LT-Net: Label Transfer by Learning Reversible Voxel-wise Correspondence for One-shot Medical Image Segmentation

Shuxin Wang* [1,2], Shilei Cao* [2], Dong Wei* [2], Renzhen Wang[4], Kai Ma[2],
Liansheng Wang† [1], Deyu Meng[3,4] and Yefeng Zheng† [2]

[1] Xiamen University [2] Jarvis Lab, Tencent
[3] Macau University of Science and Technology [4] Xi'an Jiaotong University
sxwang@stu.xmu.edu.cn, lswang@xmu.edu.cn, {eliasslcao, donwei, kylekma, yefengzheng}@tencent.com,
wrzhen@stu.xjtu.edu.cn,dymeng@mail.xjtu.edu.cn

Abstract

We introduce a one-shot segmentation method to alleviate the burden of manual annotation for medical images. The main idea is to treat one-shot segmentation as a classical atlas-based segmentation problem,[1] where voxel-wise correspondence from the atlas to the unlabelled data is learned. Subsequently, segmentation label of the atlas can be transferred to the unlabelled data with the learned correspondence. However, since ground truth correspondence between images is usually unavailable, the learning system must be well-supervised to avoid mode collapse and convergence failure. To overcome this difficulty, we resort to the forward-backward consistency, which is widely used in correspondence problems, and additionally learn the backward correspondences from the warped atlases back to the original atlas. This cycle-correspondence learning design enables a variety of extra, cycle-consistency-based supervision signals to make the training process stable, while also boost the performance. We demonstrate the superiority of our method over both deep learning-based one-shot segmentation methods and a classical multi-atlas segmentation method via thorough experiments.

1. Introduction

Precise segmentation of medical images delineates different anatomical structures and abnormal tissues throughout the body, which can be utilized for clinical diagnosis, treatment planning, *etc*. With sufficient well-annotated data, deep convolutional neural networks (DCNNs) achieved ground-breaking performance in such segmentation tasks [34, 31, 19]. However, obtaining 3D annotations of medical images for fully supervised training of DCNNs is labor-intensive and error-prone. Therefore, DCNN-based segmentation methods that require only one or few examples of annotation for training are highly desirable to enable efficient development and deployment of practical solutions.

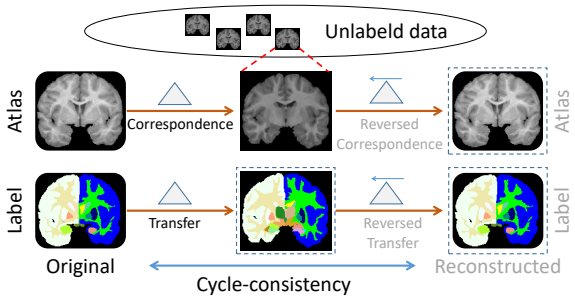

Figure 1: We tackle the one-shot medical image segmentation problem by resorting to the concept of classical atlas-based segmentation, where voxel-wise correspondence from the atlas to the unlabelled data is learned. In addition, we innovatively learn the backward correspondences from the *warped* atlases back to the original atlas to provide cycle-consistency-based supervision.

The lack of large-scale real-world annotations is a long-standing problem in medical image segmentation. Before the era of deep learning, a large body of literatures on medical image segmentation focused on the *atlas*-based segmentation [6, 33, 27, 28, 41, 7]. The key idea is that one or more labelled reference volumes (i.e., atlas) are non-rigidly registered [6, 33] to a target volume, or provided to learn patch-wise corresponding relationship [41, 7] with the target volume, and then the labels of the atlases are propagated to the

*Equal contributions. Shuxin Wang contributed to this work during an internship at Tencent.

†Corresponding authors.

[1]We follow Zhao *et al.* [44] to use the phrase "one-shot" for its generalized meaning that only one exemplar annotation is needed for the proposed model to learn to segment medical images.

target volume as segmentation. An intriguing characteristic of the atlas-based methods is that they only need one or several annotated data, naturally matching the recently rising concept of few-shot learning in deep learning. Classical state-of-the-art (SOTA) atlas-based segmentation methods [41, 7] rely on abundant texture features of local descriptors. Powered by the convolutional operations repeatedly conducted in local regions, DCNNs are especially good at extracting multi-scale local semantic features. Therefore, it is intuitive and appealing to apply DCNNs to develop advanced atlas-based methods for medical image segmentation.

Recent studies [44, 26, 13, 42, 10, 43, 38] showed that the principle of classical atlas-based segmentation could be implemented with DCNNs, and decent performance was achieved. Among all those works, two of them are specifically related to ours in regard to one-shot learning [44, 42]. In the first work, Zhao *et al.* [44] proposed to learn a set of spatial and appearance transformations from the atlas to unlabelled images. By applying randomly combined spatial and appearance transformations to different unlabelled images, the model could synthesize a diverse set of labelled data. In this sense, it provided extra labelled data for training of the segmenter. One of the limitations of this work is that the segmentation accuracy was indirectly boosted by data augmentation, resulting in extra overhead for training the networks responsible for learning both kinds of transformations. The second work [42] proposed a framework that jointly trained two networks for image registration and segmentation, assuming that these two tasks would help each other since they were highly related. However, in many clinical scenarios, registration is not required when segmentation is demanded.

Different from these two works, we propose to *directly* imitate the classical atlas-based segmentation with a deep learning framework, which takes both the atlas and the target image as input, and predicts the correspondence map from the former to the latter. In this way, the segmentation label can be transferred from the atlas to the unlabelled target image with the predicted correspondence. For efficient learning of the correspondence, we enhance the backbone of VoxelMorph [3] via the addition of a discriminator network and adversarial training [17].

