deformation fields generated by the traditional methods [12, 13] or artificially generated transformations [14, 15] to guide the learning process of the neural network. However, these deformation fields are not accurate enough and may be inconsistent with the real deformations in the clinical environment. Therefore, unsupervised learning-based methods are proposed to achieve registration by exploiting the similarity between images as an intrinsic supervision. Since no extra manual annotations are needed and human workload can be greatly reduced, unsupervised learning-based methods have attracted extensive research attention.

The encoder-decoder architecture of U-shaped network [16] is commonly used for most of unsupervised learning-based methods, where the moving image is warped only once by the deformation field predicted by the network as shown in Fig. 1 (a). However, the performance of UNet-based methods [17, 18, 19, 20, 21] deteriorates greatly when handling large deformations. To overcome this drawback, the module-based and structure-based methods are proposed. For the module-based methods, self-attention [22] or diffusion [23] modules are introduced to enhance the encoder parts of networks as depicted in Fig. 1 (b). Although the feature representation abilities of

*Corresponding author. Tel.: +86-27-87792366; Fax: +86-27-87792072;
Email addresses: `m202272336@hust.edu.cn` (Yibo Wang), `d202380982@hust.edu.cn` (Mingwei Wen), `zxmboshi@hust.edu.cn` (Xuming Zhang)

[1] Yibo Wang and Mingwei Wen contribute equally to this article.

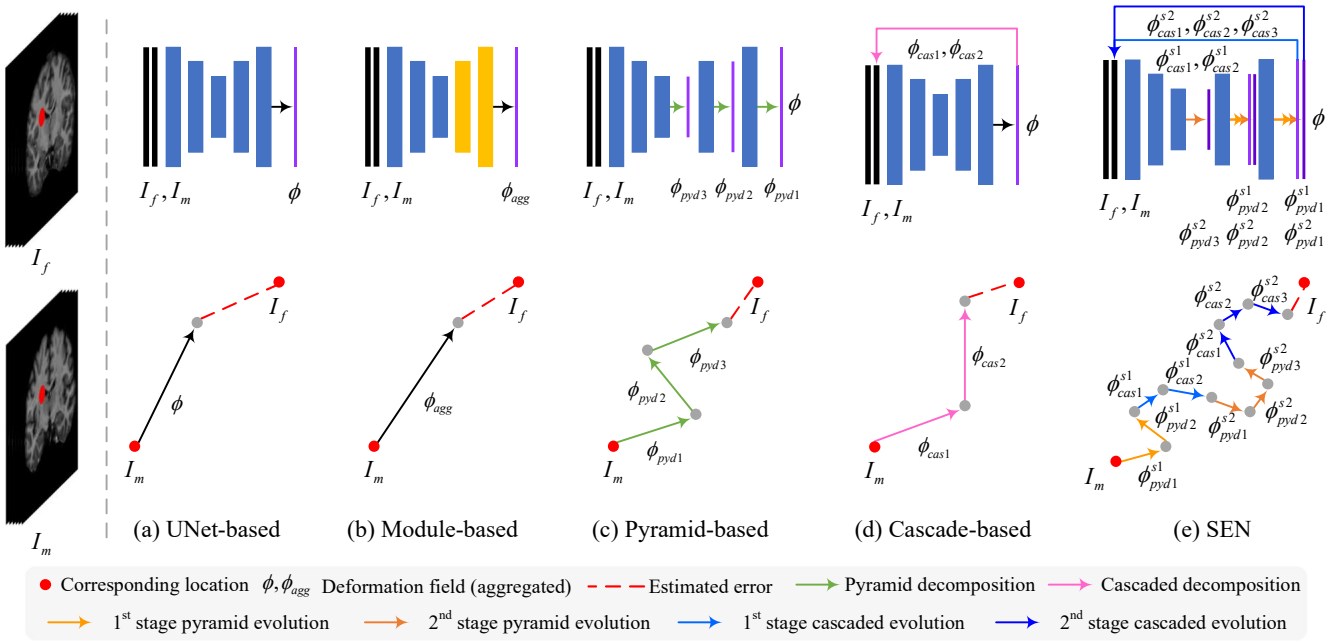

Figure 1. Comparison of various registration methods. The fixed image I_f and the moving image I_m with corresponding anatomical locations are depicted in the left panel. The right panel is divided into two parts, where the network architectures of different methods are presented in the top row and their deformation estimation schemes are illustrated in the bottom row. (a) the UNet-based method: a deformation field ϕ is directly estimated by the network. (b) the module-based method: an aggregated deformation field ϕ_{agg} is directly estimated by the decoder-enhanced network. (c) the pyramid structure based method: an entire deformation field ϕ is decomposed into one-stage multi-level deformation fields ϕ_{pyd1}, ϕ_{pyd2}, ϕ_{pyd3}, which are generated from feature layers at corresponding resolution levels. (d) the cascaded structure based method: an entire deformation field ϕ is decomposed into multiple one-stage deformation fields of the same size ϕ_{cas1}, ϕ_{cas2} using corresponding sub-networks. (e) our proposed SEN: an entire deformation field ϕ is estimated by multi-stage pyramid and cascaded evolution. ϕ^{s1}_{pyd1}, ϕ^{s1}_{pyd2} and ϕ^{s2}_{pyd1}, ϕ^{s2}_{pyd2}, ϕ^{s2}_{pyd3} are the decomposed deformation fields of different levels in the first and second stage pyramid evolution, respectively. ϕ^{s1}_{cas1}, ϕ^{s1}_{cas2} and ϕ^{s2}_{cas1}, ϕ^{s2}_{cas2}, ϕ^{s2}_{cas3} are the decomposed deformation fields of different sub-networks in the first and second stage cascaded evolution, respectively.

these methods are improved, they involve complex computations, which is disadvantageous for their applications to scenarios demanding real-time registration. In terms of the structure-based methods, pyramid and cascaded structures are two mainstream choices to learn deformations progressively. The pyramid structure based methods [24, 25, 26, 27] generate multi-level deformation fields from the decoder and warp the moving features in a coarse-to-fine manner as illustrated in Fig. 1 (c). By splitting the whole network into multiple sub-networks, the cascaded structure based methods [28, 29, 30] warp the moving image step-by-step by using the deformation fields generated by the sub-networks as presented in Fig. 1 (d).

Although the pyramid or cascaded structure based methods outperform the direct ones by adopting the progressive warping procedure, they also suffer from the challenge of optimizing the entire network at once. Regarding the pyramid networks [24, 26], they optimize the parameters of all layers by backwarding the loss which is only related to the finest-level deformation field. Disadvantageously, the coarse-level layers maybe fail to capture large deformations due to the lack of effective guidance and the fine-level ones may be effected by errors generated by the insufficiently trained coarse-level layers. As regards the cascaded networks [28, 29], they optimize the parameters of multiple sub-networks at the same time, which is difficult to take each one into full consideration especially when the number of sub-networks is huge. Meanwhile, errors raised by the preceding sub-network are difficult to remove and the distribution of features may be corrupted when they are passed to the following sub-networks. Therefore, how to arrange the optimization order of each level of feature layer in the pyramid networks and each sub-network in the cascaded networks to smooth their learning process is still an open question.

Previous studies [25, 31] have proven that it is an effective strategy to separate the optimization procedure of the entire network and progressively update the parameters of each part. LapIRN [25] feeds the multi-resolution images into three sub-networks with the same structure and optimizes their parameters recursively, which can be viewed as the application of progressive learning to a cascaded network. AMNet [31] downsamples the input images to generate a multi-scale image pyramid and uses an image at a specific resolution to optimize the corresponding layer, which can be seen as an example of progressive learning for a pyramid network. However, these methods ignore the learning discrepancy between different sub-networks or multi-level layers and only adopt the channel concatenation to connect the two separate learning processes. In fact, the network itself can be a good guidance to ease this problem. Exploiting the knowledge contained in the network is a possible solution to the learning gap issue.

In this paper, to address the problems of insufficient parameter optimization and learning gaps in the existing pyramid and cascaded networks, we propose a novel self-distilled evolution-

ary network (SEN) to provide an effective solution to the unsupervised learning-based registration. To reasonably arrange the optimization order of each part within an entire network, SEN adopts an evolution-based learning strategy, where evolution means that each part is progressively updated in multiple training stages. To fill the discrepancy between parts of SEN with different learning statuses, we introduce the idea of knowledge distillation, where the well-trained parts provide guidance to the new ones in a self-distillation manner. Based on these two ideas, we progressively update a single UNet to be an entire SEN which is composed of multiple sub-networks with multi-level layers. As shown in Fig. 1 (e), we firstly add the coarse-level layers into the UNet only with the fine-level layers through the 1st stage depth-wise pyramid evolution and update the new generated layers in the feature-shifted distillation manner. Then, we cascade a new UNet with the same structure to the existing one through the 1st stage broad-wise cascaded evolution and optimize the new generated sub-network using the distribution-consistent distillation scheme. Subsequently, we apply the similar 2nd stage pyramid and cascaded evolution as well as distillation to the existing networks. Finally, an entire SEN with three well-trained sub-networks containing fully optimized multi-level layers is generated. With the pyramid and cascaded evolution and distillation, each part within a SEN can be updated effectively and easily, thus ensuring better registration performance.

The main contributions of our work can be summarized as:

- An innovative self-distilled evolutionary network has been proposed to improve the unsupervised learning-based registration, addressing the problems of insufficient training and learning gap of the existing methods.
- A depth-wise pyramid evolution strategy along with distillation mode has been designed to ease the learning process of multi-scale deformations.
- A broad-wise cascaded evolution procedure as well as corresponding distillation manner has been developed to attenuate the difficulty of deformation decomposition.
- Extensive experiments have been done on both monomodal and multi-modal datasets to demonstrate the efficiency and superiority of SEN.

The remainder of this paper is structured as follows. The related work is reviewed in Section 2. The proposed method is elaborated in Section 3. The experimental settings and the corresponding results are provided in Section 4 and Section 5, respectively. The discussion is provided in Section 6. Finally, the conclusion is made in Section 7.

2. Related Work

2.1. Deformable Image Registration

The traditional methods treat the registration problem as an optimization task and drive a moving image to match the fixed image by minimizing a similarity-based energy function. Typically, these methods include elastic model [5], B-spline based free-form deformation (FFD) model [6], and Demons [7]. To ensure the topology preservation and ensure invertibility, many diffeomorphic methods have also been proposed, including diffeomorphic FFD model [8], diffeomorphic Demons [9], large deformation diffeomorphic metric mapping (LDDMM) [10], and SyN [11]. Due to the insufficient feature representation ability and complicated iterative computation, these methods cannot work well in cases where high accuracy and real-time performance are demanding.

Learning-based methods build a neural network to simulate the mapping function from the input image pair to the output deformation field, where the registration is achieved by optimizing the network parameters. These methods can be further classified into supervised and unsupervised ones. Supervised learning-based methods [12, 13, 14, 15] rely on the ground-truth deformation field to supervise their registrations. Specifically, Fan *et al.* proposed a BIRNet [13] to solve brain magnetic resonance (MR) image registration by utilizing deformation fields generated by the traditional methods. Eppenhof *et al.* [15] introduced the synthetic transformations as supervision to register computer tomography (CT) images. However, since the ground-truth deformation fields are usually inaccessible in the clinical situations and the synthetic transformations cannot reflect the real deformations, supervised learning-based methods are unsuitable for image registration in clinical settings.

By comparison, unsupervised learning-based methods [17, 18, 19, 20, 21] have attracted extensive research attention since no manual annotations are needed so that human workload can be greatly reduced. VoxelMorph [17] serves as a milestone in the field of unsupervised learning-based registration, where a UNet [16] is used to estimate deformation fields directly. Nevertheless, when a large deformation is encountered, a simple U-shaped structure is not sufficient for accurate image registration. To solve this problem, the module-based and structure-based variants have been proposed. Regarding the module-based methods, some modules with strong feature representation ability are added to the naive UNet. For example, TransMorph [22] introduced the shifted-window Transformer [32] to enhance the encoder stage. DiffuseMorph [23] improved the performance of deformation estimation network by adding an auxiliary diffusion network [33]. However, these additional modules usually involve time-consuming computation, leading to the slow inference speed of module-based methods.

