Abstract

In medical image segmentation, supervised deep networks’ success comes at the cost of requiring abundant labeled data. While asking domain experts to annotate only one or a few of the cohort’s images is feasible, annotating all available images is impractical. This issue is further exacerbated when pre-trained deep networks are exposed to a new image dataset from an unfamiliar distribution. Using available open-source data for ad-hoc transfer learning or hand-tuned techniques for data augmentation only provides suboptimal solutions. Motivated by atlas-based segmentation, we propose a novel volumetric self-supervised learning for data augmentation capable of synthesizing volumetric image-segmentation pairs via learning transformations from a single labeled atlas to the unlabeled data. Our work’s central tenet benefits from a combined view of one-shot generative learning and the proposed self-supervised training strategy that cluster unlabeled volumetric images with similar styles together. Unlike previous methods, our method does not require input volumes at inference time to synthesize new images. Instead, it can generate diversified volumetric image-segmentation pairs from a prior distribution given a single or multi-site dataset. Augmented data generated by our method used to train the segmentation network provide significant improvements over state-of-the-art deep one-shot learning methods on the task of brain MRI segmentation. Ablation studies further exemplified that the proposed appearance model and joint training are crucial to synthesize realistic examples compared to existing medical registration methods. The code, data, and models are available at https://github.com/devavratTomar/SST/.

1. Introduction

Automated medical image segmentation, for example, to localize anatomical structures, is of great importance for disease diagnosis or treatment planning. Fully supervised deep neural networks (DNNs) [31, 38] achieve state-of-the-art results when trained on large amounts of labeled data. However, acquiring abundant labeled data is often not feasible, as manual labeling is tedious and costly. Using available open-source data for domain adaptation [41, 5, 4] or hand-tuned approaches for augmentation only provides suboptimal solutions. Furthermore, the cross-modality adaptation methods [41, 42, 6] usually rely on fully labeled source datasets. Medical images can vary significantly from institution to
would be of crucial importance. However, these approaches have not been explored much for low labeled data regime medical image segmentation, e.g., one-shot scenarios. Another classical medical imaging approach often used to reduce the need for labeled data for synthesis and segmentation purposes is an atlas-based approach [7, 2, 26]. In this approach, an atlas is used to register each image and warp one into another (labels undergo the same transformation). One atlas (with its label) is enough for the procedure; however, one can utilize more if available to improve accuracy. Nevertheless, heterogeneity of medical images causes inaccurate warping and, consequently, erroneous segmentation. This heterogeneity issue is even more pronounced for the multi-site medical dataset. Recently, atlas-based approaches empowered by deep convolutional neural networks (CNNs) [50, 14, 49] enable the development of one-shot learning segmentation models.

Among recent one-shot learning methods, two are the most relevant to ours [50, 44]. In the first work, Zhao et al. [50] proposed a learning framework to synthesize labeled samples using CNN to warp images based on atlas. However, [50] only recreates exact styles/deformations presented in the dataset without inducing diversity. Furthermore, training includes two separate steps for spatial and appearance transformations bringing extra computational overhead. In the second work [44], a one-shot segmentation method has been proposed based on the forward-backward consistency strategy to stabilize the training process. Nonetheless, since the atlas’ style does not match the unlabeled image style, this results in imprecise registration and imperfect segmentation. In summary, our main contributions are as follows:

- We propose a novel volumetric self-supervised contrastive learning to learn style representation from unlabeled data facilitating registration and consequently segmentation task in the presence of intra-site and inter-site heterogeneity of MR images (see Fig. 1);

- Unlike current state-of-the-art methods, our method does not require input volumes at inference time to synthesize new images. Instead, it can generate diversified and unlimited image-segmentation pairs by just sampling from a prior distribution;

- Previous methods train the spatial and appearance models separately, resulting in sub-optimal solutions compared to our joint optimization of all modules. Our method achieved state-of-the-art one-shot segmentation performance on two brain T1-weighted MRI datasets and improved the generalization ability for cross-site adaptation scenarios.