Learning correspondence plays an important role in many computer vision tasks, e.g., optical flow [29, 40], tracking [32, 22], patch matching, [2], registration [3, 13, 26, 42], and so on. Among those inspiring works, forward-backward consistency is widely used in the correspondence problem. Specifically, our framework learns not only the forward correspondences from the atlas to unlabelled images, but also the backward correspondences from the *warped* atlases back to the original atlas (see Fig. 1). The introduction of the reverse correspondences naturally complements the full cycle of bidirectional warping, enabling extra, cycle-consistency-based supervision signals to make the learning process with only one annotation more robust and meanwhile preserve the anatomical consistency. In addition, we impose supervision in three involved spaces, namely, the image space, the transformation space, and the label space, which has been verified effective in our experiments.

In summary, we propose a label transfer network (LT-Net) to propagate the segmentation map from the atlas to unlabelled images by learning the reversible voxel-wise correspondences. Our main contributions are as follows:

- To deal with the lack of annotations, our method addresses the one-shot segmentation problem by resorting to the idea of classical atlas-based segmentation. Powered by the representation ability of the DCNN, the proposed method boosts the performance of image matching in feature space, providing anatomically meaningful correspondence for the label transfer.
- We extend correspondence learning to our one-shot segmentation framework in an end-to-end manner, where forward-backward cycle-consistency takes an important role to provide extra supervision in the image, transformation, and label spaces.
- We demonstrate the superiority of our method over both deep learning-based one-shot segmentation methods [44, 3] and a classical multi-atlas segmentation method [21] in segmenting 28 anatomical structures from a brain magnetic resonance imaging (MRI) dataset. We also demonstrate the benefits of the cycle-consistency supervision in each individual space via ablation studies.

2. Related Work

One-shot learning: Early works about one-shot learning mainly focused on image classification [14, 15] based on the assumption that previously learned categories could be leveraged to help forecast a new category when very few examples are available. Along the years, this concept has been used in various branches of machine learning and computer vision problems, such as imitation learning [12, 16], object segmentation [36, 4, 30, 35], neural architecture search [45, 11], and so on. Most recently, Zhao *et al.* [44] developed a one-shot medical image segmentation framework based on data augmentation using learned transformations from the reference atlas to unlabelled images. Specifically, both the spatial and appearance transformation models were learned and then utilized to synthesize additional labelled samples for data augmentation. Our work also explores the one-shot setting for medical image segmentation to alleviate the burden of manual annotation. The

main difference is that we directly target the segmentation in our network design, and incorporate the forward-backward consistency in the framework to ensure abundant supervision for learning.

Atlas-based segmentation: Atlas-based segmentation is a classical topic in medical image analysis, evolved from single atlas-based [33, 44, 26, 13, 42, 10] to sophisticated, multi-atlas-based methods [25, 18, 41, 10, 43, 38, 21]. Recently, motivated by the success of DCNNs, researchers revitalized this classical concept with deep learning models. Using a single atlas, researchers explored this methodology in three ways: learning transformations for data augmentation [44], combining with another registration task [26, 13, 42], and learning a deformation field to resample an initial binary mask [10]. Whereas for multiple atlases, recent works attempted to implement key components of multi-atlas segmentation with DCNNs, e.g., atlas selection [43], label propagation [38], and label fusion [43, 9].

Our work approaches the one-shot medical image segmentation problem via single atlas-based segmentation with a correspondence-learning generative adversarial network (GAN) framework. It falls into the single-atlas category, which offers two advantages. First, for complex organs, e.g., the brain, to annotate extra few samples in detail can be a considerable burden. Second, there is no need to consider the intricate processes involved in the multi-atlas approach, such as label fusion or atlas selection. In spite of relying on a single atlas, our proposed framework outperforms an advanced multi-atlas method [21] using up to five atlases (cf. Section 6.4).

Correspondence in computer vision: Correspondence plays an important role in computer vision. Actually, many fundamental vision problems, from optical flow [29, 40] and tracking [32, 22] to patch matching [2] and registration [3, 13, 26, 42], require some notion of visual correspondence [39]. Optical flow and registration can be seen as pixel/voxel-level correspondence problems, whereas tracking and patch matching can be seen as patch-level correspondence problems. By treating atlas-based segmentation as a correspondence problem, we draw lessons from these research areas to guide the design of our framework.

Forward-backward consistency: Forward-backward consistency has been widely adopted in many computer vision problems, especially in the correspondence learning problem. For example, forward-backward consistency has been the evaluation metric [22] as well as the measure of uncertainty [32] for tracking. Recent methods on unsupervised optical flow estimation [29, 40] employed forward and backward consistency to define an occluded region, which was excluded for training. Besides, forward-backward consistency is an important building block for CycleGAN, which is the most popular framework for image-to-image translation [46]. To the best of our knowledge, our work is the first to employ cycle-consistency in one-shot atlas-based segmentation within a deep learning framework.

3. Basic Framework for Correspondence Learning

Preliminaries: We first recap the basic concept of atlas-based image segmentation, where the segmentation of an unseen subject can be estimated by a registration process. Let (l, l_s) be a labelled image pair, where $l \in \mathbb{R}^{h\times w\times c}$ is the atlas image, $l_s \in \mathbb{R}^{h\times w\times c}$ is its corresponding segmentation map, and h, w, c are the numbers of voxels along the coronal, sagittal, and axial directions, respectively. In practice, input images are defined within a 3D space $\Omega \in \mathbb{R}^3$, which also applies to the unlabelled image pool $\{u^{(i)}|u^{(i)} \in \mathbb{R}^{h\times w\times c}\}$. In the following, we use u to denote an unlabelled image for an uncluttered notion. Let Δp_F (F standing for forward is used to differentiate the backward operations introduced in Section 4) denote the forward correspondence map that warps l towards u during the registration process. Specifically, Δp_F can be considered as a spatially varying function defined over Ω, that maps coordinates of u to those of l by displacement vectors. We use $l \circ \Delta p_F$ to denote the application of Δp_F on l (i.e., warp l towards u according to Δp_F):

$$\bar{u} = l \circ \Delta p_F, \tag{1}$$

where $\circ$ is a warp operation, and $\bar{u}$ is the deformed atlas. The segmentation map l_s can be warped the same way as the atlas:

$$\bar{u}_s = l_s \circ \Delta p_F, \tag{2}$$

where $\bar{u}_s$ is the synthetic segmentation of u. If Δp_F registers l and u well, $\bar{u}_s$ is expected to be an accurate segmentation of u. In this sense, we treat the atlas-based segmentation as a label transfer process.