Different from the module-based methods, the structure-based ones boost the registration performance by exploiting new network structures, i.e., pyramid and cascaded structures. As regards pyramid structure based methods [24, 25, 26, 27], multi-scale deformation fields are generated and used to warp the moving image or features step-by-step. Specifically, Hu *et al.* [24] proposed a dual-stream pyramid network called Dual-PRNet to integrate moving and fixed features in a coarse-to-fine manner. Furthermore, they developed Dual-PRNet++ [26] by introducing a new cross-correlation layer into the decoder stage to enhance its resilience to the large deformation. LapIRN [25], developed by Mok *et al.*, adopted a Laplacian pyramid framework to learn deformations with multi-resolution image pyramid. For the cascaded structure based methods [28, 29, 30], an entire network is separated into many sub-networks, each of which is arranged in a cascaded manner. Concretely, De

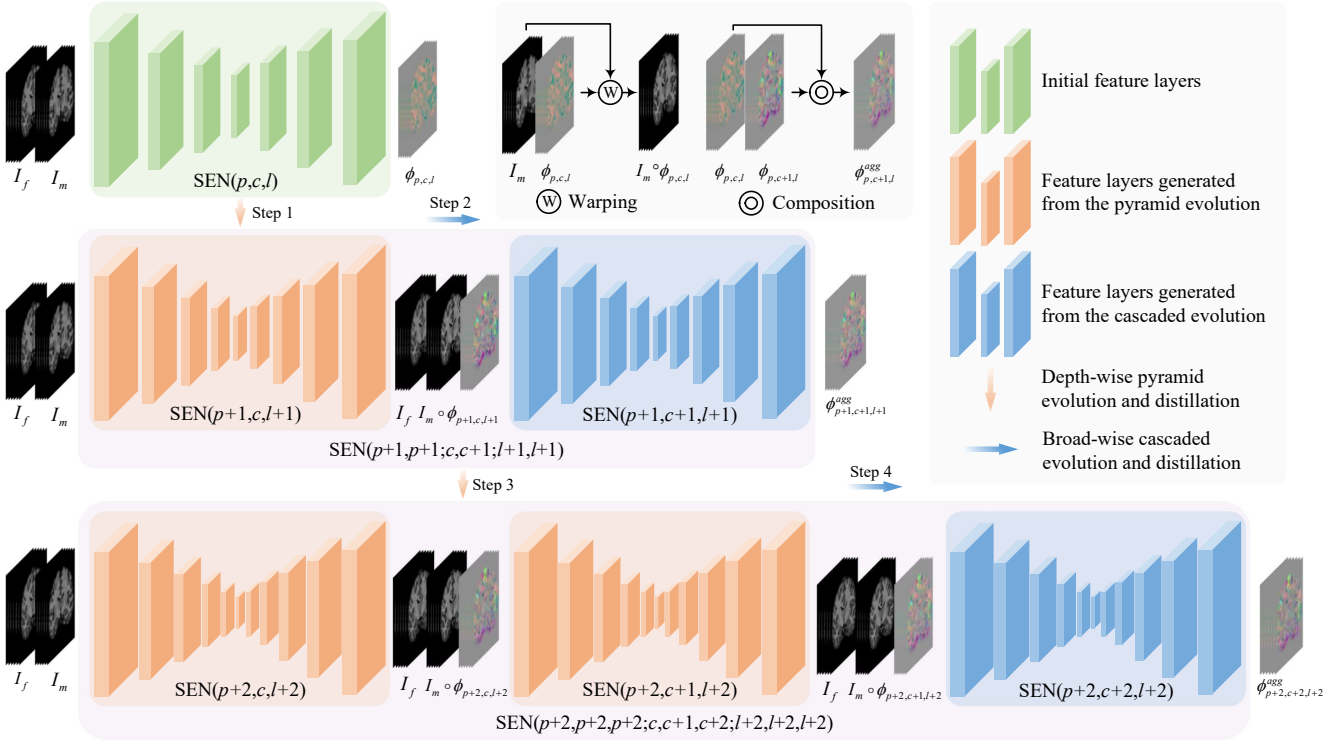

Figure 2. An overview of the proposed self-distilled evolutionary network (SEN). A naive UNet SEN(p, c, l) is initialized as the base network and will be progressively updated through multiple evolution-based steps. After several pyramid and cascaded evolution and distillation, each feature layer and sub-network within SEN is well optimized, yielding a fast and accurate registration network SEN($p+2, p+2, p+2; c, c+1, c+2; l+2, l+2, l+2$).

Vos *et al.* designed the DLIR [28] to achieve both affine and deformable image registration, where the sub-networks were trained sequentially. Zhao *et al.* proposed a RCN to warp the moving image progressively, where the VTN [34] and VoxelMorph [17] were treated as the sub-networks. Hu *et al.* [30] proposed the level-wise and stage-wise recursion, which could be viewed as feature-level and network-level cascade, respectively. However, these methods optimize numerous network parameters at once and ignore the relationship between different layers and sub-networks, which increases the difficulty of network optimization. Differently, our proposed SEN progressively update the parameters of layers and sub-networks by adopting the depth-wise pyramid evolution and the broad-wise cascaded evolution, thereby smoothing the learning process of the entire network and boosting the registration performance.

2.2. Knowledge Distillation

Knowledge distillation (KD) aims to improve a light-weighted student network's performance via transferring "dark" informative knowledge (i.e., probability or feature maps) from a cumbersome teacher model. It was firstly proposed by Hinton *et al.* [35] to enhance a neural network's classification accuracy by learning well-performing teacher model's probability distribution. Since the compact student network trained using KD strategy shows significant improvement in accuracy without extra computation cost, many researchers have explored its application to image classification [36, 37, 38], semantic segmentation [39, 40, 41] and object detection [42, 43]. Specifically, Liu *et al.* [38] built a relational instance graph to transfer more structured knowledge to student model. Likewise, Hou *et al.* [40] explored the relationship between different feature maps and applied cosine similarity to build graphed knowledge to boost the student network's segmentation performance. Li *et al.* [43] adopted a simple but effective approach to achieve precise object detection via directly transferring feature maps from the teacher to the student.

Meanwhile, there are also a large number of KD-related approaches in medical image segmentation [44, 45, 46] and registration [47, 48]. Zhang *et al.* [46] introduced KD scheme and self-attention mechanism to solve cross-modal prostate cancer segmentation problem. Tran *et al.* [48] utilized generative adversarial network (GAN) to distill knowledge for a light-weighted student network in medical image registration. However, these methods have several shortcomings. Firstly, the large teacher model still involves much training time. Secondly, the capability gap between the teacher model and the student network always exists and it is difficult to eliminate. Thirdly, the student network relies on the teacher model, which means that the performance of the former is greatly restricted by the learning capability of the latter. Therefore, finding a more economic way to leverage knowledge within one network will be a feasible means to improve the KD-related methods.

Self-distillation (SD) is a special type of KD, where the teacher model and the student network have the same architecture and are trained simultaneously. Compared with KD-based methods, the SD-related approaches have several advantages. Firstly, since no teacher model exists, the extra training time

can be saved. Secondly, the student network will teach itself and will not be restricted by the teacher model's learning capability. Thirdly, a single network is easy to train and implement. Therefore, many SD-based methods have been proposed in the field of both natural and medical image processing, such as classification [49, 50, 51], segmentation [52, 41, 53, 54] and registration [55]. Specifically, Zhang *et al.* [49] adopted the feature-based loss and logits-based loss to improve a single network's classification performance. SDHNet, presented by Zhou *et al.* [55], adopted the value-based and gradient-based distillation scheme to teach a registration network itself. Differently, we introduce the idea of SD into a novel evolution-based learning strategy and utilize the knowledge within a network itself to guide its optimization. In this way, the learning gap within layers and sub-networks can be effectively filled with the proposed depth-wise pyramid distillation and broad-wise cascaded distillation, ensuring a more smooth learning process.

3. Method

3.1. Overview

The overall framework of the proposed SEN is illustrated in Fig. 2. Given a pair of moving image I_m and fixed image I_f, SEN aims to predict a deformation field ϕ to generate the warped moving image $I_m(\phi) = I_m \circ \phi$ which can be spatially aligned to I_f. Different from the existing one-stage learning methods, SEN relies on multi-stage learning processes to achieve better registration performance.

We consider SEN(p, c, l) as the initial network, where p, c and l denote the pyramid index, the cascaded index and the level of one network, respectively. A pair of p and c absolutely decides one model's training order during the whole learning process. At the beginning of training, p and c are both set to 0 and increase step-by-step with multiple pyramid and cascaded evolution. $l \in \{1, 2, ..., 5\}$ is the number of encoder-decoder convolution pairs, and the size of coarsest feature map within one network is $1/2^{l-1}$ that of the input image. Here, we take the evolutionary path from SEN(p, c, l) to SEN$(p+2, p+2, p+2; c, c+1, c+2; l+2, l+2, l+2)$ through double pyramid and cascaded evolution as an example to introduce the workflow of SEN.

Firstly, we train the initial SEN(p, c, l) until convergence by minimizing the difference between $I_m(\phi_{p,c,l})$ and I_f, where $\phi_{p,c,l}$ is the predicted deformation field by SEN(p, c, l). Then, we transfer the trained SEN(p, c, l) to SEN$(p+1, c, l+1)$ by adopting our proposed depth-wise pyramid evolution. To fill the learning gap between SEN(p, c, l) and SEN$(p+1, c, l+1)$, we apply the corresponding pyramid distillation to assist the training of SEN$(p+1, c, l+1)$. Compared with SEN(p, c, l), SEN$(p+1, c, l+1)$ can achieve higher registration accuracy as it contains more convolutions for extracting deeper features. The detailed mechanism of depth-wise pyramid evolution and distillation will be elaborated in Section 3.2.

Secondly, we append a new SEN$(p+1, c+1, l+1)$ following the trained SEN$(p+1, c, l+1)$ based on the proposed broad-wise cascaded evolution. To maintain the distribution consistency between SEN$(p+1, c, l+1)$ and SEN$(p+1, c+1, l+1)$, a similar cascaded distillation is adopted to smooth the learning process

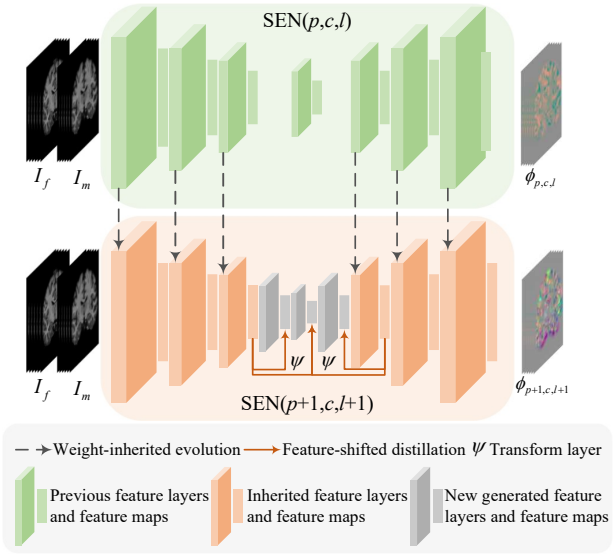

Figure 3. The illustration of depth-wise pyramid weight-inherited evolution and feature-shifted distillation.

of SEN$(p+1, c+1, l+1)$. The corresponding workflow of broad-wise cascaded evolution and distillation will be introduced in Section 3.3. Different from the previous pyramid evolution, an integrated network SEN$(p+1, p+1; c, c+1; l+1, l+1)$ is built, where the interior sub-networks are connected based on the warping and composition [29]. Specifically, a warped moving image $I_m(\phi_{p+1,c,l+1})$ serves as the input of SEN$(p+1, c+1, l+1)$. The initial $\phi_{p+1,c+1,l+1}$ is composited with $\phi_{p+1,c,l+1}$ to yield the aggregated deformation field $\phi^{agg}_{p+1,c+1,l+1}$ to further boost the registration performance of the entire network.

Finally, we adopt one more pyramid and cascaded evolution and distillation on the base of SEN$(p+1, p+1; c, c+1; l+1, l+1)$ to generate SEN$(p+2, p+2, p+2; c, c+1, c+2; l+2, l+2, l+2)$. It should be noted that all interior sub-networks in the integrated network follow the rule of pyramid evolution and they are connected by warping and composition operations.

In the following subsections, we firstly make the elaborated analysis of the proposed depth-wise pyramid evolution and distillation as well as broad-wise cascaded evolution and distillation. Then, we introduce the symmetric UNet which is designed to meet the evolutionary requirement of SEN. Next, we summarize the loss function and overall learning process of SEN. Finally, we describe the diffeomorphic version of SEN.

3.2. Depth-wise Pyramid Evolution and Distillation

Previous pyramid structure based methods mainly keep their image-level or feature-level network architecture unchanged during the whole training process [25, 26]. However, it is actually difficult for the network itself to optimize such a large number of interior layers at once. Meanwhile, with the increasing training iterations, it is easier for the network to update the unlearned layers under the guidance of trained ones. To ease the network optimization procedure and boost the accuracy, we propose the pyramid weight-inherited evolution and feature-shifted distillation as depicted in Fig. 3.

Given a trained SEN(p, c, l), the depth-wise pyramid weight-inherited evolution involves the following three steps. Firstly,

we copy the weights of all encoder-decoder pairs in SEN(p, c, l) to initialize those of the same structures in SEN$(p+1, c, l+1)$. Secondly, we append a new $(l+1)$-level encoder-decoder pair following the l-level one to enhance the depth-exploiting ability of SEN. Note that we choose to build the unlearned encoder-decoder pair rather than use the trained intermediate layer to initialize them even though they have the same structure. Thirdly, we add a new intermediate layer between the $(l+1)$-level encoder and decoder in SEN$(p+1, c, l+1)$ connecting these two parts into a whole. Compared with SEN(p, c, l), the evolved SEN$(p+1, c, l+1)$ not only inherits its updated parameters but also has deeper feature extraction layers, making it possible to achieve better registration performance.