2. Related Work

**Data Augmentation.** Various data augmentation techniques have been developed to compensate for the extreme scarcity of labeled data encountered in medical image segmentation [9, 34, 50]. Augmentation approaches range from geometric transformations [31, 33] to data-driven augmentation methods [17, 27]. While former methods are often difficult to tune as they have limited capacities to cover real variations and complex structural differences found in images, latter approaches often learn generative models to synthesize images to combine with real images and train the segmentation model. However, variations in shape and anatomical structures can negatively impact their performance, especially when little training data is available. In this regard, image registration has been effective in approximating deformations between unlabeled images so that the augmented images with plausible deformations can be obtained [50, 48, 8]. The augmented images then allow training deep segmentation models with few labeled examples in a semi-supervised manner. Disappointingly, heterogeneity of medical images often yields inaccurate warping between the moving image and the fixed image and, consequently, inaccurate segmentation. Olut et al. [34] proposed an augmentation method to leverage statistical deformation models from unlabeled data via principal component analysis. Shen et al. [39] proposed a geometric based image augmentation method that generates realistic images via interpolation from the geodesic subspace to estimate the space of anatomical variability. He et al. [19] proposed a Deep Complementary Joint Model (DeepRS) for medical image registration and segmentation.

**Image Segmentation in Low Data Regimes.** Self-supervised learning and few-shot learning are two facets of the same problem: *training a deep model in low labeled data regime*. These approaches have been used for sparsely annotated images for segmentation but without much success. The former often rely on many training classes to avoid overfitting [43, 40], while the latter requires fine-tuning on sufficient labeled data before testing [16, 45]. Deep atlas-based models [3, 50, 44, 13, 2, 26] using a single atlas or multi-atlas tackled weakly-supervised medical image segmentation. Balakrishnan et al. [3] developed VoxelMorph, which aims to estimate pairwise 3D image registration through a learned CNN-based spatial transformation function. In a one-shot learning context, it learns to register a labeled atlas to any
Figure 2: **Schematic description of the training phase.** The **Appearance Model** applies the target image’s style to the source image using its style code predicted by the **Style Encoder**, which is trained in parallel using self-supervised volumetric contrastive loss. Then, the **Flow Model** non-linearly warps the style translated source image to the target image. This allows backpropagation of supervision signals to all three models. Independently, **Flow AAE** maps the flow fields corresponding to the base image generated by using the trained Flow model and Appearance model to a normally distributed latent space. \( D_{\text{style}} \) and \( D_{\text{flow}} \) are trained in an adversarial manner to ensure that the style codes and the flow codes are normally distributed.

other unlabeled volume. This model suffers from variance in voxel intensity confusing the spatial transformer. In this regard, a similar unsupervised method [13] has been proposed that combines a conventional probabilistic atlas-based segmentation with deep learning for MRI scan segmentation. More recently, few deep models [50, 44] explore the one-shot setting for medical image segmentation. Nevertheless, these methods either use samples from a single site (hospital) [44] or aggregate data from multiple sites [50] without cross-dataset transfer learning capability.

3. Method

We first recap the concept of our proposed one-shot atlas-based medical image segmentation. We formulate the synthesis of novel volumetric images and their corresponding segmentation labels as a learned random spatial and style deformation of the given single labeled volumetric atlas image (referred to as **base image**) from learned latent space. We employ a **Style Encoder** (Sec. 3.1) that learns to cluster similar styled images together in a self-supervised manner using volumetric contrastive loss by adapting Momentum Contrast [18] while imposing a normal distribution prior on the latent style codes using adversarial training [28]. The **Appearance Model** (Sec. 3.2) is trained to generate different styles of the base image without changing its spatial structure. For learning the spatial deformation correspondences (referred to as **flow**) between the base image and the target image, we employ **Flow Model** (Sec. 3.3) that is trained on the task of registering two different image volumes with the same style. This is achieved by first changing the style of moving image (referred to as **source image**) to match the target image’s style (as obtained by **Style Encoder**) using the Appearance Model, followed by morphing it into the target image. As discussed later in Sec 4.1, matching the source image’s style to the target image improves the registration accuracy. We employ an additional adversarial autoencoder [28] that encodes the Flow model’s output for the base image into a Gaussian prior flow latent space to learn the distribution of the spatial deformation fields corresponding to the base image. At test time, we sample flow latent code and style latent code from Gaussian prior to generate new deformation fields and style appearances for the base image, respectively. Fig. 2 and Fig. 3 show an overview of the training procedure and the data generation at test time, respectively. To quantify the images’ quality and their corresponding segmentation labels generated using our approach, we train a separate 3D U-Net [12] on them. We test our model on the real image/segmentation pairs and compare it with the performance of a fully supervised model trained using real data. The loss terms used for training Style Encoder, Appearance Model, and Flow Model, described in the subsequent subsections.