Atlas-based segmentation with deep learning: To model the registration function with a DCNN, a generator network G_F is often adopted to match the local spatial information between l and u, and output Δp_F. For example, VoxelMorph [3] used a U-Net [34] as G_F to learn the image correspondence. The parameters of the network are optimized by minimizing two unsupervised loss functions: the image similarity loss $\mathcal{L}_{\text{sim}}(u, \bar{u})$ and the transformation smoothness loss $\mathcal{L}_{\text{smooth}}(\Delta p_F)$. To introduce robustness against global intensity variations in medical images, we use the locally normalized cross-correlation (CC) loss [42, 44] for $\mathcal{L}_{\text{sim}}$, which encourages coherence in local regions. For $\mathcal{L}_{\text{smooth}}$, it is formulated with first-order derivatives of Δp_F:

$$\mathcal{L}_{\text{smooth}}(\Delta p_F) = \sum_{t\in\Omega} \|\nabla(\Delta p_F(t))\|_2, \tag{3}$$

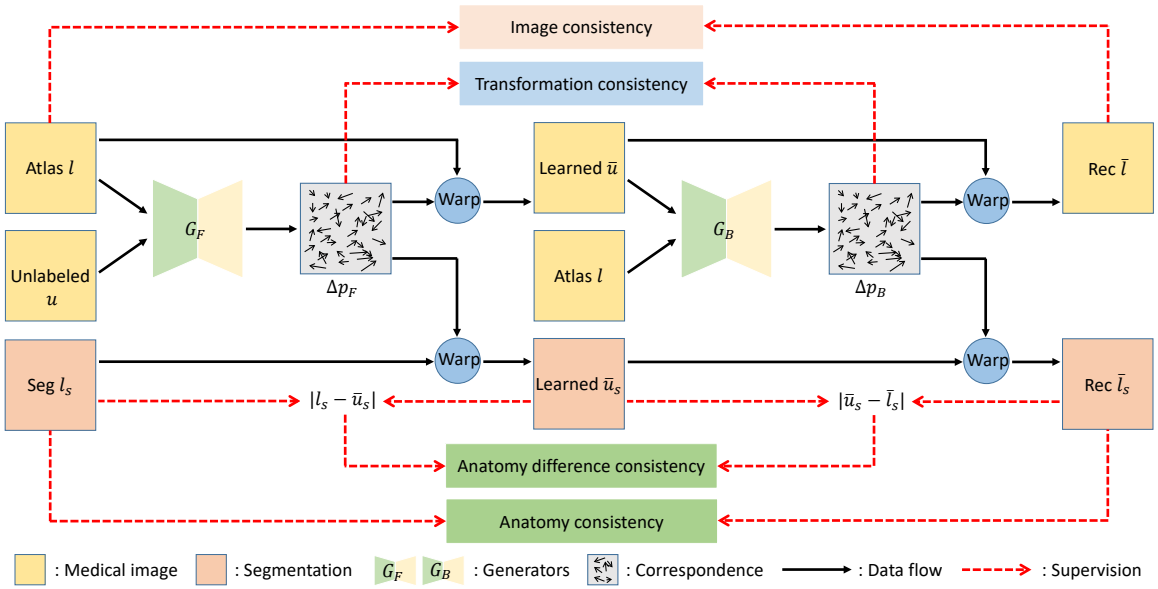

Figure 2: The overview of our proposed label transfer network (LT-Net). We innovatively introduce forward-backward cycle consistency design into the atlas-based segmentation workflow. This facilitates us to explore new supervision signals, which provide more robust driving forces to guide the correspondence learning. Specifically, we adopt cycle-consistency losses in image, transformation, and label spaces and their effectiveness has been verified in the experiments.

where $t \in \Omega$ iterates over all spatial locations in Δp_F, and we approximate $\|\nabla(\Delta p(t))\|_2$ with spatial gradient differences between neighboring voxels along x, y, z directions [3]. Minimizing $\mathcal{L}_{\text{sim}}$ encourages $\bar{u}$ to approximate u, whereas minimizing $\mathcal{L}_{\text{smooth}}$ regularizes Δp_F to be smooth. In addition, smoothness regularization can be considered as a strategy to alleviate the overfitting problem while encoding the anatomical priori.

Auxiliary GAN loss: Besides the two basic losses used by VoxelMorph, we introduce a GAN [17] into our basic framework to offer additional supervision. The GAN subnet in our framework comprises G_F and a discriminator network D. A vanilla GAN would make the discriminator D differentiate Δp_F from the true underlying correspondence map. In practice, however, it is usually infeasible to obtain the true correspondence between a pair of clinical images. Instead, we make D distinguish the synthetic image $\bar{u}$ from u. In this sense, $\bar{u}$ serves as a delegate of Δp_F, and G_F is trained to generate Δp_F that can be used to synthesize $\bar{u}$ authentically enough to confuse D; meanwhile, D becomes more skilled at flagging synthesized images. This delegation strategy provides indirect supervision to G_F and Δp_F, and allows the networks to be trained end-to-end with a large number of unlabelled images. Consequently, the image adversarial loss $\mathcal{L}_{\text{GAN}}$ is defined as

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(l, u, \bar{u}) = \mathbb{E}_{u\sim p_d(u)}[\|D(u)\|_2] \\ + \mathbb{E}_{l\sim p_d(l), u\sim p_d(u)}[\|D(\bar{u}) - \mathbf{1}\|_2],\end{aligned} \tag{4}$$

where G_F and D are trained alternatively to compete in a two-player min-max game with the objective function $\min_{G_F} \max_D \mathcal{L}_{\text{GAN}}(G_F, D)$. [2]

