**Data Augmentation Using Learned Transformations for One-Shot Medical Image Segmentation**

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th><strong>Citation</strong></th>
<th>Zhao, Amy et al. “Data Augmentation Using Learned Transformations for One-Shot Medical Image Segmentation.” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019, Long Beach, California, Institute of Electrical and Electronics Engineers (IEEE), June 2019. © 2019 The Author(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>As Published</strong></td>
<td>10.1109/CVPR.2019.00874</td>
</tr>
<tr>
<td><strong>Publisher</strong></td>
<td>Institute of Electrical and Electronics Engineers (IEEE)</td>
</tr>
<tr>
<td><strong>Version</strong></td>
<td>Original manuscript</td>
</tr>
<tr>
<td><strong>Citable link</strong></td>
<td><a href="https://hdl.handle.net/1721.1/129978">https://hdl.handle.net/1721.1/129978</a></td>
</tr>
<tr>
<td><strong>Terms of Use</strong></td>
<td>Creative Commons Attribution-Noncommercial-Share Alike</td>
</tr>
<tr>
<td><strong>Detailed Terms</strong></td>
<td><a href="http://creativecommons.org/licenses/by-nc-sa/4.0/">http://creativecommons.org/licenses/by-nc-sa/4.0/</a></td>
</tr>
</tbody>
</table>
Data augmentation using learned transformations
for one-shot medical image segmentation

Amy Zhao
MIT
xamyzhao@mit.edu

Guha Balakrishnan
MIT
balakg@mit.edu

Frédéric Durand
MIT
fredo@mit.edu

John V. Guttag
MIT
guttag@mit.edu

Adrian V. Dalca
MIT, MGH
adalca@mit.edu

Abstract

Image segmentation is an important task in many medical applications. Methods based on convolutional neural networks attain state-of-the-art accuracy; however, they typically rely on supervised training with large labeled datasets. Labeling medical images requires significant expertise and time, and typical hand-tuned approaches for data augmentation fail to capture the complex variations in such images.

We present an automated data augmentation method for synthesizing labeled medical images. We demonstrate our method on the task of segmenting magnetic resonance imaging (MRI) brain scans. Our method requires only a single segmented scan, and leverages other unlabeled scans in a semi-supervised approach. We learn a model of transformations from the images, and use the model along with the labeled example to synthesize additional labeled examples. Each transformation is comprised of a spatial deformation field and an intensity change, enabling the synthesis of complex effects such as variations in anatomy and image acquisition procedures. We show that training a supervised segmenter with these new examples provides significant improvements over state-of-the-art methods for one-shot biomedical image segmentation.

1. Introduction

Semantic image segmentation is crucial to many biomedical imaging applications, such as performing population analyses, diagnosing disease, and planning treatments. When enough labeled data is available, supervised deep learning-based segmentation methods produce state-of-the-art results. However, obtaining manual segmentation labels for medical images requires considerable expertise and time. In most clinical image datasets, there are very few manually labeled images. The problem of limited labeled data is exacerbated by differences in image acquisition procedures across machines and institutions, which can produce wide variations in resolution, image noise, and tissue appearance [45].

To overcome these challenges, many supervised biomedical segmentation methods focus on hand-engineered pre-processing steps and architectures [53, 57]. It is also common to use hand-tuned data augmentation to increase the number of training examples [2, 55, 57, 63, 65]. Data augmentation functions such as random image rotations or random nonlinear deformations are easy to implement, and are effective at improving segmentation accuracy in some settings [55, 57, 63, 65]. However, these functions have limited ability to emulate real variations [26], and can be highly sensitive to the choice of parameters [25].

We address the challenges of limited labeled data by learning to synthesize diverse and realistic labeled examples. Our novel, automated approach to data augmentation
leverages unlabeled images. Using learning-based registration methods, we model the set of spatial and appearance transformations between images in the dataset. These models capture the anatomical and imaging diversity in the unlabeled images. We synthesize new examples by sampling transformations and applying them to a single labeled example.