3.1. Style Encoder

Style Encoder aims to incentivize content-invariant image-level representation that brings together similar styled images and pushes apart dissimilar styled images. To do so, we
propose a new volumetric contrastive learning [10, 18, 11] based strategy for training. In particular, we adapt Momentum Contrast [18] for volumetric medical images for clustering task of images’ styles as opposed to the original classification task [18]. More importantly, we use the learned spatial transformer [3] to generate the positive keys (preserving the styles) instead of standard augmentation, e.g., random cropping used in the original formulation. Without loss of generality, we keep a dictionary of keys \( \{k_1, k_2, ..., k_K\} \) that represents different styles. During a training step, we sample a volumetric image \( q \) (called \textit{query}) from the training set and generate its corresponding positive \textit{key} volumetric image \( k_+ \) by warping \( q \) to a randomly selected volumetric image from the training set using a pre-trained spatial transformer \( T \). This ensures that \( q \) and \( k_+ \) have the same style (with different structural geometries) that is different from the style keys in the dictionary. The dictionary’s style keys are generated by a separate model (\textit{key-Style Encoder}) whose weights are updated as a moving average of the weights of the Style Encoder with momentum \( m = 0.99 \). The volumetric contrastive loss is computed as below:

\[
L_{\text{vol-cl}} = - \log \frac{\exp(q, k_+ / \tau)}{\sum_{i=0}^{K} \exp(q, k_i / \tau)}
\]

(1)

where \( k_+ = T(q) \) and \( \tau \) is a temperature hyper-parameter [47]. The sum is over one positive and \( K \) negative samples.

### 3.2. Appearance Model

The Appearance Model is responsible for translating the style of the source image \( s \) to that of the target image \( t \), given the style latent code as predicted by the Style Encoder. We feed the target image’s style code to the adaptive instance normalization (AdaIN) layers [20] of the model to perform style transfer. Thus, the target styled source image \( \tilde{s} \) is obtained as:

\[
\tilde{s} = A(s, E_{\text{style}}(t))
\]

(2)

where \( A \) denotes the Appearance model, and \( E_{\text{style}} \) represents the Style Encoder. The Appearance Model loss \( L_{\text{app}} \) consists of two components: \( L_{\text{app}} = L_{\text{style}} + L_{\text{id}} \), where \( L_{\text{style}} \) and \( L_{\text{id}} \) denote the style consistency loss and style identity loss, respectively that are described below.

#### Style Consistency Loss.

We include a style consistency loss that guides the Appearance model to generate images with the same spatial structure as the source image but with a different style in a consistent cyclic manner. Given the style codes of the source and target images, the following style consistency loss is computed as:

\[
L_{\text{style}} = L_{\text{SSIM-L_1}}(s, A(s, E_{\text{style}}(s)))
\]

(3)

where \( L_{\text{SSIM-L_1}} \) computes multi-scale structural similarity index [46] and \( L_1 \) distance between the two images as:

\[
L_{\text{SSIM-L_1}}(u, v) = \|u - v\|_1 + (1 - \text{SSIM}(u, v))
\]

(4)

#### Style Identity Loss.

We also include a regularization loss term, called style identity, that enforces the Appearance Model to generate the same image using its own style.

\[
L_{\text{id}} = L_{\text{SSIM-L_1}}(s, A(s, E_{\text{style}}(s)))
\]

(5)

### 3.3. Flow Model

The Flow Model builds upon a spatial transformer network that warps a moving image \( I_m \) to the fixed image \( I_f \). The spatial transformer \( \mathcal{F} \) generates a correspondence map \( \delta p \), referred to as \textit{flow}, which is used to register \( I_m \) onto \( I_f \). This warping operation is defined as:

\[
\begin{align*}
\delta p &= \mathcal{F}(I_m, I_f) \\
y &= \delta p \odot I_m 
\end{align*}
\]

(6)

where \( \odot \) denotes the warping operator, and \( y \) is the predicted image. Once we know the correspondence map \( \delta p \) between the base image \( b \) and the target image \( t \), we can transfer the segmentation label of the base image \( b_{\text{seg}} \) onto the target image using the same warping operation.

\[
t_{\text{seg}} = \mathcal{F}(b, t) \odot b_{\text{seg}}
\]

(7)

To learn the distribution of deformation fields corresponding to the base image predicted by the Flow Model, we train a separate adversarial autoencoder [28] on its output. This allows us to encode the flow \( \delta p \) in the latent space, which can be used later to generate novel volumetric images and corresponding segmentation labels (Sec. 3.6).