4. Learning Reversible Cycle Correspondence

In the previous section, we introduce our baseline method for atlas-based one-shot segmentation. Our proposed framework is built on this baseline and adds a cycle consistency constraint to further boost the segmentation performance. We name the proposed framework as label transfer network (LT-Net). Unlike previous works [44, 42] that only learned the forward correspondences from the atlas to unlabelled images, we in addition learn the backward correspondences from the *warped* atlases back to the original atlas. As far as the authors are aware of, this work is the first attempt that utilizes the cycle correspondence for one-shot (atlas-based) segmentation with deep learning. Specifically, we propose a backward correspondence learning path: $\Delta p_B = G_B(\bar{u}, l)$, where Δp_B is the backward correspondence map, and G_B is the backward generator (see Δp_B and G_B in Fig. 2). With the newly added Δp_B, we can *revert* the synthetic image $\bar{u}$ to reconstruct the atlas using the warp operation:

$$\bar{l} = \bar{u} \circ \Delta p_B, \tag{5}$$

and we call $\bar{l}$ the reconstructed atlas. Accompanied with the backward learning path, a straightforward addition to the

[2] Our basic framework for correspondence learning is illustrated and explained in more details in the supplementary material.

network's supervision is to impose transformation smoothness loss on Δp_B as well. Hence, our complete transformation smoothness loss becomes

$$\mathcal{L}_{\text{smooth}} = \mathcal{L}_{\text{smooth}}(\Delta p_F) + \mathcal{L}_{\text{smooth}}(\Delta p_B). \quad (6)$$

More importantly, the completion of the correspondence cycle enables a variety of supervision signals to boost the performance upon unidirectional correspondence learning. Concretely, we propose four novel, cycle-consistency-driven supervision losses (cf. the supervision blocks in Fig. 2) in three spaces, namely, the image space, the transformation space, and the label space. These supervision losses are all devised by straightforward intuitions, as described below.

- In the image space, the reconstructed and original atlas images ($\bar{l}$ and l) should be the same (the *image consistency*).
- In the transformation space, conceptually the forward and backward warpings should be the inverse function of each other, so that the atlas warped toward the unlabelled image can be warped back to what it originally is (the *transformation consistency*).
- Lastly, in the label space, the true segmentation l_s and the reconstructed segmentation $\bar{l}_s$ should be the same (the *anatomy consistency*). In addition, they must differ from the synthetic segmentation $\bar{u}_s$ in the same way (the *anatomy difference consistency*).

Despite being conceptually simple, the comprehensive inclusion and combination of these supervision signals in our framework are proved to be effective in the experiments—our LT-Net outperforms the current SOTA by significant margins, and the ablation studies demonstrate benefits of the supervision in individual spaces. In the following, we describe each loss in detail.

Cycle-consistency supervision in image space: Enabled by our novel forward-backward cycle correspondence learning framework, we can revert the synthetic image $\bar{u}$ to reconstruct the atlas. We employ an L1 loss to enforce the consistency between the true atlas and the reconstructed one, which is defined as

$$\mathcal{L}_{\text{cyc}}(l, \bar{l}) = \mathbb{E}_{l\sim p_d(l)}[\|\bar{l} - l\|_1]. \quad (7)$$

Cycle-consistency supervision in transformation space: In terms of forward-backward consistency, the correspondences should be reversible, meaning that a voxel warped from one position to another in the forward path should be warped back to its original position in the backward path. Therefore, we define a transformation consistency loss to enforce this constraint as

$$\mathcal{L}_{\text{trans}}(\Delta p_F, \Delta p_B) = \sum_{t\in\Omega} \rho(\Delta p_F(t) + \Delta p_B(t + \Delta p_F(t))), \quad (8)$$

where $\rho(x) = (x^2 + \epsilon^2)^\gamma$ is a robust generalized Charbonnier penalty function [37] and widely used as a photometric loss in optical flow estimation [29, 40]. In this work, we use the same setting with $\epsilon = 0.001, \gamma = 0.45$ as [29].

Cycle-consistency supervision in label space: In many applications, matching the images solely based on intensity is under-constrained and may lead to wrong correspondences. The corresponding anatomical structure may shift or twist away from one position to another, as long as the warped and target images appear similar. Enforcing smoothness constraint on the correspondence map (as in VoxelMorph [3]) is a common way of alleviating this problem. In this work, we further explore driving forces in the label space to guide the correspondence learning towards an anatomically meaningful direction.

When considering supervision signal in the label space, an anatomy cycle-consistency constraint naturally comes up within our framework. Let $\bar{l}_s = \bar{u}_s \circ \Delta p_B$ denote the reconstructed segmentation map of $\bar{l}$. To model the dissimilarity between $\bar{l}_s$ and the original segmentation map l_s, a Dice loss [31] is adopted which is defined as

$$\mathcal{L}_{\text{anatomy\_cyc}}(l_s, \bar{l}_s) = 1 - \frac{2\sum_{t\in\Omega} l_s(t)\bar{l}_s(t)}{\sum_{t\in\Omega} l_s^2(t) + \sum_{t\in\Omega} \bar{l}_s^2(t)}. \quad (9)$$

Since our target is to learn the correspondence which can be used to transfer the segmentation map of the atlas to each of the unlabelled images, we also propose an anatomy difference consistency loss to indirectly regularize quality of the synthetic segmentation map $\bar{u}_s$. As aforementioned, this loss is based on a simple intuition: the anatomy differences between the atlas and the unlabelled image in the forward and backward paths should be cyclically consistent in the label space. The loss is thus formulated as

$$\mathcal{L}_{\text{diff\_cyc}}(l_s, \bar{u}_s, \bar{l}_s) = \sum_{t\in\Omega} \rho(|l_s(t) - \bar{u}_s(t)| - |\bar{u}_s(t) - \bar{l}_s(t)|). \quad (10)$$