When the evolution is finished, how to optimize the new generated layers is a challenging problem. Here, we propose the depth-wise pyramid distillation to smooth their learning procedure. Let $\mathbf{F}_l^{enc}$ and $\mathbf{F}_l^{dec}$ denote the feature maps generated from the trained l-level encoder and decoder layers in SEN$(p+1, c, l+1)$, respectively. Likewise, the corresponding feature maps $\mathbf{F}_{l+1}^{enc}$ and $\mathbf{F}_{l+1}^{dec}$ will be yielded from the unlearned $(l+1)$-level encoder and decoder layers. Since $\mathbf{F}_l^{enc}$ and $\mathbf{F}_{l+1}^{enc}$ are both responsible for encoding image features, it is reasonable for us to use $\mathbf{F}_l^{enc}$ to guide $\mathbf{F}_{l+1}^{enc}$. The similar strategy will be used in the decoder's learning process. Based on the above analysis, the distillation loss within the same type of feature maps can be written as:

$$\mathcal{L}_{pyd}^{same} = \|\psi(\mathbf{F}_l^{enc}) - \mathbf{F}_{l+1}^{enc}\|_2 + \|\psi(\mathbf{F}_l^{dec}) - \mathbf{F}_{l+1}^{dec}\|_2, \quad (1)$$

where $\|\cdot\|_2$ is the Frobenius $l2$-norm; ψ is the transformation layer to match the feature channels and spatial resolution. Concretely, it is composed of a $1\times1\times1$ convolutional layer followed by a LeakyReLU layer and a batch normalization layer.

For the intermediate layer used to connect the encoder and decoder layers, its feature map $\mathbf{F}_{l+1}^{inter}$ is similar to both $\mathbf{F}_{l+1}^{enc}$ and $\mathbf{F}_{l+1}^{dec}$. Therefore, we choose to deploy both $\mathbf{F}_{l+1}^{enc}$ and $\mathbf{F}_{l+1}^{dec}$ to guide $\mathbf{F}_{l+1}^{inter}$. Correspondingly, the distillation loss among the different types of feature maps can be formulated as:

$$\mathcal{L}_{pyd}^{diff} = \|\psi(\mathbf{F}_{l+1}^{enc}) - \mathbf{F}_{l+1}^{inter}\|_2 + \|\psi(\mathbf{F}_{l+1}^{dec}) - \mathbf{F}_{l+1}^{inter}\|_2. \quad (2)$$

Finally, the overall depth-wise pyramid distillation loss $\mathcal{L}_{pyd}$ can be computed using $\mathcal{L}_{pyd}^{same}$ and $\mathcal{L}_{pyd}^{diff}$. Note that we will not add any auxiliary balance weight as $\mathcal{L}_{pyd}^{same}$ and $\mathcal{L}_{pyd}^{diff}$ are equally important for the learning process of SEN:

$$\mathcal{L}_{pyd} = \mathcal{L}_{pyd}^{same} + \mathcal{L}_{pyd}^{diff}. \quad (3)$$

With this scheme, the optimization process of SEN is decomposed into multiple depth levels. SEN can focus on one specific deformation at one time and gradually incorporate greater depth-exploiting ability with the evolution goes. Meanwhile, the discrepancy between the trained layers and newly generated ones can be effectively filled with our proposed distillation manner. Therefore, the depth-wise pyramid evolution and distillation overcome the drawback of static depth-level modeling in the existing pyramid structure based approaches.

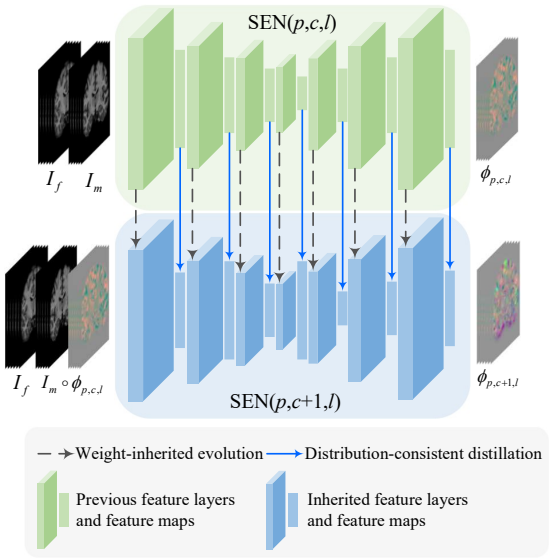

Figure 4. The illustration of broad-wise cascaded weight-inherited evolution and distribution-consistent distillation.

3.3. Broad-wise Cascaded Evolution and Distillation

Existing cascade-based methods always set up a huge model by recursively cascading multiple sub-networks and optimize them simultaneously during the whole training process [29, 55]. In this way, these methods are difficult to effectively optimize their interior parameters as the learned gradient cannot pass through a lot of sub-networks without decay. At the same time, they ignore the intrinsic distribution consistency of the predictions from sub-networks, rendering the generated deformation field not robust enough. To overcome the two drawbacks, we propose the cascaded weight-inherited evolution and distribution-consistent distillation, as illustrated in Fig. 4.

Compared with the pyramid evolution, the broad-wise cascaded weight-inherited evolution is much easier as no extra interior layers are introduced. Differently, a new SEN$(p, c+1, l)$ is initialized with the weights of a trained SEN(p, c, l) and closely appended after it. Since SEN$(p, c+1, l)$ has exactly the same structure as SEN(p, c, l), the learned parameters of the later can transfer to the corresponding layers of the former completely. Meanwhile, the integrated model SEN$(p, p; c, c+1; l, l)$ is built by connecting SEN(p, c, l) with SEN$(p, c+1, l)$. For simplicity, we ignore the composition operation here.

To maintain the consistency of predicted deformation fields within the integrated network, we introduce the distribution-consistent distillation to regularize SEN$(p, c+1, l)$. Specifically, we deploy the probability distribution maps of SEN(p, c, l) to guide those of SEN$(p, c+1, l)$. Let $\{\mathbf{F}_{c,i}^{enc}\}_{i=1}^{l}$, $\mathbf{F}_c^{inter}$, $\{\mathbf{F}_{c,i}^{dec}\}_{i=1}^{l}$ represent the feature maps produced by the encoder, intermediate and decoder layers in SEN(p, c, l), respectively. Following this idea, $\{\mathbf{F}_{c+1,i}^{enc}\}_{i=1}^{l}$, $\mathbf{F}_{c+1}^{inter}$,$\{\mathbf{F}_{c+1,i}^{dec}\}_{i=1}^{l}$ are the corresponding feature maps produced by the layers in SEN$(p, c+1, l)$. Considering the paired feature maps $\{\mathbf{F}_{c,i}^{enc}\}_{i=1}^{l}$ and $\{\mathbf{F}_{c+1,i}^{enc}\}_{i=1}^{l}$, we minimize their probability difference to achieve the distribution consistency, which can be formulated as:

$$\mathcal{L}_{cas}^{enc} = \sum_{i=1}^{l} \mathrm{KL}(\mathrm{Softmax}(\mathbf{F}_{c,i}^{enc}), \mathrm{Softmax}(\mathbf{F}_{c+1,i}^{enc})), \quad (4)$$

where KL(·) represents the Kullback-Leibler divergence and Softmax(·) denotes the Softmax operation along the channel and all spatial axes of the feature map.

For the paired decoders, the distillation loss on their feature maps can be written as:

$$\mathcal{L}_{cas}^{dec} = \sum_{i=1}^{l} \mathrm{KL}(\mathrm{Softmax}(\mathbf{F}_{c,i}^{dec}), \mathrm{Softmax}(\mathbf{F}_{c+1,i}^{dec})). \quad (5)$$

Compared with the computation of feature maps from the encoder and decoder, the distribution-consistent regularization on the intermediate feature maps is much easier, which can be defined as:

$$\mathcal{L}_{cas}^{inter} = \mathrm{KL}(\mathrm{Softmax}(\mathbf{F}_{c}^{inter}), \mathrm{Softmax}(\mathbf{F}_{c+1}^{inter})). \quad (6)$$

Finally, the overall broad-wise cascaded distillation loss $\mathcal{L}_{cas}$ can be computed by adding $\mathcal{L}_{cas}^{enc}$, $\mathcal{L}_{cas}^{dec}$, and $\mathcal{L}_{cas}^{inter}$, where no extra weight is deployed as done in the pyramid distillation:

$$\mathcal{L}_{cas} = \mathcal{L}_{cas}^{enc} + \mathcal{L}_{cas}^{inter} + \mathcal{L}_{cas}^{dec}. \quad (7)$$

Through this strategy, the optimization process of SEN is decoupled into many cascaded steps. SEN can convert a complex deformation to contiguous decomposed deformations and learn each of them using one specific sub-network. Besides, the distribution consistency of multiple predictions within an integrated SEN is ensured, making the final predicted deformation field more reliable. Based on the above analysis, the problems of existing cascade-based methods can be effectively solved by the proposed broad-wise cascaded evolution and distillation.

3.4. Symmetric UNet

Different from the most popular registration method VoxelMorph [19], we design a completely symmetric UNet as the base network of the proposed SEN as illustrated in Fig. 5. Concretely, the symmetric UNet takes the concatenated $I_m(\phi^{pre})$ and I_f as the input images and yields ϕ^{cur} as the output deformation field. Apart from the input and output layers, the framework of symmetric UNet is composed of three parts: the encoder layers, the intermediate layer, and the decoder layers. Among all encoder layers, the convolutional layer with a kernel size of 3×3×3 and a stride of 2 followed by a LeakyReLU layer is chosen to generate the hierarchical encoder feature maps. For the decoder layers, the $3\times3\times3$ convolutional layer and the upsampling layer are used to receive the encoder feature maps through the skip connections, generating the corresponding decoder feature maps. The intermediate layer is a $1\times1\times1$ convolutional layer without activation, serving as a connection between the encoder and decoder layers. It worth mentioning that the number of encoder-decoder layers and the size of intermediate layer are dynamic, and they depend on the specific training order of SEN. In Fig. 5, we just take the case of $l = 5$ as an example and the similar principle will be followed for other cases.

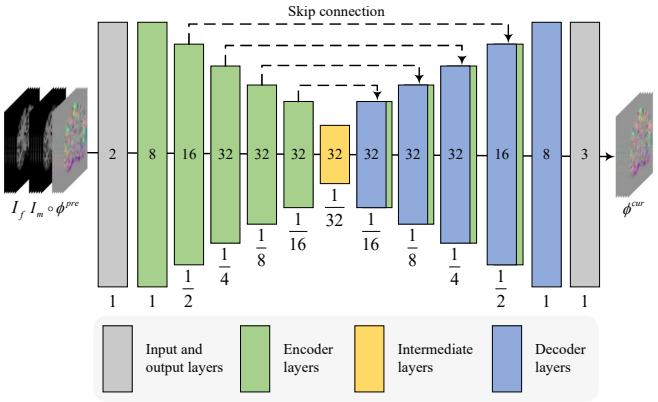

Figure 5. The illustration of symmetric UNet, which serves as the base network of SEN. The channels and resolutions of feature maps are presented in the middle and bottom of each rectangle.

3.5. Loss Function

The overall loss function of SEN consists of four aspects: the similarity term $\mathcal{L}_{sim}$, the regularization term $\mathcal{L}_{reg}$, the depth-wise pyramid distillation term $\mathcal{L}_{pyd}$, and the broad-wise cascaded distillation term $\mathcal{L}_{cas}$.

For the similarity term $\mathcal{L}_{reg}$, we adopt the negative cross-correlation (NCC) to minimize the difference between $I_m(\phi)$ and I_f, which can be formulated as:

$$\mathcal{L}_{sim}(I_m(\phi), I_f) = -\sum_{v\in\Omega} \frac{\left(\sum_{v_i}(I_f(v) - \bar{I}_f(v))(I_m(\phi(v)) - \bar{I}_m(\phi(v)))\right)^2}{\left(\sum_{v_i}(I_f(v) - \bar{I}_f(v))\right)^2\left(\sum_{v_i}(I_m(\phi(v)) - \bar{I}_m(\phi(v)))\right)^2}, \quad (8)$$

where $v \in \Omega$ is the voxel coordinate in the image domain and $\bar{I}$ represents the mean value in a $9\times9\times9$ local window centered at the voxel v. A higher NCC value indicates better alignment, and thus $\mathcal{L}_{sim}$ drives $I_m(\phi)$ to match I_f spatially.

For the regularization term $\mathcal{L}_{reg}$, we deploy the Frobenius $l2$-norm to penalize the gradient of ϕ, ensuring that a smooth deformation field is generated without folds:

$$\mathcal{L}_{reg}(\phi) = \sum_{v\in\Omega} \|\nabla\phi(v)\|_2. \quad (9)$$

With the defined depth-wise pyramid distillation term $\mathcal{L}_{pyd}$ and the broad-wise cascaded distillation term $\mathcal{L}_{cas}$, the final loss function of SEN can be written as:

$$\mathcal{L} = \mathcal{L}_{sim} + \lambda_{reg}\mathcal{L}_{reg} + \lambda_{pyd}\mathcal{L}_{pyd} + \lambda_{cas}\mathcal{L}_{cas}, \quad (10)$$

where λ_{reg}, λ_{pyd}, and λ_{cas} are the hyperparamters to balance the weights of regularization and distillation terms.

Optionally, to achieve better registration performance, we introduce the segmentation loss to train SEN in a semi-supervised manner. It is achieved by maximizing the overlap between segmentation maps, which can be formulated as:

$$\mathcal{L}_{seg}(S_m(\phi), S_f) = \frac{1}{K}\sum_{k=1}^{K} \frac{2\cdot|S_m^k(\phi) \cap S_f^k|}{|S_m^k(\phi)| + |S_f^k|}, \quad (11)$$

where $S_m(\phi) = S_m \circ \phi$ and S_f represent the warped moving segmentation map and the fixed segmentation map, respectively.

Furthermore, the overall forward and backward algorithm to train SEN is provided in Algorithm 1.