The Flow Model loss \( L_{\text{flow}} \) consists of two components: \( L_{\text{flow}} = L_{\text{recon}} + \lambda_{\text{reg}} L_{\text{flow-reg}} \), where \( L_{\text{recon}} \) and \( L_{\text{flow-reg}} \) denote the reconstruction loss and the flow regularization loss that are described below.

#### Reconstruction Loss.

In contrast to the Normalized Cross-Correlation loss [3] commonly used for the voxel registration, we include pixel similarity based reconstruction loss between the target image \( t \) and the spatially warped target styled source image (referred to as \textit{predicted image}) obtained by Flow Model. Using the pixel-wise similarity loss is justified as the Appearance Model changes the source image’s style to match the style of the target image, thus allowing adequate registration of the two images.

\[
L_{\text{recon}} = L_{\text{SSIM-L_1}}(t, \mathcal{F}(\tilde{s}, t) \odot \tilde{s})
\]

(8)
Flow Regularization. We regularize the flow $\delta p$ by penalizing its spatial gradient, thus ensuring smoothness of the correspondence map generated by our spatial transformer.

$$L_{\text{reg}}^{\text{flow}} = \| \nabla F(s, t) \|_1$$

We prefer the $L_1$ norm as it helps to stabilize training and results in less noisy flow compared to the $L_2$ norm.

3.4. Adversarial Loss

We introduce two latent discriminators, called $D_{\text{style}}$ and $D_{\text{flow}}$, for enforcing a prior distribution on the latent style codes and the flow codes generated during training, respectively. We use the adversarial loss function described in LS-GAN [29] for training the two discriminators along with Style Encoder and Flow Encoder in an adversarial manner, as shown in Fig. 2.

$$L_{\text{adv}}^{\text{style}} = \mathbb{E}_{t \sim X_{\text{data}}} \left[ (D_{\text{style}}(E_{\text{style}}(t)) - 1)^2 \right] + \mathbb{E}_{n \sim \mathcal{N}} [D_{\text{style}}(n)^2]$$

where $t$ is sampled from the training images $X_{\text{data}}$, and $\mathcal{N}$ is the normal distribution. Similarly, we train the adversarial autoencoder (AAE) [28] on flow fields generated by Flow Model (Sec. 3.3) using LS-GAN and $l_1$ reconstruction loss.

3.5. Training Objective

Finally, the proposed training loss $L_{\text{total}}$ joins the loss terms used to train Style Encoder, Appearance Model, and Flow Model:

$$L_{\text{total}} = L_{\text{vol-cl}} + \lambda_1 L_{\text{app}} + \lambda_2 L_{\text{flow}} + \lambda_3 L_{\text{adv}}$$

where $\lambda$’s are the weights of different losses. We observed that pre-training the Style Encoder alone using the loss: $L_{\text{vol-cl}} + \lambda_3 L_{\text{adv}}$ improves the overall convergence rate and reduces the optimization’s complexity. After pre-training the Style Encoder, we jointly optimize it along with Appearance Model and Flow Model by minimizing $L_{\text{total}}$.

3.6. Generation Phase

As shown in Fig. 3, we can sample new volumetric images and their segmentation maps by mapping Gaussian noise to novel style codes and flow fields using the trained Appearance Model ($A$) and Flow Decoder ($G_{\text{flow}}$). The Appearance Model performs style deformation of the base image while the Flow Decoder produces random flow field which is used to warp the styled base image and its segmentation, thus generating new image-segmentation pairs.

$$X = G_{\text{flow}}(n_{\text{flow}}) \circ A(b, n_{\text{style}})$$
$$Y = G_{\text{flow}}(n_{\text{flow}}) \circ b_{\text{seg}}$$

where $b$ and $b_{\text{seg}}$ are the base image and base segmentation and $n_{\text{flow}}, n_{\text{style}} \sim \mathcal{N}$. $X$ and $Y$ are the generated image-segmentation pairs.

4. Experiments

This section introduces the implementation details, experimental settings, dataset, results, and ablation studies.