5. Optimization Objective and Implementation

Given the definitions of the supervision signals above, our complete objective for optimization is defined as

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{\text{GAN}} + \mathcal{L}_{\text{sim}} + \lambda_1\mathcal{L}_{\text{cyc}} + \lambda_2(\mathcal{L}_{\text{anatomy\_cyc}} \\ &+ \mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{trans}} + \mathcal{L}_{\text{diff\_cyc}}),\end{aligned} \quad (11)$$

where λ_1 and λ_2 are the weights to balance the importance of the different losses. We use the same weight for the last four losses in Eq. (11), since they are comparable in magnitude and we find the results insensitive to their relative weights in our primitive experiments. We set $\lambda_1 = 10$ following CycleGAN [46], and consequently set $\lambda_2 = 3$

to make the corresponding loss values at the same level as $\mathcal{L}_{\text{cyc}}$. The supervision signals in the image, transformation, and label spaces affect each other and restrict each other, pushing the learning system towards an anatomically meaningful direction.

We implement all models using Keras [5] with a TensorFlow [1] backend. For the generator networks in both the forward and backward paths, we adopt the same 3D U-Net architecture as VoxelMorph [3] for a fair comparison later. For the discriminator network, we use an extended 3D version of PatchGAN [20] to determine whether an image patch is real or synthesized. All networks are optimized from scratch using the Adam solver [24]. The learning rate is initialized to 0.0002 and remains unchanged during the training process. Each mini-batch processes a pair of volumes (one atlas and one unlabelled image) per GPU while running two Tesla P40 GPUs in parallel. During testing, the forward correspondence map Δp_F from the atlas to a test unlabelled image $u^{(i)}$ is predicted by G_F, then the segmentation map for $u^{(i)}$ is produced with Eq. (2).

6. Experiments

We demonstrate the superiority of our LT-Net on the task of brain MRI segmentation. Above all, the effectiveness of the cycle correspondence learning framework is evaluated (Section 6.2). As aforementioned, the forward-backward consistency is a classical constraint in correspondence problems. By introducing a backward correspondence path, extra meaningful supervision signals can be exploited to drive the learning process towards a more robust and anatomically meaningful direction. Hence, within the cycle correspondence framework, we subsequently examine the effects of the several newly proposed cycle consistency losses with ablation studies: the transformation consistency loss in the transformation space, the anatomy consistency and difference consistency losses in the label space, and the combination of the losses from both spaces (Section 6.3). Next, we compare our method with a classical multi-atlas method (Section 6.4), demonstrating that the traditional idea of atlas-based segmentation in computer vision can be further boosted using deep learning. Finally, we compare with a SOTA method for one-shot medical image segmentation, demonstrating the superiority of our framework to other DCNN-based methods (Section 6.5). Examples of the synthesized images and warped segmentation maps for unlabelled images are also presented for visual evaluation.

6.1. Dataset and Evaluation Metric

Dataset: We use a publicly available dataset from the Child and Adolescent NeuroDevelopment Initiative (CANDI) at the University of Massachusetts Medical School [23]. The dataset comprises 103 T1-weighted MRI scans (57 males and 46 females) with anatomic segmentation labels. The subjects come from four diagnostic groups: healthy controls, schizophrenia spectrum, bipolar disorder with psychosis, and bipolar disorder without psychosis. We use 28 anatomical structures (tabulated in the supplementary material) that were used in VoxelMorph [3]. The volume size ranges from $256 \times 256 \times 128$ to $256 \times 256 \times 158$ voxels. For computation efficiency, we crop a $160 \times 160 \times 128$ region around the center of the brain, which is large enough to contain the whole brain. We randomly select 20 volumes as test data, and use the others for training. Among the training data, the volume which is most similar to the anatomical average is selected as the only annotated atlas (the same strategy as adopted in VoxelMorph [3]).

Table 1: Mean Dice scores (%) (with standard deviations in parentheses) for VoxelMorph [3], and its extended versions which gradually incorporate the image adversarial loss $\mathcal{L}_{\text{GAN}}$ and cycle-consistency loss $\mathcal{L}_{\text{cyc}}$. Min and Max represent the minimum and maximum Dice scores (%) in the test dataset.

	Mean (std)	Min	Max
VoxelMorph	76.0 (9.7)	61.7	80.1
+ $\mathcal{L}_{\text{GAN}}$	79.0 (3.1)	72.7	81.9
+ $\mathcal{L}_{\text{GAN}}$ + $\mathcal{L}_{\text{cyc}}$	**79.2** (**2.8**)	72.7	82.1

Evaluation metric: We use the Dice similarity coefficient [8] to evaluate the segmentation accuracy of each model, which measures the overlap between manual annotations and predicted results.

6.2. Effectiveness of Forward-backward Consistency

We adapt VoxelMorph [3]—the SOTA DCNN registration model—for our problem setting and use it as the initial performance baseline. Specifically, we train it as a single-atlas model to learn the forward correspondence, and warp the atlas's segmentation map according to the learned correspondence for each unlabelled image. As introduced in Section 3, our basic framework adopts the same backbone as VoxelMorph for forward correspondence learning, but adds a GAN for additional supervision. Then, built on top of the basic framework, our LT-Net introduces a backward correspondence learning path to form a complete cycle correspondence framework. Enabled by the cyclic structure, we further add an image cycle-consistency loss $\mathcal{L}_{\text{cyc}}$ between the atlas and the reconstructed one. For detailed comparisons, we first add $\mathcal{L}_{\text{GAN}}$ alone, and then $\mathcal{L}_{\text{cyc}}$ together.

The quantitative comparison results are shown in Table 1. The results show that 3.0% and 3.2% improvements are achieved when gradually adding $\mathcal{L}_{\text{GAN}}$ and $\mathcal{L}_{\text{cyc}}$. This indicates that $\mathcal{L}_{\text{GAN}}$ can boost the performance for our correspondence learning problem, which is in accordance with

Table 2: Ablation study on the newly proposed supervision signals. We show the mean Dice scores (%) with standard deviations. Besides, Min and Max represent the minimum and maximum Dice scores (%) in the test dataset.