3.6. Diffeomorphic Version of SEN

Apart from the aforementioned displacement version SEN, we also provide the diffeomorphic version SEN-diff to ensure topology preservation and deformation invertibility. Similar to the existing methods [20, 25], we replace the displacement field

Algorithm 1 Self-distilled Evolution Learning Strategy

Input: Moving image I_m, fixed image I_f, initial pyramid index p, cascaded index c, network level l, untrained model SEN(p, c, l) with uniformly initialized parameters $\theta_{p,c,l}$, number of pyramid evolution N_p, number of cascaded evolution N_c, training iterations per evolution N_{iter}, hyperparameters λ_{reg}, λ_{pyd}, and λ_{cas}.

Output: Well-trained SEN$(p + N_p, p + N_p, ..., p + N_p; c, c + 1, ..., c + N_c; l + N_p, l + N_p, ..., l + N_p)$ with its parameter set $\{\theta_{p+N_p,c,l+N_p}, \theta_{p+N_p,c+1,l+N_p}, ..., \theta_{p+N_p,c+N_c,l+N_p}\}$ for each sub-network.

while N_p is not **0** and N_c is not **0 do**
  // Consider the situation that SEN(p, c, l) has not been trained
  if SEN(p, c, l) has not been trained **then**
    Set λ_{pyd}, λ_{cas} to **0**.
    for $i = 1$ **to** N_{iter} **do**
      $\phi_{p,c,l} \leftarrow$ Feed I_m and I_f to SEN(p, c, l) and generate the deformation field.
      $\mathcal{L} \leftarrow$ Compute the overall loss with I_m, I_f, and $\phi_{p,c,l}$ via Eq. 10.
      $\nabla\theta_{p,c,l} \leftarrow$ Backward $\mathcal{L}$ and compute the gradient of $\theta_{p,c,l}$.
      $\theta_{p,c,l} \leftarrow \theta_{p,c,l} + \nabla\theta_{p,c,l}$, update $\theta_{p,c,l}$.
    end for
  end if
  // Apply the depth-wise pyramid evolution and distillation once
  Set λ_{cas} to **0**.
  SEN$(p + 1, c, l + 1) \leftarrow$ Update SEN(p, c, l) using the depth-wise pyramid evolution.
  for $i = 1$ **to** N_{iter} **do**
    $\phi_{p+1,c,l+1} \leftarrow$ Feed I_m and I_f to SEN$(p + 1, c, l + 1)$ and generate the deformation field.
    $\mathcal{L} \leftarrow$ Compute the overall loss with I_m, I_f, and $\phi_{p+1,c,l+1}$ via Eq. 10.
    $\nabla\theta_{p+1,c,l+1} \leftarrow$ Backward $\mathcal{L}$ and compute the gradient of $\theta_{p+1,c,l+1}$.
    $\theta_{p+1,c,l+1} \leftarrow \theta_{p+1,c,l+1} + \nabla\theta_{p+1,c,l+1}$, update $\theta_{p+1,c,l+1}$.
  end for
  $N_p \leftarrow N_p - 1$, update N_p.
  // Apply the broad-wise cascaded evolution and distillation once
  Set λ_{pyd} to **0**.
  SEN$(p + 1, p + 1; c, c + 1; l + 1, l + 1) \leftarrow$ Update SEN$(p + 1, c, l + 1)$ using the broad-wise cascaded evolution.
  for $i = 1$ **to** N_{iter} **do**
    $\phi_{p+1,c+1,l+1} \leftarrow$ Feed I_m and I_f to SEN$(p + 1, p + 1; c, c + 1; l + 1, l + 1)$ and generate the deformation field.
    $\mathcal{L} \leftarrow$ Compute the overall loss with I_m, I_f, and $\phi_{p+1,c+1,l+1}$ via Eq. 10.
    $\{\nabla\theta_{p+1,c,l+1}, \nabla\theta_{p+1,c+1,l+1}\} \leftarrow$ Backward $\mathcal{L}$ and compute the gradient of $\{\theta_{p+1,c,l+1}, \theta_{p+1,c+1,l+1}\}$.
    $\{\theta_{p+1,c,l+1}, \theta_{p+1,c+1,l+1}\} \leftarrow \{\theta_{p+1,c,l+1} + \nabla\theta_{p+1,c,l+1}, \theta_{p+1,c+1,l+1} + \nabla\theta_{p+1,c+1,l+1}\}$, update $\{\theta_{p+1,c,l+1}, \theta_{p+1,c+1,l+1}\}$.
  end for
  $N_c \leftarrow N_c - 1$, update N_c.
end while

with the stationary velocity field and further integrate it with several time steps to obtain the deformation field. Specifically, the relation between the stationary velocity field $\mathcal{V}$ and the deformation field ϕ follows an ordinary differential equation, which can be formulated as:

$$\frac{\mathrm{d}\phi^{(t)}}{\mathrm{d}t} = \mathcal{V}(\phi^{(t)}),\ s.t.\phi^{(0)} = Id, \tag{12}$$

where Id represents the identical transformation. Meanwhile, we apply the squaring and scaling operation [56] to integrate $\mathcal{V}$ over unit time with $K = 7$ time steps to obtain ϕ. Note that the training procedure of SEN-diff is consistent with that of SEN.

4. Experiment

4.1. Datasets

We conduct comprehensive experiments to evaluate our proposed method on four mono-modal datasets including LPBA40, OASIS, SLIVER, LSPIG and one multi-modal dataset MMWHS2017.

The LPBA40 [57] dataset consists of 40 T1-weighted MR brain scans, each of which is correspondingly segmented into 56 anatomical structures. We adopt the subject-to-subject registration strategy here, which means that two images will be randomly chosen from the dataset and one of chosen images (i.e., the moving image) will be aligned to another one (i.e., the fixed image). We firstly select 25, 5, 10 images from the dataset, then generate 600, 20, 90 image pairs for training, validation, and testing, respectively. After the preprocessing steps like center-cropping, the resultant image resolution is $160 \times 192 \times 160$.

The OASIS [58] dataset is a well-known public registration dataset, which is composed of 414 T1-weighted MR brain scans along with 35 annotated regions for each scan. Similar to LPBA40, we choose 290, 40, 84 images from the dataset,

yielding 83810, 1560, 6972 image pairs in the training, validation and testing datasets, respectively. We follow the standard preprocessing steps as done in VoxelMorph [19]. Finally, the image resolution of this dataset is $160 \times 192 \times 224$.

The SLIVER [59] dataset includes 20 CT scans of human liver with manual segmentation labels. We deploy MSD [60] and BFH [29] as the training set and treat SLIVER as the validation and testing sets. The number of images in MSD and that in BFH are 993 and 92, respectively. We follow the same registration strategy and preprocessing steps of LPBA40, ensuring that the resolution of this dataset is $128 \times 128 \times 128$.

The LSPIG [29] dataset involves 20 CT scans of pig liver along with corresponding anatomical segmentation. Similar to the SLIVER dataset, we directly use the trained weight on MSD and BFH datasets to evaluate various methods on the LSPIG dataset. Since the liver structures of human and pig are similar, the LSPIG dataset can be viewed as a good choice to test our method's generalization ability. The final resolution of this dataset is same to that of SLIVER.

The MMWHS2017 [61] dataset is derived from the Multi-modality Whole Heart Segmentation 2017 Challenge, which contains 20 pairs of MR and CT heart scans. All scans are labeled with seven structures, including the left ventricle (LV) and right ventricle (RV) blood cavity, the left atrium (LA) and right atrium (RA) blood cavity, the myocardium of the left ventricle (MYO), the ascending aeorta (AA), and the pulmonary artery (PA). We choose 15 pairs of MR-CT images to generate augmented 180 image pairs using the random flipping and rotation for network training, and use the left 5 image pairs for validating and testing all methods. Since images in this dataset have been aligned by the challenge holder, we add the simulated deformations to images using the B-spline based FFD model [6], thereby ensuring the similar registration procedure to other studies.

4.2. Evaluation Metrics

We evaluate the registration performance of various methods using the Dice [62], 95% maximum Hausdorff distance (HD95) [63] and Folds [19]. Dice is designed to measure the overlap degree between the warped segmentation map and the ground-truth label. The closer to 1 the Dice value is, the better registration is achieved. It should be noted that we average the Dice values from multiple anatomical regions as the final result for the LPBA40, OASIS and MMWHS2017 datasets. Meanwhile, we adopt HD95 to measure the difference between the warped segmentation outlines and the ground-truth ones. The lower HD95 score is, the better alignment of outlines is guaranteed. Moreover, we calculate the negative percentage of the predicted deformation field ($|J(\phi)| < 0$) as Folds ratio to evaluate the diffeomorphic property. The closer to 0 the Folds ratio is, the fewer folds of the deformation field are generated and the better topology is preserved.

4.3. Baseline Methods

We compare our proposed method with two traditional methods and seven unsupervised learning based methods. For the traditional methods, we implement SyN [11] and NiftyReg [64], which are commonly used registration solutions in research and clinic environment. Regarding the unsupervised learning based methods, we choose the naive U-shaped networks, including VoxelMorph (VM) [17], its diffeomorphic version VoxelMorph-diff (VM-diff)[18] and VTN [34] as well as the module-based and structure-based approaches. For the module-based method, TransMorph is chosen due to its representative attention mechanism. For the structure-based methods, LapIRN, its diffeomorphic version LapIRN-diff [25] and DualPRNet++ (DualPR) [26] are chosen as the pyramid structure based baseline methods and multi-cascaded VTN (mC-VTN) is chosen as the cascaded structure based baseline method. Moreover, to highlight the superiority of our proposed self-distillation scheme, a self-distilled based network SDHNet [55] is also compared.

4.4. Implementation Details

We implement all baseline methods and our proposed SEN based on the PyTorch framework. A high-performance computer with an Intel I7-6950X CPU and two NVIDIA GeForce RTX 3090 GPUs is used to run all experiments. We choose the Adam optimizer with default momentum values of 0.9 and 0.99 to update the parameters of all methods. For the baseline methods, we set the initial learning rate to $1e^{-4}$ and run 10^4 steps with the cosine decay. For the proposed SEN, the total optimization steps are same as those of baseline methods for the fair comparison and specific iterations for each evolution are determined by averaging the overall steps. We set the initial level of SEN to 3 and apply pyramid and cascaded evolution and distillation twice in turn to get a SEN$(2, 2, 2; 0, 1, 2; 5, 5, 5)$ as the default network in the following experiments. Based on the ablation study on the validation set, λ_{reg}, λ_{pyd}, and λ_{cas} are set to 0.1, 0.001, and 0.001, respectively. The default settings of SEN-diff are consistent with those of SEN.

5. Results

In this section, we provide quantitative and qualitative comparisons of our method with various baseline methods on the brain, liver, and heart datasets. The ablation studies of our method are also presented.

5.1. Comparison with Baseline Methods on Brain Datasets

We firstly provide quantitative comparison results in Table 1. Compared with the traditional methods SyN and NiftyReg, our proposed SEN achieves better registration performance in terms of Dice and HD95 scores on LPBA40 and OASIS datasets. Besides, SEN largely reduces the inference time from 2752/32 s to 0.38 s, which means that SEN has the potential to be implemented in the real-time registration scenario. Compared with UNet-based methods, SEN still performs better in terms of both similarity-based and smoothness-based metrics. Specifically, SEN outperforms VM with the improvements of Dice values by about 1.7 % on the LPBA40 dataset and 14.4 % on the OASIS dataset, respectively. As regards the Folds ratio, SEN-diff achieves the consistent performance while VM-diff degrades to some extent, which means that our proposed learning strategy can better preserve the topology of deformation fields.

Table 1. Comparison results of SEN and different baseline methods on the LPBA40 and OASIS datasets. The allocated memory of GPU and time are measured on the OASIS dataset during the testing phase. The standard deviations are listed within brackets. **Bold** font indicates the best value.