Dataset, Preprocessing and Evaluation Metric. We evaluate our method and other baselines on multi-study datasets from publicly available datasets: CANDI [23] and a large-scale dataset, OASIS [30], each contains 3D T1-weighted MRI volumes. CANDI dataset consists of scans from 103 patients with manual anatomical segmentation labels, whereas the OASIS dataset has 2044 scans with segmentation labels obtained by the FreeSurfer [15] pipeline. As in VoxelMorph [3] and LT-Net [44], 28 anatomical structures are used in our experiments. All the dataset images are first prepossessed by removing the brain’s skull region, followed by center cropping the volumes to $128 \times 160 \times 160$. We set aside 20% brain images and their segmentation as a test set, which was untouched during training. The remaining 80% brain images are then used for training and validation, with a split of 90%-10%, in which there is no patient ID overlap among the subsets. For each dataset, there are different acquisition details and health conditions. The most similar image to the anatomical average is selected as the only annotated atlas (base image) used for the training. We only use validation set labels for choosing the best model and hyper-parameters. For the OASIS dataset, the model is trained with a mixture of healthy subjects and diseased patients and is then evaluated on test cases constitutes of both sets. We assess our method’s efficiency by training a 3D U-Net-based segmentation model on the generated volumetric image-segmentation pairs and evaluating its perfor-
Dice las) [22], VoxelMorph [3], Bayesian [13] and DataAug [50].

All our models are based on 3D

Cerebral-WM.

Bayesian

Dice score in %) of MABMIS (2 atlas) [22], VoxelMorph [3], Bayesian [13], DataAug [50], LT-Net** [44], and Ours across various brain structures on CANDI and OASIS datasets. Abbreviations used: white matter (WM), cortex (CX), ventricle (Vent), and cerebrospinal fluid (CSF). ** as reported in the published paper. * fully supervised model. The last four rows (OASIS→CANDI and CANDI→OASIS) indicate the results of our method with (w) and without (w/o) cross-site style adaptation.

Table 1: Comparison of segmentation performance (mean Dice score in %) of MABMIS (2 atlas) [22], VoxelMorph [3], Bayesian [13], DataAug [50], LT-Net** [44], and Ours across various brain structures on CANDI and OASIS datasets. Abbreviations used: white matter (WM), cortex (CX), ventricle (Vent), and cerebrospinal fluid (CSF). ** as reported in the published paper. * fully supervised model. The last four rows (OASIS→CANDI and CANDI→OASIS) indicate the results of our method with (w) and without (w/o) cross-site style adaptation.

Figure 4: Comparison of volume-wise segmentation accuracy (Dice score %) of our method with MABMIS (2 atlas) [22], VoxelMorph [3], Bayesian [13] and DataAug [50]. We outperform the second best baseline by a margin of 3.1% on CANDI and 4.1% on OASIS dataset (p-value of $5.8 \times 10^{-4}$, $3.2 \times 10^{-16}$ respectively using paired t-test).

Figure 4: Comparison of volume-wise segmentation accuracy (Dice score %) of our method with MABMIS (2 atlas) [22], VoxelMorph [3], Bayesian [13] and DataAug [50]. We outperform the second best baseline by a margin of 3.1% on CANDI and 4.1% on OASIS dataset (p-value of $5.8 \times 10^{-4}$, $3.2 \times 10^{-16}$ respectively using paired t-test).

Model is a lighter version of VoxelMorph [3], while AAE Model has a 3D convolutonal encoder-decoder architecture. We implement all the models in PyTorch [36]. For training Style Encoder, Appearance Model, Flow Model, and Flow Autoencoder, we used Adam [24] optimizer with a learning rate of $2 \times 10^{-4}$ and ($\beta_1 = 0.9, \beta_2 = 0.999$). We use the same optimizer with the same learning rate for training the latent style code and flow code’s discriminators but with ($\beta_1 = 0.5, \beta_2 = 0.999$). A hyper-parameter search was conducted to find the optimal values. We choose the loss term weights as $\lambda_1 = 5.0, \lambda_2 = 1.0, \lambda_3 = 0.1$ and $\lambda_{reg} = 0.1$. The temperature coefficient $\tau$ for volumetric contrastive loss is set to 0.7. We use a batch size of 32 for pre-training the Style Encoder and 4 when all the models are trained end-to-end. We utilize the same experimental setup for all baseline experiments for a fair comparison.