	Mean (std)	Min	Max
Baseline	79.2 (2.8)	72.7	82.1
+ $\mathcal{L}_{\text{trans}}$	80.9 (2.7)	73.6	83.2
+ $\mathcal{L}_{\text{anatomy\_cyc}}$	80.5 (2.5)	74.2	83.1
+ $\mathcal{L}_{\text{trans}}$ + $\mathcal{L}_{\text{anatomy\_cyc}}$	81.4 (2.6)	74.4	83.8
+$\mathcal{L}_{\text{trans}}$ + $\mathcal{L}_{\text{anatomy\_cyc}}$ + $\mathcal{L}_{\text{diff\_cyc}}$	**82.3** (**2.5**)	75.6	84.2

Table 3: Comparison of our LT-Net with a multi-atlas method MABMIS [21] using increasing numbers of atlases. We show the mean Dice scores (%) with standard deviations. Besides, Min and Max represent the minimum and maximum Dice scores (%) in the test dataset.

	No. of Atlases	Mean (std)	Min	Max
MABMIS	2	70.4 (4.4)	63.0	75.9
MABMIS	3	74.5 (5.2)	63.2	80.4
MABMIS	4	77.8 (3.9)	72.0	82.5
MABMIS	5	81.1 (3.6)	76.0	85.6
LT-Net	1	**82.3** (**2.5**)	75.6	84.2

the practical experience that image adversarial losses usually perform well in image-to-image translation tasks. It is worth noting that $\mathcal{L}_{\text{cyc}}$ does not bring substantial further improvement upon the GAN setting. However, as we have mentioned earlier and will experimentally show next, the benefit of the cycle design is that it enables us to incorporate extra supervision signals to the learning framework, which can further improve the performance. Next, we treat the VoxelMorph backbone plus $\mathcal{L}_{\text{GAN}}$ and $\mathcal{L}_{\text{cyc}}$ as a new baseline, and design experiments to examine the effectiveness of the newly proposed supervision signals.

6.3. Ablation Study on the Supervision Signals

Cycle-consistency in transformation space: The correspondence learned from the atlas to the unlabelled image is used to synthesize $\bar{u}$ by warping the atlas in the forward path, whereas in the backward path another correspondence is learned from $\bar{u}$ back to the atlas. The forward and backward correspondences should be cycle-consistent. We conduct an ablation study with respect to the transformation consistency loss $\mathcal{L}_{\text{trans}}$ and show the results in Table 2. From the table, we can observe that $\mathcal{L}_{\text{trans}}$ brings a 1.7% improvement compared to the baseline. This may imply that intensity matching at the image level—despite the cycle correspondence setting—is not enough to prevent the overfitting by DCNNs, and the introduction of supervision in other spaces (e.g., the label space) has the potential for further improvement in performance.

Cycle-consistency in label space: With the forward correspondence, the segmentation map of the atlas can be warped to synthesize the segmentation map for each unlabelled image. Inversely, the synthetic segmentation map can be warped back to restore the segmentation map of the atlas using the backward correspondence. Ideally, the segmentation maps of the atlas before and after the dual warping should be the same, and we enforce this constraint with the anatomy consistency loss $\mathcal{L}_{\text{anatomy\_cyc}}$. Table 2 quantitatively displays the effect of this supervision. We can observe that $\mathcal{L}_{\text{anatomy\_cyc}}$ brings a 1.3% improvement when compared with the baseline, and an extra 0.9% improvement when further combined with the transformation consistency loss. As expected, $\mathcal{L}_{\text{anatomy\_cyc}}$ boosts the performance, since it can ensure the integrity and internal coherence of the anatomical structure.

The anatomy cycle-consistency loss does not consider the middle-cycle synthesized segmentation map $\bar{u}_s$ for each unlabelled image, which, however, is the ultimate goal of our LT-Net. To place more emphasis on $\bar{u}_s$, the anatomy difference consistency loss $\mathcal{L}_{\text{diff\_cyc}}$ is proposed in the label space to regularize the differences between the segmentation maps of the atlas and that of the unlabelled image. The results in Table 2 show that by indirectly regularizing the segmentation maps of the unlabelled images, we achieve a 0.9% further improvement.

6.4. Comparison with a Classical Multi-atlas Method

Traditional multi-atlas methods once achieved SOTA results for atlas-based segmentation. We compare our LT-Net with MABMIS [21], which consists of a tree-based groupwise registration method and an iterative groupwise segmentation method. The results are shown in Table 3. We can observe that our method using only one atlas outperforms MABMIS using up to five atlases. In addition, classical multi-atlas segmentation is notorious for being time-consuming. While MABMIS requires $\sim$14 minutes to segment one case with an Intel® Core i3-4150 CPU (using two atlases), our LT-Net only needs $\sim$4 seconds with a single Tesla P40 GPU.

6.5. Comparison with SOTA Methods

Besides VoxelMorph, we also compare our proposed LT-Net with DataAug [44], a SOTA method for one-shot medical image segmentation relying on registration-based data augmentation. In addition, we train a fully supervised U-Net [34] using a labelled training pool of 83 subjects, which is served as the upper bound for the one-shot methods. The results are shown in Table 5. Using only one annotated data for training, our framework achieves 95.1% of the upper bound on the mean Dice score, yet with an apparently lower standard deviation. Besides, we can observe that our LT-Net outperforms both VoxelMorph and DataAug by margins of

Table 4: Segmentation accuracy (mean Dice scores, %) of VoxelMorph [3], DataAug [44], U-Net [34] and our proposed LT-Net across various brain structures. Labels consisting of left and right structures are combined (*e.g.*, hippocampus). Abbreviations: white matter (WM), cortex (CX), ventricle (Vent), and cerebrospinal fluid (CSF).