Methods	LPBA40			OASIS			GPU	Time
	Dice (%)	Folds (%)	HD95	Dice (%)	Folds (%)	HD95	(MB)	(s)
SyN	78.7 (3.9)	**0**	7.1 (1.6)	73.8 (1.4)	**0**	4.3 (2.3)	-	2752
NiftyReg	80.6 (3.2)	0.3 (0.1)	7.2 (2.3)	76.2 (1.3)	0.4 (0.1)	4.0 (2.5)	-	32
VM	81.6 (2.4)	0.7 (0.2)	5.6 (1.4)	73.4 (1.6)	1.8 (0.3)	3.1 (3.2)	2246.4	1.6e-3
VM-diff	80.9 (2.7)	**0**	5.7 (1.4)	72.4 (1.7)	0.7 (0.2)	3.0 (2.8)	2247.5	0.13
VTN	82.1 (2.0)	2.4 (0.3)	5.3 (1.1)	77.7 (1.8)	3.8 (0.2)	2.9 (2.4)	2264.7	1.5e-3
TransMorph	82.2 (2.1)	3.1 (0.2)	5.4 (1.2)	81.1 (2.1)	3.5 (0.1)	3.1 (3.8)	6481.8	0.95
DualPR	82.1 (2.0)	1.4 (0.3)	5.5 (1.1)	73.8 (1.9)	2.5 (0.3)	3.3 (3.7)	6590.3	0.59
LapIRN	82.5 (2.2)	2.5 (0.5)	5.8 (1.5)	80.5 (1.8)	2.7 (0.3)	3.0 (3.8)	7268.2	5.9e-3
LapIRN-diff	82.4 (2.1)	**0**	5.4 (1.5)	81.8 (1.5)	2.3 (0.2)	2.9 (4.2)	7268.2	0.81
2C-VTN	82.3 (2.1)	2.9 (0.5)	5.4 (1.1)	81.3 (1.6)	3.3 (0.2)	2.7 (4.2)	2654.8	0.39
3C-VTN	82.6 (2.1)	3.0 (0.4)	5.3 (1.4)	82.2 (1.4)	3.1 (0.2)	2.6 (3.3)	3043.8	0.64
SDHNet	82.5 (1.9)	2.1 (1.1)	5.3 (1.4)	81.5 (1.4)	2.0 (0.2)	3.8 (1.5)	2236.5	1.37
SEN	**83.0** (1.8)	1.1 (0.4)	**4.2** (1.1)	**84.0** (1.3)	1.5 (0.2)	**2.1** (1.2)	2948.5	0.38
SEN-diff	82.9 (1.3)	**0**	4.7 (1.3)	82.4 (1.6)	**0**	2.3 (1.3)	2948.5	0.57

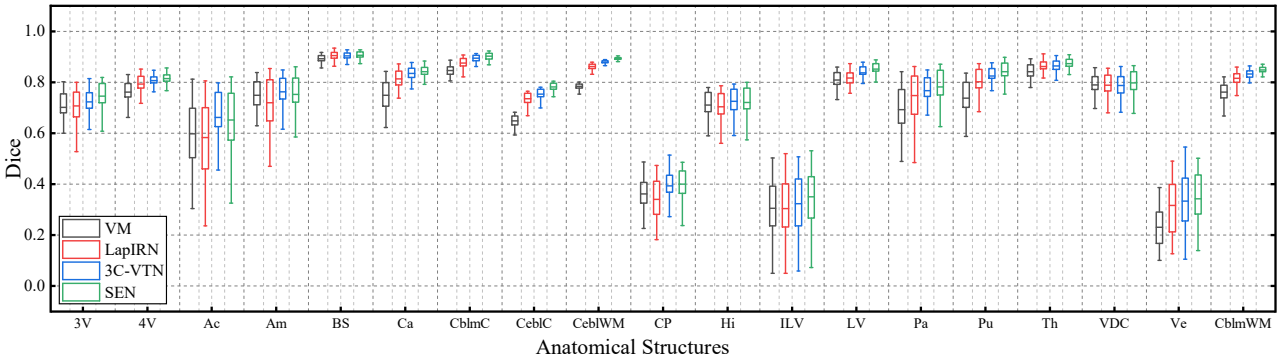

Figure 6. Boxplots of VM, LapIRN, 3C-VTN and our SEN in terms of Dice values on the OASIS dataset. All anatomical structures are presented in the horizontal axis, including: 3rd ventricle (3V), 4th ventricle (4V), Accumbens (Ac), amygdala (Am), brain stem (BS), caudate (Ca), cerebellum-cortex (CblmC), cerebral cortex (CeblC), cerebellum-white-matter (CblmWM), cerebral-white-matter (CeblWM), choroid-plexus (CP), hippocampus (Hi), Inf-Lat-Ventricle (ILV), lateral-ventricle (LV), pallidum (Pa), putamen (Pu), thalamus (Th), Ventral-DC (VDC), vessel (Ve).

In comparison with TransMorph, SEN yields higher registration accuracy while requiring less GPU memory allocation. Compared with all structure-based methods, SEN provides better Dice and HD95 values. Concretely, SEN improves the Dice score from 82.1 % to 83.0 % and reduces the HD95 from 5.5 to 4.2 compared with DualPR on the LPBA40 dataset. Meanwhile, SEN improves the Dice score from 82.2 % to 84.0 % and reduces the HD95 from 2.6 to 2.1 compared with 3C-VTN on the OASIS dataset. Moreover, SEN outperforms the self-distillation based method SDHNet in terms of Dice values and Folds ratio, which indicates that our proposed distillation manner is more effective for the network learning process.

Fig. 6 presents the boxplots of Dice values among all anatomical structures on the OASIS dataset to verify the robustness of our proposed SEN. Clearly, SEN outperforms other methods on both large and small anatomical structures. As regards some large regions like BS, CblmC, and CeblWM, SEN achieves higher Dice values as the broad-wise cascaded evolution and distillation facilitate realizing the large deformation decomposition effectively. For some small regions like Pa, Pu, and Th, SEN also performs best among all evaluated methods, which indicates that the depth-wise pyramid evolution and distillation familiarize SEN with small deformations.

Furthermore, we provide qualitative visualization of various methods on the LPBA40 and OASIS datasets in Fig. 7. The results in the 1st row show that the warped image generated by SEN is the closest to the fixed image. From the overlapped images in the 3rd row, it can be seen that SEN can align both large and small brain regions well, which is meaningful for assisting in brain disease diagnosis and surgical navigation. In the 2nd and 4th rows, the error maps predicted by SEN include fewer mis-alignments of image details, which indicates that our method can provide the best registered result among all evaluated methods.

5.2. Comparison with Baseline Methods on Liver Datasets

To further evaluate the registration performance of SEN on the mono-modal dataset, we compare our SEN with other baseline methods on two CT datasets in Table 2. On the SLIVER dataset, SEN outperforms SyN and NiftyReg by boosting the Dice scores from 93.7%/95.4% to 99.2%. SEN provides about 8.5% improvement in the Dice score and about 40.0% reduction in the Folds rate over VTN. Compared with TransMorph,

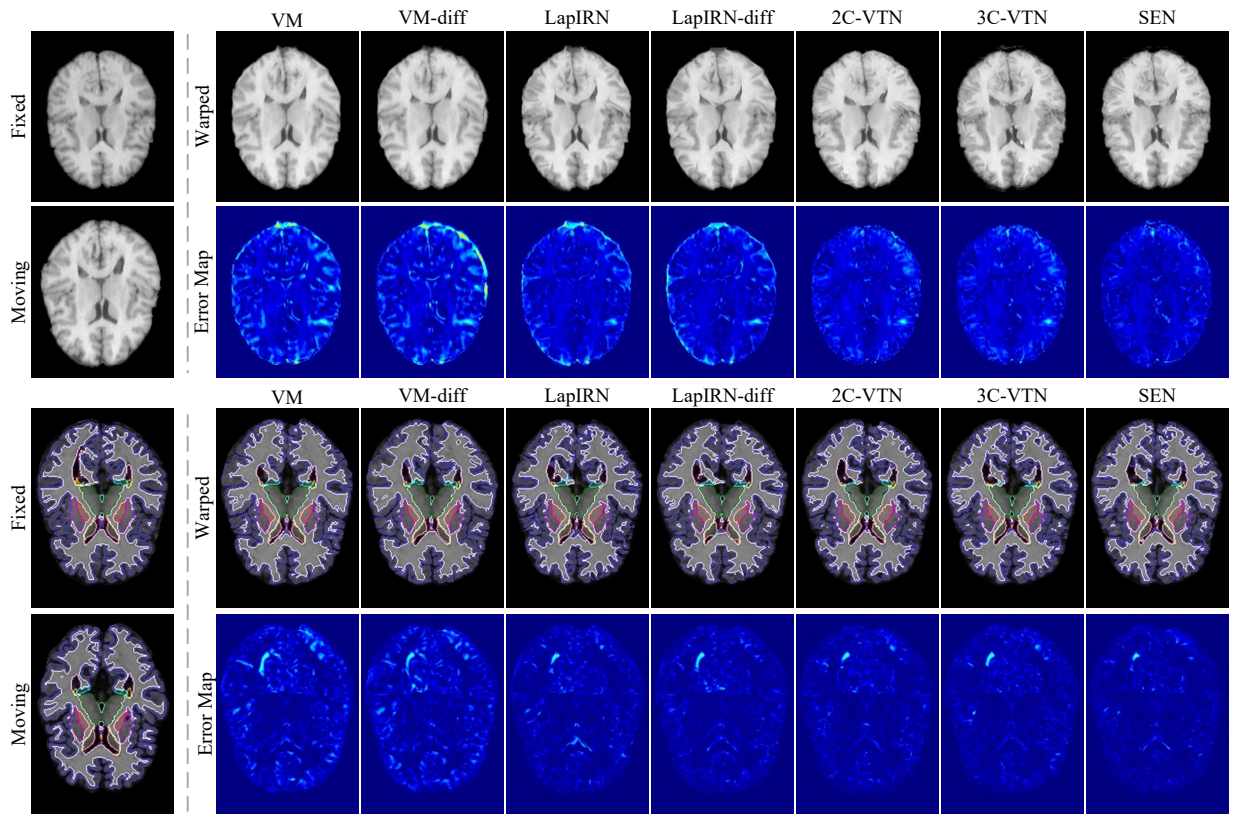

Figure 7. Qualitative illustrations of various methods on the LPBA40 dataset (top) and OASIS dataset (bottom). The 1st and 3rd rows show the fixed image and warped images generated by various methods on the LPBA40 and OASIS datasets, respectively. The 2nd and 4th rows show the error maps, which measure the difference between the fixed image and warped images generated by various methods on the two datasets. Images from the OASIS dataset are highlighted with several representative anatomical structures.

SEN not only largely improves the registration accuracy (Dice: 94.2%→99.2%) but also preserves the smoothness of the deformation field better (Folds: 4.2%→1.5%). The comparison among two structure based methods and SEN shows that the performance of our method is still superior. Specifically, SEN outperforms DualPR and LapIRN by increasing the Dice scores from 96.4%/97.2% to 99.2%. In contrast to 2C-VTN and 3C-VTN, SEN achieves higher Dice value (98.6%/98.8%→99.2%) while greatly reducing the Folds ratio (4.0%/3.9%→1.5%). Besides, SEN still performs better than SDHNet by improving the Dice value from 98.3% to 99.2%.

To verify the adaptability of our proposed SEN, we further conduct comparison experiments on the LSPIG dataset as presented in Table 2. Compared with the SLIVER dataset, the LSPIG dataset is more challenging as the liver structure of pigs is largely different from that of humans. Obviously, the performance of SEN and SEN-diff on this dataset is consistent with that on the SLIVER dataset. Specifically, SEN provides the Dice score of 89.6%, outperforming VM by 6.0% and VTN by 23.1%. Meanwhile, SEN outperforms DualPR and 3C-VTN by yielding 2.4% and 1.6% improvements of the Dice values. Different from the significant degradation of SDHNet (Dice: 98.3%→66.5%), the performance of SEN is more stable (Dice: 99.2%→89.6%), which proves that our proposed distillation manner is more robust.

Similarly, we illustrate the registered images and segmentation maps on the SLIVER and LSPIG datasets in Fig. 8. The comparison of the fixed image and warped images in the 1st row shows that SEN achieves the best registration performance as its warped image matches the fixed one most closely. In the 2nd row, different from the insufficient alignment of VM, the segmentation map generated by SEN can correspond to the liver region in the fixed image well, which indicates that the proposed evolution strategy is more powerful than the direct U-shaped network. The performance of SEN in the 3rd row is consistent with that in the 1st row, which further verifies the robustness of SEN. The analysis of the segmentation maps in the 4th row further demonstrates the significant superiority of SEN to other methods. For example, the segmentation map generated by SEN involves fewer artifacts than that by LapIRN in the 2rd column and matches the liver region in the fixed image better than that by 3C-VTN in the 3rd column.

5.3. Comparison with Baseline Methods on Heart Datasets

To verify the effectiveness of SEN in the multi-modal image registration, we compare SEN with all baseline methods on the MMWHS2017 dataset as presented in Table 3. According to the different modalities of moving image and fixed image, we further categorize this dataset into two subsets, i.e. MR→CT and CT→MR subsets, where the former aligns MR images to CT images while the later works in an opposite way. Obviously, our proposed SEN achieves better Dice and HD95 scores

Table 2. Comparison results of SEN and different baseline methods on the SLIVER and LSPIG datasets. The standard deviations are listed within brackets. **Bold** font indicates the best value.