Comparison with SOTA Methods. Our method surpasses the state-of-the-art methods in most semantic classes and, on average, on both CANDI and OASIS datasets (see Fig. 4). A qualitative comparison of our method with several baselines on coronal brain slices’ segmentation task is shown in Fig. 5. MABMIS [22] and VoxelMorph [3] produce visibly noisy and inaccurate boundaries compared to other methods. Bayesian [13] and DataAug [50] struggle with outer
contours, where one crops it unnecessarily, and another includes background. Furthermore, VoxelMorph and DataAug encounter difficulties in identifying small segmentation regions, whereas Bayesian overproduces them. Instead, our method handles outer/inner regions and smaller anatomical regions well and compares closely to supervised/ground truth results. The qualitative observations are backed by quantitative metrics (see Table 1).

4.1. Ablations

Ablation Study on Appearance Model. The Appearance Model plays a crucial role in generating diversified styled images and improving the Flow Model’s image registration efficiency during training. Table 2 shows the effect of including-excluding the Appearance Model ($L_{app}$) on the registration accuracy of segmentation labels predicted by flow model and supervised segmentation accuracy of the 3D U-Net using the generated images. For both these scenarios, the average Dice score improves after the Appearance Model’s inclusion. This is expected as registering two images with similar intensity distributions is easier than registering images with different intensities. We can only generate similar styled images without the Appearance model for the generation phase, which leads to a poor generalization of 3D U-Net on the test set. A qualitative evaluation of registration on the CANDI test set is shown in Fig. 6.

Ablation Study on Joint Training. We observed that training our model end-to-end is critical for improved registration accuracy (Table 2). We experimented with the pre-training Appearance model, Flow Model, and Style encoder separately using the losses defined in Sec. 3 followed by fine-tuning and found significant improvement when Appearance Model and Flow Model are trained end-to-end. However,
Table 2: Effect of style transfer and joint training. Mean Dice score of the segmentation task with and without Appearance Model and joint training using the method as (a) registration of the base image to the target image (b) 3D U-Net trained on generated image-segmentation pairs.

<table>
<thead>
<tr>
<th>Ablation Type</th>
<th>Method</th>
<th>Mean±std</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o ( L_{\text{app}} )</td>
<td>Reg. (a)</td>
<td>71.6±3.2</td>
</tr>
<tr>
<td>w/o Joint opt.</td>
<td>Reg. (a)</td>
<td>73.6±3.3</td>
</tr>
<tr>
<td>w/ Joint opt., ( L_{\text{app}} )</td>
<td>Reg. (a)</td>
<td>78.9±2.4</td>
</tr>
<tr>
<td>w/o ( L_{\text{app}} )</td>
<td>Sup. (b)</td>
<td>54.9±20.4</td>
</tr>
<tr>
<td>w/ ( L_{\text{app}} )</td>
<td>Sup. (b)</td>
<td>83.5±3.0</td>
</tr>
</tbody>
</table>

Table 1, the domain gap between the two data sites would severely impede the generalization ability of the trained supervised model on the source data site (w/o style adaptation) while performing style adaptation between OASIS and CANDI dataset provides substantial improvements. We use the same test samples from CANDI and OASIS datasets for evaluation.

Exploring Diversity of the Generated Data. We evaluate 3D U-Net’s performance on the segmentation task using different data sizes generated by our approach. For the best performing 3D U-Net, we report the accuracy obtained by training it on 1850 generated image-segmentation pairs. Fig. 7 shows a box-plot of the Dice score (%) of the segmentation accuracy of 3D U-Net trained on different sizes of samples generated using our proposed method trained on CANDI and OASIS. This suggests that our method has learned to generate more diverse and realistic data. Besides, the effects of different flow codes and styles on generated images can be observed in Fig. 8.

5. Conclusion and Future Work

We proposed the novel volumetric contrastive loss used for style transfer by leveraging unlabeled data for one-shot medical image segmentation. We presented a generic method adapted for cross-site one-shot segmentation scenarios to generate arbitrarily diversified volumetric image-segmentation pairs using trained appearance models from one data site and a flow model from another data site. We demonstrated state-of-the-art one-shot segmentation performance on two T1-weighted brain MRI datasets under various settings and ablations. We shed light on our method’s efficacy in closing the gap with a fully-supervised segmentation model in the extreme case of only one labeled atlas. Our method uses neither tissue nor modality-specific information and can be adjusted to other modalities or anatomy. As future work, our method can be easily extended for the few-shot scenario using several atlases.
References