	Cerebral-WM	Cerebral-CX	Lateral-Vent	Cerebellum-WM	Cerebellum-CX	Thalamus-Proper	Caudate	Putamen	Pallidum	3rd-Vent	4th-Vent	Brain-Stem	Hippocampus	Amygdala	CSF	VentralDC
VoxelMorph	81.7	87.1	76.7	74.0	86.3	84.7	79.6	83.3	74.0	62.9	69.8	87.1	59.2	66.5	50.9	76.6
DataAug	90.5	93.3	87.5	82.3	93.2	87.3	81.3	82.8	73.7	66.3	72.2	90.5	72.5	69.2	63.3	80.3
LT-Net	85.8	90.9	83.1	80.0	91.6	87.9	85.5	88.4	80.5	68.4	79.7	92.4	71.6	71.6	67.1	82.3
U-Net (upper bound)	92.0	93.1	91.8	87.9	93.1	90.6	88.1	88.7	82.5	79.0	84.9	92.2	80.3	75.3	69.8	85.9

Table 5: Comparison of our LT-Net with VoxelMorph [3], DataAug [44] and fully supervised U-Net [34]. We show the mean Dice scores (%) with standard deviations. Besides, Min and Max represent the minimum and maximum Dice scores (%) in the test dataset.

	Mean (std)	Min	Max
VoxelMorph	76.0 (9.7)	61.7	80.1
DataAug	80.4 (4.3)	73.8	84.0
LT-Net	82.3 (2.5)	75.6	84.2
U-Net (upper bound)	86.5 (6.3)	83.7	89.2

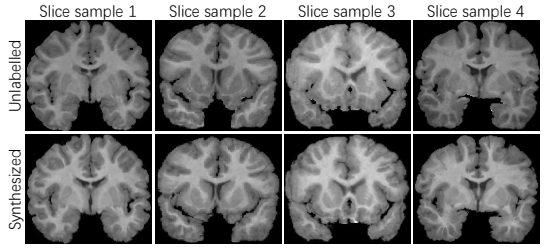

Figure 3: Example coronal MR slices from the synthetic images obtained by warping the atlas with the learned forward correspondences. Each column is a different patient.

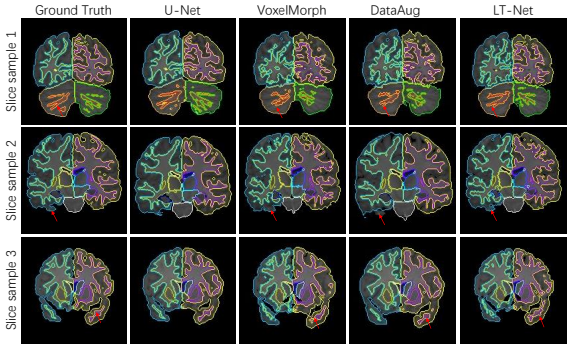

Figure 4: Example coronal MR slices of brain structure annotations and segmentation maps predicted by VoxelMorph [3], DataAug [44], U-Net [34] and our LT-Net. Each row is a different patient. Red arrows point to the flaws (best viewed zoomed-in) in the predictions made by the other one-shot methods, as compared to those by LT-Net.

6.3% and 1.9%, respectively. Table 4 shows the segmentation accuracy across various brain structures.

We visualize some example slices of the synthetic volumes $\bar{u}$ from different patients in Fig. 3. We observe that the synthesized images are close to the unlabelled images in terms of the anatomical structures. In addition, Fig. 4 shows some example slices of brain structure annotations and segmentation maps predicted by U-Net, VoxelMorph, DataAug, and our proposed LT-Net. Compared to the other two one-shot methods, LT-Net predicts brain structures in a way that is more anatomically meaningful.

7. Conclusion

In this study, we traced back to two classical ideas—atlas-based segmentation and correspondence—in computer vision and applied them to one-shot medical image segmentation with DCNNs. First, we bridged the conceptual gap between atlas-based segmentation and the generic idea of one-shot segmentation. This provided us with some critical thinkings for the design of our deep network. Second, we adopted the forward-backward consistency strategy from other correspondence problems, which subsequently enabled the design of a few novel supervision signals in three involved spaces (namely, the image space, the transformation space, and the label space) to make the learning well-supervised and effectively-guided. We hope this work would inspire the future development of one-shot learning for medical image segmentation in the era of deep learning.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 61671399), the Fundamental Research Funds for the Central Universities (Grant No. 20720190012), the Key Area Research and Development Program of Guangdong Province, China (Grant No. 2018B010111001) and the Science and Technology Program of Shenzhen, China (No. ZDSYS201802021814180).

References

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

[2] Christian Bailer, Kiran Varanasi, and Didier Stricker. CNN-based patch matching for optical flow with thresholded hinge embedding loss. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 3250–3259, 2017.

[3] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. VoxelMorph: A learning framework for deformable medical image registration. *IEEE Trans. Medical Imaging*, 2019.

[4] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 221–230, 2017.

[5] François Chollet. Keras. `https://github.com/fchollet/keras`, 2015.

[6] D. Louis Collins, Colin J. Holmes, Terrence M. Peters, and Alan C. Evans. Automatic 3-D model-based neuroanatomical segmentation. *Human Brain Mapping*, 3(3):190–208, 1995.

[7] Pierrick Coupé, José V. Manjón, Vladimir Fonov, Jens Pruessner, Montserrat Robles, and D. Louis Collins. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. *NeuroImage*, 54(2):940–954, 2011.

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

[9] Zhipeng Ding, Xu Han, and Marc Niethammer. VoteNet: A deep learning label fusion method for multi-atlas segmentation. In *Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention*, pages 202–210, 2019.

[10] Nicola K. Dinsdale, Mark Jenkinson, and Ana I. L. Namburete. Spatial warping network for 3D segmentation of the hippocampus in MR images. In *Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention*, pages 284–291, 2019.

[11] Xuanyi Dong and Yi Yang. One-shot neural architecture search via self-evaluated template network. *arXiv preprint arXiv:1910.05733*, 2019.