Methods	SLIVER		LSPIG	
	Dice (%)	Folds (%)	Dice (%)	Folds (%)
SyN	93.7 (1.1)	**0**	83.5 (1.0)	**0**
NiftyReg	95.4 (1.2)	0.4 (0.3)	86.1 (0.7)	2.5 (1.5)
VM	93.0 (2.7)	2.6 (0.5)	83.6 (3.0)	3.4 (0.8)
VM-diff	97.0 (0.7)	1.3 (0.5)	86.5 (3.7)	2.0 (0.5)
VTN	91.4 (2.8)	2.5 (0.6)	66.5 (6.7)	2.2 (0.3)
TransMorph	94.2 (2.0)	4.2 (0.2)	86.8 (5.4)	3.0 (0.3)
DualPR	96.4 (0.8)	2.6 (0.3)	87.5 (3.2)	4.2 (1.1)
LapIRN	97.2 (0.9)	3.1 (0.4)	88.2 (2.7)	7.6 (0.7)
LapIRN-diff	97.0 (0.7)	2.2 (0.4)	85.4 (3.0)	3.2 (0.3)
2C-VTN	98.6 (0.2)	4.0 (0.7)	74.1 (6.5)	2.3 (0.2)
3C-VTN	98.8 (0.1)	3.9 (0.7)	88.2 (3.0)	4.2 (0.8)
SDHNet	98.3 (0.2)	2.5 (0.6)	69.5 (6.1)	2.7 (0.9)
SEN	**99.2** (0.1)	1.5 (0.6)	**89.6** (3.4)	1.3 (0.3)
SEN-diff	97.5 (0.6)	1.1 (0.4)	88.7 (3.4)	1.0 (0.2)

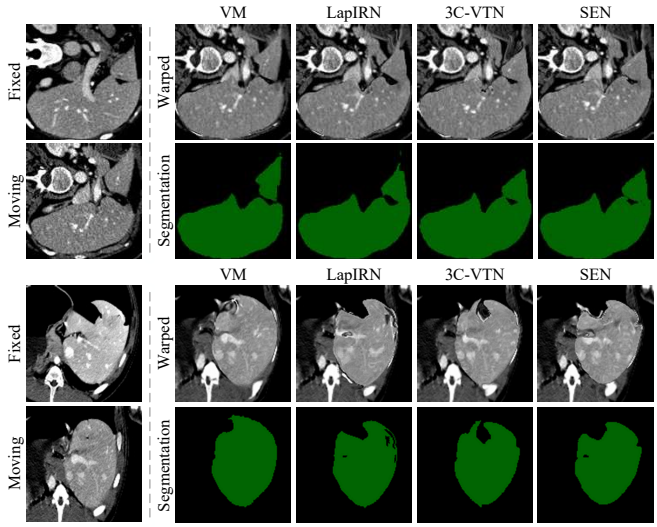

Figure 8. Qualitative illustrations of various methods on the SLIVER dataset (top) and LSPIG dataset (bottom). The 1st and 3rd rows show the fixed image and warped images generated by various methods on the SLIVER and LSPIG dataset, respectively. The 2nd and 4th rows show the moving image and warped segmentation map generated by various methods on the two datasets.

than SyN and NiftyReg on both MR→CT and CT→MR subsets. SEN outperforms VM by increasing the Dice values from 71.4% to 77.1% and from 70.6% to 78.3% on the two subsets, respectively. As regards the diffeomorphic versions, our proposed SEN-diff achieves better topology preservation than VM-diff as it reduces the Folds rate from 0.8% to 0.6% and reduces HD95 from 5.3 to 4.4 on the MR→CT subset.

Compared with TransMorph, SEN not only achieves higher registration accuracy (Dice: 75.4%/77.7% → 77.1%/78.3%) but also involves fewer parameters and lower computation complexity (Params/FLOPs: 46.5/136.82 → 0.88/65.18). SEN outperforms DualPR and LapIRN by boosting the Dice value from 76.5%/71.8% to 77.1% on the MR→CT subset. Compared

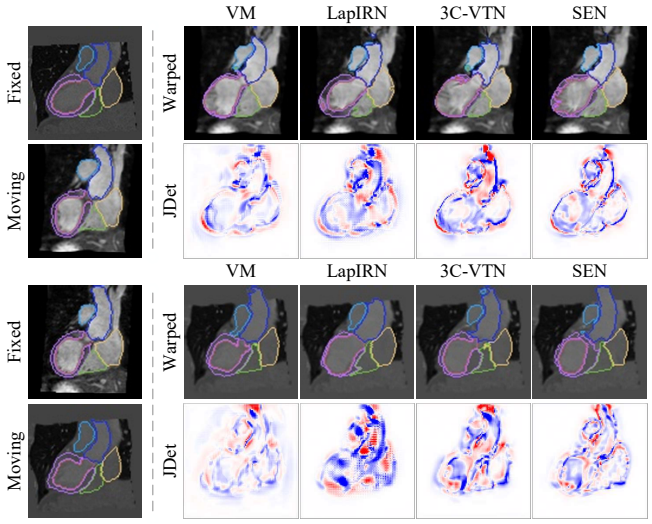

Figure 9. Qualitative illustrations of various methods on the MMWHS2017 dataset, including MR→CT (top) and CT→MR (bottom) subsets. The 1st and 3rd rows show the fixed image and warped images generated by various methods on the MR→CT and CT→MR subsets, respectively. The 2nd and 4th rows show the moving image and jacobian determinant of deformation fields generated by various methods on the two subsets, where red color denotes the fold. Important anatomical structures are highlighted in images.

with 2C-VTN and 3C-VTN, SEN achieves better performance on the CT→MR subset as it yields higher Dice value and involves less computation. SEN also demonstrates the superiority to SDHNet in that it improves the Dice score about 0.9% and 1.4% on the two subsets, respectively.

Besides, we provide qualitative visualizations of some representative methods on the MMWHS2017 dataset in Fig. 9. By jointly observing the 1st and 3rd rows, it reveals that our proposed SEN can align major cardiac structures like LA, LV, RA, RV on both MR→CT and CT→MR subsets well. From the results in the 2nd and 4th rows, we can see that there are fewer red regions in the jacobian determinant of SEN, which means that SEN reduces the folds and preserves the deformation fields effectively.

5.4. Ablation Studies

Here, we further investigate whether distillation and evolution involved in SEN are useful and explore how they contribute to the success of SEN through extensive ablation studies on both OASIS and MMWHS2017 CT→MR datasets.

5.4.1. Effectiveness of Distillation

We firstly employ five networks (corresponding to five rows in Table 4) to investigate the influence of the proposed distillation on the performance of SEN. Specifically, the 1st row represents a vanilla SEN(0;0;3) trained in an end-to-end way. Four networks from the 2nd to 5th rows denote four SEN(1,1;0,1;4,4) produced using single pyramid and cascaded evolution under different distillation settings. The network in the 2nd row is trained without using any distillation strategy. The networks in the 3rd and 4th rows are updated only using pyramid or cascaded distillation. The network in the 5th row

Table 3. Comparison results of SEN and different baseline methods on the MMWHS2017 dataset. The number of floating point operations (FLOPs) is measured during the testing phase. The standard deviations are listed within brackets. **Bold** font indicates the best value.

Methods	MR→CT			CT→MR			Params	FLOPs
	Dice (%)	Folds (%)	HD95	Dice (%)	Folds (%)	HD95	(MB)	(G)
SyN	73.6 (3.4)	**0**	7.3 (3.3)	73.4 (2.3)	**0**	7.6 (3.4)	-	-
NiftyReg	72.9 (3.8)	**0**	5.9 (3.4)	72.5 (1.6)	**0**	6.6 (4.1)	-	-
VM	71.4 (7.5)	0.9 (0.2)	5.1 (2.7)	70.6 (6.6)	1.3 (0.2)	6.1 (4.3)	0.28	31.28
VM-diff	70.0 (7.8)	0.8 (0.1)	5.3 (3.0)	70.3 (6.2)	0.9 (0.1)	6.1 (4.4)	0.28	31.28
VTN	73.0 (7.7)	1.0 (0.2)	5.1 (3.0)	72.6 (6.7)	1.3 (0.2)	5.7 (3.8)	0.92	63.52
TransMorph	75.4 (7.6)	1.7 (0.2)	5.1 (3.3)	77.7 (5.0)	1.8 (0.1)	5.4 (3.9)	46.5	136.82
DualPR	76.5 (5.4)	1.3 (0.3)	4.8 (3.4)	77.3 (5.2)	2.0 (0.5)	5.6 (2.2)	1.00	170.55
LapIRN	71.8 (8.3)	1.5 (0.4)	5.5 (3.8)	77.8 (6.4)	1.6 (0.3)	4.9 (3.5)	0.68	402.85
LapIRN-diff	72.2 (6.7)	0.5 (0.1)	5.0 (3.3)	76.9 (6.6)	0.9 (0.1)	4.7 (3.2)	0.68	402.85
2C-VTN	76.4 (6.5)	1.3 (0.2)	5.1 (3.2)	75.9 (6.9)	1.6 (0.2)	5.5 (3.8)	1.84	127.05
3C-VTN	76.2 (6.6)	1.3 (0.3)	5.0 (3.2)	76.6 (6.4)	1.7 (0.2)	5.1 (3.7)	2.76	190.57
SDHNet	76.4 (6.1)	0.5 (0.2)	4.6 (2.9)	77.2 (5.9)	0.7 (0.2)	4.5 (3.3)	4.12	46.18
SEN	**77.1** (4.6)	1.1 (0.2)	**4.2** (2.3)	**78.3** (5.5)	1.4 (0.2)	**4.1** (1.8)	0.88	65.18
SEN-diff	77.0 (4.4)	0.6 (0.1)	4.4 (2.7)	78.1 (5.4)	1.5 (0.2)	4.2 (1.6)	0.88	65.18

Table 4. Ablation study of the effectiveness of distillation on the OASIS and MMWHS2017 CT→MR datasets. Pyd means the pyramid distillation. Cas denotes the cascaded distillation. The standard deviations are listed within brackets. **Bold** font indicates the best value.

Networks	Pyd	Cas	Dice (%)	
			OASIS	CT→MR
SEN(0,0,3)	-	-	69.5 (2.1)	67.8 (6.4)
SEN(1,1;0,1;4,4)	×	×	80.1 (1.8)	74.7 (6.9)
SEN(1,1;0,1;4,4)	×	✓	80.3 (1.3)	74.8 (6.5)
SEN(1,1;0,1;4,4)	✓	×	80.5 (1.9)	74.9 (6.6)
SEN(1,1;0,1;4,4)	✓	✓	**80.7** (1.5)	**75.2** (6.4)

Table 5. Ablation study of the modes of distillation on the OASIS and MMWHS2017 CT→MR datasets. Pyd means the online pyramid distillation. Cas denotes the online cascaded distillation. The standard deviations are listed within brackets. **Bold** font indicates the best value.

Networks	Pyd	Cas	Dice (%)	
			OASIS	CT→MR
SEN(0,0,3)	-	-	69.5 (2.1)	67.8 (6.4)
SEN(1,1;0,1;4,4)	×	×	77.2 (1.7)	73.5 (6.6)
SEN(1,1;0,1;4,4)	✓	×	79.3 (1.5)	74.5 (6.6)
SEN(1,1;0,1;4,4)	×	✓	79.2 (1.4)	74.4 (6.5)
SEN(1,1;0,1;4,4)	✓	✓	**80.7** (1.5)	**75.2** (6.4)

is obtained using both pyramid and cascaded distillation. Table 4 clearly shows that our proposed evolution strategy is effective as the evolved networks show a significant increase in the Dice score compared with the naive one (Dice: 69.5% → 80.7%). Meanwhile, the comparison of Dice values in the 2nd, 3rd and 4th rows demonstrate that the proposed pyramid distillation and cascaded distillation both contribute to the success of SEN as they improve the Dice score about 2.5‰ and 5‰, respectively. When comparing the 5th row with other four rows, we can see that using both pyramid and cascaded distillation is the best choice since the network in the 5th row achieves the highest Dice scores of 80.7% and 75.2% on the OASIS and MMWHS2017 datasets, respectively.

5.4.2. Modes of Distillation

After verifying the proposed distillation is useful, we further employ five networks corresponding to five rows in Table 5 to investigate which mode of distillation is the best. Here, the modes mainly include online and offline distillation, where the former means that the weights of trained layers or sub-networks are unlocked while the later means that they are locked during the optimization of new ones. In Table 5, the network in the 1st row is a vanilla SEN(0;0;3) trained in an end-to-end way. The networks from the 2nd to 5th rows are four SEN(1,1;0,1;4,4) obtained through single pyramid and cascaded evolution under different modes of distillation. By comparing the results listed in the 2nd, 3rd, and 4th rows, it is obvious that the online distillation is more suitable. Concretely, the network in the 3rd row surpasses that in the 2nd row in terms of the Dice score, which indicates that the online pyramid distillation can better enhance the learning process of multi-scale deformation. Meanwhile, the network in the 4th row outperforms that in the 2nd row by improving the Dice value by about 2.6% on the OASIS dataset and 2.5% on the MMWHS2017 CT→MR dataset, which means that the online cascaded distillation can facilitate the learning process of large deformation decomposition. From the above comparisons, it can be seen that the online mode is useful for both pyramid and cascaded distillation. Furthermore, the results reported in the 5th row confirm that the online mode still works well when the pyramid and cascaded distillations are implemented simultaneously. Therefore, we adopt the online distillation while updating the network via evolution in the following ablation studies.

5.4.3. Number of Evolution

We explore the effect of the number of evolution on the performance of SEN by employing nine networks corresponding to nine rows in Table 6. We categorize these nine networks

Table 6. Ablation study of the number of evolution on the OASIS and MMWHS2017 CT→MR datasets. $\xrightarrow{\text{Pyd}}$ means updating network through the pyramid evolution. $\xrightarrow{\text{Cas}}$ means updating network through the cascaded evolution. The standard deviations are listed within brackets. **Bold** font indicates the best value.