[12] Yan Duan, Marcin Andrychowicz, Bradly Stadie, OpenAI Jonathan Ho, Jonas Schneider, Ilya Sutskever, Pieter Abbeel, and Wojciech Zaremba. One-shot imitation learning. In *Advances in Neural Information Processing Systems*, pages 1087–1098, 2017.

[13] Mohamed S. Elmahdy, Jelmer M. Wolterink, Hessam Sokooti, Ivana Išgum, and Marius Staring. Adversarial optimization for joint registration and segmentation in prostate CT radiotherapy. In *Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention*, pages 366–374. Springer, 2019.

[14] Li Fei-Fei, Robert Fergus, and Pietro Perona. A Bayesian approach to unsupervised one-shot learning of object categories. In *Proc. Int'l Conf. Computer Vision*, pages 1134–1141. IEEE, 2003.

[15] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Machine Intell.*, 28(4):594–611, 2006.

[16] Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot visual imitation learning via meta-learning. *arXiv preprint arXiv:1709.04905*, 2017.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[18] Rolf A. Heckemann, Joseph V. Hajnal, Paul Aljabar, Daniel Rueckert, and Alexander Hammers. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115–126, 2006.

[19] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of Digital Imaging*, pages 1–15, 2019.

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[21] Hongjun Jia, Pew-Thian Yap, and Dinggang Shen. Iterative multi-atlas-based multi-image segmentation with tree-based registration. *NeuroImage*, 59(1):422–430, 2012.

[22] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Forward-backward error: Automatic detection of tracking failures. In *20th International Conference on Pattern Recognition*, pages 2756–2759. IEEE, 2010.

[23] David N. Kennedy, Christian Haselgrove, Steven M. Hodge, Pallavi S. Rane, Nikos Makris, and Jean A. Frazier. CANDIShare: A resource for pediatric neuroimaging data. *Neuroinformatics*, 10(3):319–322, Oct. 2011.

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

[25] Arno Klein, Brett Mensh, Satrajit Ghosh, Jason Tourville, and Joy Hirsch. Mindboggle: Automated brain labeling with multiple atlases. *BMC Medical Imaging*, 5(1):7, 2005.

[26] Bo Li, Wiro J. Niessen, Stefan Klein, Marius de Groot, M. Arfan Ikram, Meike W. Vernooij, and Esther E. Bron. A hybrid deep learning framework for integrated segmentation and registration: Evaluation on longitudinal white matter tract changes. In *Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention*, pages 645–653. Springer, 2019.

[27] Maria Lorenzo-Valdés, Gerardo I. Sanchez-Ortiz, Raad Mohiaddin, and Daniel Rueckert. Atlas-based segmentation and tracking of 3D cardiac MR images using non-rigid registration. In *Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention*, pages 642–650. Springer, 2002.

[28] Jyrki MP. Lötjönen, Robin Wolz, Juha R. Koikkalainen, Lennart Thurfjell, Gunhild Waldemar, Hilkka Soininen,

Daniel Rueckert, Alzheimer's Disease Neuroimaging Initiative, et al. Fast and robust multi-atlas segmentation of brain magnetic resonance images. *NeuroImage*, 49(3):2352–2365, 2010.

[29] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proc. AAAI Conf. Artificial Intelligence*, pages 7251–7259, 2018.

[30] Claudio Michaelis, Ivan Ustyuzhaninov, Matthias Bethge, and Alexander S. Ecker. One-shot instance segmentation. *CoRR*, abs/1811.11507, 2018.

[31] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[32] Pan Pan, Fatih Porikli, and Dan Schonfeld. Recurrent tracking using multifold consistency. In *Proceedings of the Eleventh IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2009.

[33] Dzung L. Pham, Chenyang Xu, and Jerry L. Prince. Current methods in medical image segmentation. *Annual Review of Biomedical Engineering*, 2(1):315–337, 2000.

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention*, pages 234–241. Springer, 2015.

[35] Benjamin Schnieders, Shan Luo, Gregory Palmer, and Karl Tuyls. Fully convolutional one-shot object segmentation for industrial robotics. *CoRR*, abs/1903.00683, 2019.

[36] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017.

[37] Deqing Sun, Stefan Roth, and Michael J. Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *Int. J. Computer Vision*, 106(2):115–137, 2014.

[38] Maria Vakalopoulou, Guillaume Chassagnon, Norbert Bus, R. Marini, Evangelia I. Zacharaki, Marie-Pierre Revel, and Nikos Paragios. AtlasNet: Multi-atlas non-linear deep networks for medical image segmentation. In *Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention*, pages 658–666, 2018.

[39] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2566–2576, 2019.

[40] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. Occlusion aware unsupervised learning of optical flow. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 4884–4893, 2018.

[41] Yan Wang, Chen Zu, Zongqing Ma, Yong Luo, Kun He, Xi Wu, and Jiliu Zhou. Patch-wise label propagation for MR brain segmentation based on multi-atlas images. *Multimedia Systems*, 25(2):73–81, 2019.

[42] Zhenlin Xu and Marc Niethammer. DeepAtlas: Joint semi-supervised learning of image registration and segmentation. In *Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention*, pages 420–429, 2019.

[43] Heran Yang, Jian Sun, Huibin Li, Lisheng Wang, and Zongben Xu. Neural multi-atlas label fusion: Application to cardiac MR images. *Medical Image Analysis*, 49:60–75, 2018.

[44] Amy Zhao, Guha Balakrishnan, Fredo Durand, John V. Guttag, and Adrian V. Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 8543–8553, 2019.

[45] Yizhou Zhou, Xiaoyan Sun, Chong Luo, Zheng-Jun Zha, and Wenjun Zeng. One-shot neural architecture search through a posteriori distribution guided sampling. *CoRR*, abs/1906.09557, 2019.

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. Int'l Conf. Computer Vision*, pages 2223–2232, 2017.