Networks	Dice (%)	
	OASIS	CT→MR
SEN(0,0,3)	69.5 (2.1)	67.8 (6.4)
SEN(0,0,3)$\xrightarrow{\text{Cas}}$ SEN(0,0;0,1;3,3)	78.0 (1.4)	71.4 (6.4)
SEN(0,0,3)$\xrightarrow{\text{Cas}}\xrightarrow{\text{Cas}}$ SEN(0,0,0;0,1,2;3,3,3)	81.5 (1.0)	74.5 (6.2)
SEN(0,0,3)$\xrightarrow{\text{Pyd}}$ SEN(1,0,4)	74.9 (1.8)	71.6 (6.4)
SEN(1,0,4)$\xrightarrow{\text{Cas}}$ SEN(1,1;0,1;4,4)	80.7 (1.6)	75.2 (6.4)
SEN(1,0,4)$\xrightarrow{\text{Cas}}\xrightarrow{\text{Cas}}$ SEN(1,1,1;0,1,2;4,4,4)	82.3 (1.4)	76.3 (6.3)
SEN(0,0,3)$\xrightarrow{\text{Pyd}}\xrightarrow{\text{Pyd}}$ SEN(2,0,5)	76.8 (1.9)	73.4 (6.5)
SEN(2,0,5)$\xrightarrow{\text{Cas}}$ SEN(2,2;0,1;5,5)	81.3 (1.5)	75.6 (6.5)
SEN(2,0,5)$\xrightarrow{\text{Cas}}\xrightarrow{\text{Cas}}$ SEN(2,2,2;0,1,2;5,5,5)	**83.5** (1.1)	**76.8** (6.2)

into three groups, each of which contains a base network, a single cascaded evolved network, and a double cascaded evolved network. The networks in the 2nd and 3rd group are generated through applying single and double pyramid evolution to the base network in the 1st group, respectively. On the OASIS dataset, by comparing the 1st, 4th and 7th rows, it can be seen that the Dice score is improved (Dice: 69.5%→74.9%→76.8%) through applying more pyramid evolution, which indicates that the increasing number of pyramid evolution is conductive to improving the registration performance of SEN. Besides, within each group, deploying more cascaded evolution improves the accuracy of SEN, verifying the positive effect of more cascaded evolution. For instance, in the 1st group, the double cascaded evolved network outperforms the single one as well as the base network by improving the Dice score by about 4.5% and 17.3%, respectively. Moreover, by comparing the single and double cascaded evolved networks among different groups, it is obvious that applying more pyramid and cascaded evolution has more significant influence on the registration performance of SEN than only deploying pyramid or cascaded evolution. However, employing excessive evolution will impose significant computational burden on our method, thereby greatly reducing its efficiency. Therefore, we choose the single and double evolution in the following experiments.

5.4.4. Different Evolutionary Paths

In Table 7, we employ eleven networks to further investigate the effect of different evolutionary paths on the registration performance of SEN. Generally, we categorize these eleven networks into two groups and one separate network. The base network in these two groups is a vanilla SEN(0;0;3) trained in an end-to-end manner. Meanwhile, the final generated network in these two groups is also same, i.e., SEN(2,2,2;0,1,2;5,5,5). However, the evolutionary path from the base network to the final network is different, where the pyramid evolution is placed ahead of the cascaded one for the 1st group, but the reversed operation is implemented for the 2nd group. Besides, we provide the results of a SEN(0,0,0;0,1,2;5,5,5) trained in an end-to-end way for the comparison with our proposed evolutionary strategy. From the results of the 1st group including the 1st to 5th rows and the 2nd group including the 6th to 10th rows, it can be seen that deploying the pyramid-cascaded evolutionary paths is more effective than deploying the cascaded-pyramid ones as the former outperforms the later by increasing the Dice score from 83.6% to 84.0% on the OASIS dataset. Meanwhile, the results in the 5th, 10th, and 11th rows show that the networks trained in our proposed evolutionary strategy performs better than that directly optimized in the end-to-end manner, which indicates that our strategy plays a positive role in the network optimization.

Table 7. Ablation study of the evolutionary paths on the OASIS and MMWHS2017 CT→MR datasets. $\xrightarrow{\text{Pyd}}$ means to update network through the pyramid evolution. $\xrightarrow{\text{Cas}}$ means to update network through the cascaded evolution. The standard deviations are listed within brackets. **Bold** font indicates the best value.

Networks	Dice (%)	
	OASIS	CT→MR
SEN(0,0,3)	69.5 (2.1)	67.8 (6.4)
SEN(0,0,3)$\xrightarrow{\text{Pyd}}$ SEN(1,0,4)	74.9 (1.8)	71.6 (6.4)
SEN(1,0,4)$\xrightarrow{\text{Cas}}$ SEN(1,1;0,1;4,4)	80.7 (1.5)	75.2 (6.4)
SEN(1,1;0,1;4,4)$\xrightarrow{\text{Pyd}}$ SEN(2,2;0,1;5,5)	81.5 (1.5)	75.8 (6.4)
SEN(2,2;0,1;5,5)$\xrightarrow{\text{Cas}}$ SEN(2,2,2;0,1,2;5,5,5)	**84.0** (1.3)	**78.3** (6.5)
SEN(0,0,3)	69.5 (2.1)	67.8 (6.4)
SEN(0,0,3)$\xrightarrow{\text{Cas}}$ SEN(0,0;0,1;3,3)	78.0 (1.4)	71.4 (6.4)
SEN(0,0;0,1;3,3)$\xrightarrow{\text{Pyd}}$ SEN(1,1;0,1;4,4)	80.8 (1.5)	75.0 (6.2)
SEN(1,1;0,1;4,4)$\xrightarrow{\text{Cas}}$ SEN(1,1,1;0,1,2;4,4,4)	82.7 (1.1)	76.2 (6.2)
SEN(1,1,1;0,1,2;4,4,4)$\xrightarrow{\text{Pyd}}$ SEN(2,2,2;0,1,2;5,5,5)	83.6 (1.6)	77.9 (6.4)
SEN(0,0,0;0,1,2;5,5,5)	82.8 (1.5)	75.4 (6.6)

5.4.5. Loss Weights

Finally, we employ seven networks corresponding to seven rows in Table 8 to investigate the impact of different loss

Table 8. Ablation study of the loss weights on the OASIS and MMWHS2017 CT→MR datasets. The standard deviations are listed within brackets. **Bold** font indicates the best value.

λ_{reg}	λ_{pyd}	λ_{cas}	Dice (%)	
			OASIS	CT→MR
0.05	0.001	0.001	78.9 (1.4)	74.0 (6.4)
0.2	0.001	0.001	79.8 (1.4)	74.9 (6.7)
0.1	0.0005	0.001	79.7 (1.4)	74.9 (6.5)
0.1	0.002	0.001	79.6 (1.4)	74.6 (6.3)
0.1	0.001	0.0005	79.5 (1.4)	74.7 (6.4)
0.1	0.001	0.002	79.4 (1.4)	74.7 (6.3)
0.1	0.001	0.001	**80.7** (1.5)	**75.2** (6.4)

weights on the performance of SEN. All results are produced by a SEN(1,1;0,1;4,4) evolved from a SEN(0;0;3) via single pyramid and cascaded evolution and distillation. The results listed in the 1th, 2nd, and 7th rows clearly demonstrate the effect of λ_{reg} on the registration accuracy of SEN. Specifically, setting λ_{reg} to be 0.1 is optimal as the corresponding trained network outperforms those using $\lambda_{reg} = 0.05$ and $\lambda_{reg} = 0.2$ in terms of the Dice score. Furthermore, the comparison among the 3rd, 4th, and 7th rows shows that the best choice for λ_{pyd} is 0.001 as the network under this setting improves the Dice score from 79.7% to 80.7% over that trained using $\lambda_{pyd} = 0.0005$ and from 79.6% to 80.7% over that trained using $\lambda_{pyd} = 0.002$ on the OASIS dataset. Likewise, the results in the 5th, 6th, and 7th rows show that the optimal value of λ_{cas} can be set to 0.001 as the network trained with this setting achieves the best Dice value. Therefore, we fix λ_{reg}, λ_{pyd} and λ_{cas} to 0.1, 0.001 and 0.001, respectively, to ensure the satisfactory performance of SEN.

6. Discussion

In this work, we present a self-distilled evolutionary network, SEN, to achieve fast and accurate deformable image registration via pyramid and cascaded progressive learning. Different from existing methods, SEN adopts the evolution-based training scheme to efficiently optimize each part within a network and utilizes the self-distillation based learning strategy to fill the capability gap between layers or sub-networks with different statuses. Here, we firstly provide a comprehensive analysis to investigate why our proposed SEN can outperform various baseline methods from the point of mechanism and strategy. Then, we explore the effectiveness of all designed components to further interpret the success of SEN. Next, we provide more fruitful experiments to verify the variability and robustness of SEN. Finally, we discuss the drawbacks of our existing SEN and propose potential directions for further improvement.

6.1. Self-distilled Evolution-based Learning Strategy

The current registration methods mainly rely on building the U-shaped network [17, 34], adding various modules [22, 23], or changing the pyramid or cascaded network structures [25, 26, 29] to achieve registration. However, these methods mostly adopt the one-stage training manner, which means that the entire network with plenty of parameters needs to be optimized during the whole training process, increasing the convergence time of network and restricting the registration accuracy. To solve this problem, we propose a novel evolution-based training manner by separating the whole optimization procedure into multiple stages and updating the parameter of each feature layer or sub-network per stage. In this way, all parts within a network can be sufficiently optimized to boost the registration performance. Meanwhile, we introduce the idea of knowledge distillation to assist in multi-stage training, further smoothing the whole learning process.

Benefiting from these two innovative designs, our SEN outperforms various baseline methods on both mono-modal and multi-modal datasets as shown in Tables 1, 2, 3 and Figs. 7, 8, 9. Compared with the direct UNet-based methods VM and VTN, SEN largely improves the registration accuracy in terms of the Dice score on all datasets as the depth-wise pyramid evolution familiarizes SEN with multi-scale deformations and the broad-wise cascaded evolution teaches SEN how to decompose a large deformation. Even though TransMorph improves VM with the attention mechanism, the extra matrix computation makes it difficult to converge and slows down its inference speed. Nevertheless, SEN avoids introducing complex modules and uses the evolution-based strategy to enhance a simple UNet, leading to its easy training and fast inference speed. Compared with the pyramid structure based methods LapIRN and Dual-PRNet++, SEN further improves the multi-scale deformation decomposition by using the trained fine-level layers to guide the untrained coarse-level ones in the self-distillation manner, thereby ensuring that the network only needs to learn one-level deformation per stage and errors will not pass through different levels. Different from 2C-VTN and 3C-VTN, SEN can effectively maintain the distribution consistency between different sub-networks via the probability map based distillation, prohibiting the accumulation of deformation estimation errors when the number of sub-networks increases. Besides, compared with the similar self-distillation based method SDHNet, our SEN still performs better as it relies on the evolution-based dynamic training manner rather than optimizing the parameters of a static network during the whole learning process.

6.2. The Effectiveness of Designs in SEN

The superior performance of SEN is mainly attributed to the evolution-based training strategy and the self-distillation based optimization scheme. We conduct sufficient ablation studies to investigate the contributions of these two designs to the success of SEN. Since the evolution depends on the usage of distillation, the distillation-related studies are performed before evolution-based ones. Firstly, we explore the effectiveness of pyramid and cascaded distillation as presented in Table 4, where the network trained with the guidance of hybrid distillation performs best, proving that both pyramid and cascaded distillation are important for SEN. Next, we discuss which working mode of distillation is more effective. As shown in Table 5, the network guided by the online pyramid and cascaded distillation achieves the best Dice score as SEN can adaptively accelerate the learning process of new evolved layers or sub-networks while adjusting the parameters of trained ones in this way. We also investigate the impact of evolution-related designs on SEN. As

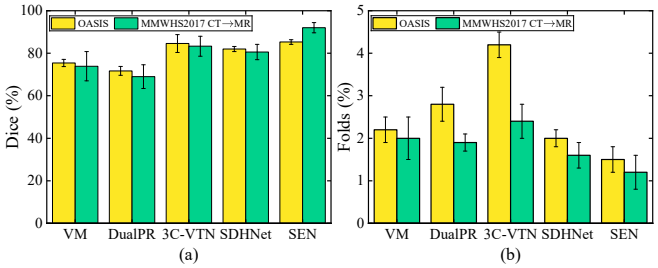

Figure 10. Barplots of various methods with the semi-supervised training in terms of Dice (a) and Folds (b) on the OASIS and MMWHS2017 CT→MR datasets.

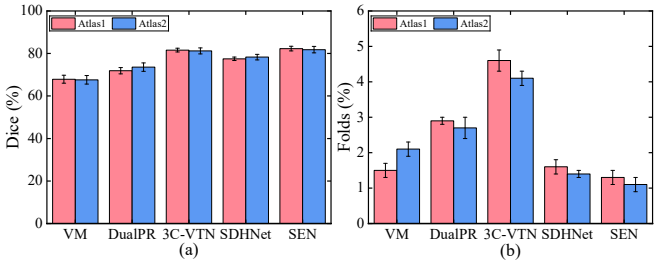

Figure 11. Barplots of the atlas-based registration by various methods in terms of Dice (a) and Folds (b) on the OASIS dataset.

depicted in Table 6, with more evolution, SEN can better handle the multi-scale deformations and large deformation decomposition, thereby improving the registration accuracy. Besides, different evolutionary paths play a vital role in the success of SEN as shown in Table 7, where SEN trained from the pyramid-cascaded evolution performs better than that trained with the cascaded-pyramid evolution, proving that learning the depth-wise knowledge rather than the broad-wise one can better boost the optimization of SEN. Finally, we explore the influence of different loss weights on the performance of SEN in Table 8, where the network trained with $\lambda_{pyd} = 0.001, \lambda_{cas} = 0.001$ performs best, verifying that the proposed pyramid and cascaded evolution and distillation are equally important for SEN.

6.3. The Variability and Robustness of SEN

Apart from SEN and its diffeomorphic version SEN-diff, we investigate more variants of SEN by using the segmentation label as an assistance to train SEN in a semi-supervised manner. For fair comparison, we follow the same settings when implementing other baseline methods. As shown in Fig. 10, SEN outperforms all baseline methods in terms of the Dice score and Folds rate on both OASIS and MMWHS2017 CT→MR datasets, which proves that the proposed self-distilled evolution-based learning strategy can better leverage the knowledge contained in the segmentation label. Based on the joint analysis of results from Tables 1, 3, and Fig. 10, it can be seen that the segmentation label can help to improve the registration performance of SEN (Dice: 84.0%/78.3%→85.3%/92.0%). It should be noticed that SEN trained in an unsupervised way is our major concern as the segmentation labels related to the semi-supervised scheme are not always available.

Moreover, we conduct the atlas-based registration on the OASIS dataset to further investigate the robustness of SEN when facing different template images. Concretely, we choose the first and second scans as the Atlas1 and Atlas2, respectively, and use all other images to match them. As shown in Fig. 11, our proposed SEN invariably performs well no matter how the atlas changes as the proposed learning strategy can adaptively learn the knowledge contained in different atlases. Meanwhile, SEN outperforms all baseline methods on two atlases. These results demonstrate that our SEN is robust enough to match different templates, which is advantageous in the clinical scenario.

6.4. Limitations and Future Work

Still, this work has several limitations and it can be further improved in the future. On the one hand, the designed symmetric UNet is a fully convolutional network whose feature extraction ability is influenced by its inability to capture long-range dependencies. Finding new light-weighted and powerful modules to enhance this network can be a good choice. On the other hand, the knowledge distillation in the existing SEN is implemented in the feature-based and probability-based way, more types of knowledge should be taken into consideration. For example, the relation-based knowledge can be explored for enhancing the registration performance of SEN.

7. Conclusion

In this paper, we propose a novel self-distillation based learning strategy to achieve fast and accurate deformable image registration via pyramid and cascaded progressive learning. We firstly build a vanilla UNet as the base network of SEN and update it step-by-step. On the one hand, we progressively train layers of different levels within each sub-network following the idea of depth-wise pyramid weight-inherited evolution and narrow the learning discrepancy between new layers and old ones via corresponding feature-shifted distillation. On the other hand, we progressively append more sub-networks to form the integrated model via the broad-wise cascaded evolution and utilize the knowledge of trained networks to guide the new one based on the distribution-consistent distillation. Through these strategies, all layers and sub-networks of SEN can be well optimized, leading to its superior performance in both mono-modal and multi-modal datasets. Extensive experiments demonstrate that SEN outperforms various baseline methods with higher registration accuracy and it has higher implementation efficiency than existing module-based and structure-based registration methods.

References

[1] H. Xie, Y. Zhang, J. Qiu, X. Zhai, X. Liu, Y. Yang, et al., Semantics lead all: Towards unified image registration and fusion from a semantic perspective, Information Fusion 98 (2023) 101835.

[2] J. Du, W. Li, K. Lu, B. Xiao, An overview of multi-modal medical image fusion, Neurocomputing 215 (2016) 3–20.

[3] T. C. Mok, A. C. Chung, Fast symmetric diffeomorphic image registration with convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4643–4652.

[4] F. Alam, S. U. Rahman, S. Ullah, K. Gulati, Medical image registration in image guided surgery: Issues, challenges and research opportunities, Biocybernetics and Biomedical Engineering 38 (2018) 71–89.

[5] S. Klein, M. Staring, K. Murphy, M. A. Viergever, J. P. W. Pluim, Elastix: A toolbox for intensity-based medical image registration, IEEE Transactions on Medical Imaging 29 (2009) 196–205.

[6] D. Rueckert, L. Sonoda, C. Hayes, D. Hill, M. Leach, D. Hawkes, Non-rigid registration using free-form deformations: Application to breast mr images, IEEE Transactions on Medical Imaging 18 (1999) 712–721.

[7] J.-P. Thirion, Image matching as a diffusion process: An analogy with maxwell's demons, Medical Image Analysis 2 (1998) 243–260.

[8] D. Rueckert, P. Aljabar, R. A. Heckemann, J. V. Hajnal, A. Hammers, Diffeomorphic registration using b-splines, in: Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI), volume 4191, 2006, pp. 702–709.

[9] T. Vercauteren, X. Pennec, A. Perchant, N. Ayache, Diffeomorphic demons: Efficient non-parametric image registration, NeuroImage 45 (2009) S61–S72.

[10] J. Glaunes, A. Qiu, M. I. Miller, L. Younes, Large deformation diffeomorphic metric curve mapping, International Journal of Computer Vision 80 (2008) 317–336.

[11] B. Avants, C. Epstein, M. Grossman, J. Gee, Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain, Medical Image Analysis 12 (2008) 26–41.

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

[13] J. Fan, X. Cao, P.-T. Yap, D. Shen, Birnet: Brain image registration using dual-supervised fully convolutional networks, Medical Image Analysis 54 (2019) 193–206.

[14] M.-M. Rohe, M. Datar, T. Heimann, M. Sermesant, X. Pennec, Svf-net: Learning deformable image registration using shape matching, in: Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI), volume 10433, 2017, pp. 266–274.

[15] K. A. J. Eppenhof, J. P. W. Pluim, Pulmonary CT registration through supervised learning with convolutional neural networks, IEEE Transactions on Medical Imaging 38 (2019) 1097–1105.

[16] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI), volume 9351, 2015, pp. 234–241.

[17] G. Balakrishnan, A. Zhao, M. R. Sabuncu, A. V. Dalca, J. Guttag, An unsupervised learning model for deformable medical image registration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9252–9260.

[18] A. V. Dalca, G. Balakrishnan, J. Guttag, M. R. Sabuncu, Unsupervised learning for fast probabilistic diffeomorphic registration, in: Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2018, pp. 729–738.

[19] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, A. V. Dalca, Voxelmorph: A learning framework for deformable medical image registration, IEEE Transactions on Medical Imaging 38 (2019) 1788–1800.

[20] A. V. Dalca, G. Balakrishnan, J. Guttag, M. R. Sabuncu, Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces, Medical Image Analysis 57 (2019) 226–236.

[21] D. Kuang, T. Schmah, Faim – a convnet method for unsupervised 3d medical image registration, in: H.-I. Suk, M. Liu, P. Yan, C. Lian (Eds.), Proceedings of Machine Learning in Medical Imaging (MLMI), volume 11861, 2019, pp. 646–654.

[22] J. Chen, E. C. Frey, Y. He, W. P. Segars, Y. Li, Y. Du, Transmorph: Transformer for unsupervised medical image registration, Medical Image Analysis 82 (2022) 102615.

[23] B. Kim, I. Han, J. C. Ye, Diffusemorph: Unsupervised deformable image registration using diffusion model, in: Proceedings of the European Conference on Computer Vision (ECCV), volume 13691, 2022, pp. 347–364.

[24] X. Hu, M. Kang, W. Huang, M. R. Scott, R. Wiest, M. Reyes, Dual-stream pyramid registration network, in: Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI), volume 11765, 2019, pp. 382–390.

[25] T. C. W. Mok, A. C. S. Chung, Large deformation diffeomorphic image registration with laplacian pyramid networks, in: Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2020, pp. 211–221.

[26] M. Kang, X. Hu, W. Huang, M. R. Scott, M. Reyes, Dual-stream pyramid registration network, Medical Image Analysis 78 (2022) 102379.

[27] J. Lv, Z. Wang, H. Shi, H. Zhang, S. Wang, Y. Wang, et al., Joint progressive and coarse-to-fine registration of brain mri via deformation field integration and non-rigid feature fusion, IEEE Transactions on Medical Imaging 41 (2022) 2788–2802.

[28] B. D. De Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, I. Ivsgum, A deep learning framework for unsupervised affine and deformable image registration, Medical Image Analysis 52 (2019) 128–143.

[29] S. Zhao, Y. Dong, E. Chang, Y. Xu, Recursive cascaded networks for unsupervised medical image registration, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10599–10609.

[30] B. Hu, S. Zhou, Z. Xiong, F. Wu, Recursive decomposition network for deformable image registration, IEEE Journal of Biomedical and Health Informatics 26 (2022) 5130–5141.

[31] T. Che, X. Wang, K. Zhao, Y. Zhao, D. Zeng, Q. Li, et al., Amnet: Adaptive multi-level network for deformable registration of 3d brain mr images, Medical Image Analysis 85 (2023) 102740.

[32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al., Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9992–10002.

[33] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), volume 33, 2020, pp. 6840–6851.

[34] S. Zhao, T. Lau, J. Luo, E. I.-C. Chang, Y. Xu, Unsupervised 3d end-to-end medical image registration with volume tweening network, IEEE Journal of Biomedical and Health Informatics 24 (2019) 1394–1404.

[35] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2015, pp. 1–9.

[36] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, L.-J. Li, Learning from noisy labels with distillation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2017, pp. 1928–1936.

[37] J. Yim, D. Joo, J. Bae, J. Kim, A gift from knowledge distillation: Fast optimization, network minimization and transfer learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7130–7138.

[38] Y. Liu, J. Cao, B. Li, C. Yuan, W. Hu, Y. Li, et al., Knowledge distillation via instance relationship graph, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7096–7104.

[39] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, Y. Yan, Knowledge adaptation for efficient semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 578–587.

[40] Y. Hou, Z. Ma, C. Liu, T.-W. Hui, C. C. Loy, Inter-region affinity distillation for road marking segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12486–12495.

[41] S. An, Q. Liao, Z. Lu, J.-H. Xue, Efficient semantic segmentation via self-attention and self-distillation, IEEE Transactions on Intelligent Transportation Systems 23 (2022) 15256–15266.

[42] G. Chen, W. Choi, X. Yu, T. Han, M. Chandraker, Learning efficient object detection models with knowledge distillation, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2017, pp. 1–10.

[43] Q. Li, S. Jin, J. Yan, Mimicking very efficient network for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6356–6364.

[44] Q. Dou, Q. Liu, P. A. Heng, B. Glocker, Unpaired multi-modal segmentation via knowledge distillation, IEEE Transactions on Medical Imaging 39 (2020) 2415–2425.

[45] K. Li, L. Yu, S. Wang, P.-A. Heng, Towards cross-modality medical image segmentation with online mutual knowledge distillation, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), volume 34, 2020, pp. 775–783.

[46] G. Zhang, X. Shen, Y.-D. Zhang, Y. Luo, J. Luo, D. Zhu, et al., Cross-modal prostate cancer segmentation via self-attention distillation, IEEE Journal of Biomedical and Health Informatics 26 (2022) 5298–5309.

[47] B. Hu, S. Zhou, Z. Xiong, F. Wu, Cross-resolution distillation for efficient 3d medical image registration, IEEE Transactions on Circuits and

Systems for Video Technology 32 (2022) 7269–7283.
[48] M. Q. Tran, T. Do, H. Tran, E. Tjiputra, Q. D. Tran, A. Nguyen, Lightweight deformable registration using adversarial learning with distilling knowledge, IEEE Transactions on Medical Imaging 41 (2022) 1443–1453.
[49] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, K. Ma, Be your own teacher: Improve the performance of convolutional neural networks via self distillation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3712–3721.
[50] C. Yang, L. Xie, C. Su, A. L. Yuille, Snapshot distillation: Teacher-student optimization in one generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2854–2863.
[51] X. Xing, Z. Chen, Y. Hou, Y. Yuan, Gradient modulated contrastive distillation of low-rank multi-modal knowledge for disease diagnosis, Medical Image Analysis 88 (2023) 102874.
[52] Y. Hou, Z. Ma, C. Liu, C. C. Loy, Learning lightweight lane detection cnns by self attention distillation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1013–1021.
[53] Z.-L. Ni, X.-H. Zhou, G.-A. Wang, W.-Q. Yue, Z. Li, G.-B. Bian, et al., Surginet: Pyramid attention aggregation and class-wise self-distillation for surgical instrument segmentation, Medical Image Analysis 76 (2022) 102310.
[54] N. Wang, S. Lin, X. Li, K. Li, Y. Shen, Y. Gao, et al., Missu: 3d medical image segmentation via self-distilling transunet, IEEE Transactions on Medical Imaging (2023) 2740–2750.
[55] S. Zhou, B. Hu, Z. Xiong, F. Wu, Self-distilled hierarchical network for unsupervised deformable image registration, IEEE Transactions on Medical Imaging 42 (2023) 2162–2175.
[56] V. Arsigny, O. Commowick, X. Pennec, N. Ayache, A log-euclidean framework for statistics on diffeomorphisms, in: Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI), volume 4190, 2006, pp. 924–931.
[57] D. W. Shattuck, M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K. L. Narr, et al., Construction of a 3d probabilistic atlas of human cortical structures, NeuroImage 39 (2008) 1064–1080.
[58] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, R. 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 (2007) 1498–1507.
[59] T. Heimann, B. Van Ginneken, M. A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, et al., Comparison and evaluation of methods for liver segmentation from ct datasets, IEEE Transactions on Medical Imaging 28 (2009) 1251–1265.
[60] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, et al., The medical segmentation decathlon, Nature Communications 13 (2022) 4128.
[61] X. Zhuang, Challenges and methodologies of fully automatic whole heart segmentation: A review, Journal of Healthcare Engineering 4 (2013) 371–408.
[62] K. Pearson, Vii. Note on regression and inheritance in the case of two parents, Proceedings of the Royal Society of London 58 (1895) 240–242.
[63] D. P. Huttenlocher, G. A. Klanderman, W. J. Rucklidge, Comparing images using the hausdorff distance, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (1993) 850–863.
[64] M. Modat, D. M. Cash, P. Daga, G. P. Winston, J. S. Duncan, S. Ourselin, Global image registration using a symmetric block-matching approach, Journal of Medical Imaging 1 (2014) 024003.