We demonstrate the utility of our method on the task of one-shot segmentation of brain magnetic resonance imaging (MRI) scans. We use our method to synthesize new labeled training examples, enabling the training of a supervised segmentation network. This strategy outperforms state-of-the-art one-shot biomedical segmentation approaches, including single-atlas segmentation and supervised segmentation with hand-tuned data augmentation.

2. Related work

2.1. Medical image segmentation

We focus on the segmentation of brain MR images, which is challenging for several reasons. Firstly, human brains exhibit substantial anatomical variations [28, 59, 76]. Secondly, MR image intensity can vary as a result of subject-specific noise, scanner protocol and quality, and other imaging parameters [45]. This means that a tissue class can appear with different intensities across images—even images of the same MRI modality.

Many existing segmentation methods rely on scan preprocessing to mitigate these intensity-related challenges. Pre-processing methods can be costly to run, and developing techniques for realistic datasets is an active area of research [14, 73]. Our augmentation method tackles these intensity-related challenges from another angle: rather than removing intensity variations, it enables a segmentation method to be robust to the natural variations in MRI scans.

A large body of classical segmentation methods use atlas-based or atlas-guided segmentation, in which a labeled reference volume, or atlas, is aligned to a target volume using a deformation model, and the labels are propagated using the same deformation [6, 13, 22, 32]. When multiple atlases are available, they are each aligned to a target volume, and the warped atlas labels are fused [36, 41, 68, 78]. In atlas-based approaches, anatomical variations between subjects are captured by a deformation model, and the challenges of intensity variations are mitigated using pre-processed scans, or intensity-robust metrics such as normalized cross-correlation. However, ambiguities in tissue appearances (e.g., indistinct tissue boundaries, image noise) can still lead to inaccurate registration and segmentations. We address this limitation by training a segmentation model on diverse realistic examples, making the segmenter more robust to such ambiguities. We focus on having a single atlas, and demonstrate that our strategy outperforms atlas-based segmentation. If more than one segmented example is available, our method can leverage them.

Supervised learning approaches to biomedical segmentation have gained popularity in recent years. To mitigate the need for large labeled training datasets, these methods often use data augmentation along with hand-engineered preprocessing steps and architectures [2, 40, 53, 57, 63, 65, 82].

Semi-supervised and unsupervised approaches have also been proposed to combat the challenges of small training datasets. These methods do not require paired image and segmentation data. Rather, they leverage collections of segmentations to build anatomical priors [21], to train an adversarial network [39], or to train a novel semantic constraint [29]. In practice, collections of images are more readily available than segmentations. Rather than rely on segmentations, our method leverages a set of unlabeled images.

2.2. Spatial and appearance transform models

Models of shape and appearance have been used in a variety of image analyses. Parametric spatial transform models have been used to align and classify handwritten digits [31, 44, 50]. In medical image registration, a spatial deformation model is used to establish semantic correspondences between images. This mature field spans optimization-based methods [4, 7, 67, 70], and recent learning-based methods [8, 9, 20, 42, 62, 72, 80]. We leverage VoxelMorph [8, 9], a recent unsupervised learning-based method, to learn spatial transformations.

Many medical image registration methods focus on intensity-normalized images or intensity-independent objective functions, and do not explicitly account for variations in image intensity. For unnormalized images, models of intensity transforms have used to remove bias field effects from MRI [44, 79]. Spatial and appearance transform models have been used together to register objects that differ in shape as well as texture. Many works build upon the framework of Morphable Models [38] or Active Appearance Models (AAMs) [15, 16], in which statistical models of shape and texture are constructed. AAMs have been used to localize anatomical landmarks [17, 58] and perform segmentation [52, 56, 77]. We build upon these concepts by using convolutional neural networks to learn models of unconstrained spatial and intensity transformations. Rather than learning transform models for the end goal of registration or segmentation, we sample from these models to synthesize new training examples. As we show in our experiments, augmenting a segmenter’s training set in this way can produce more robust segmentations than performing segmentation using the transform models directly.

2.3. Few-shot segmentation of natural images

Few-shot segmentation is a challenging task in semantic segmentation and video object segmentation. Exist-
Figure 2: An overview of the proposed method. We learn independent spatial and appearance transform models to capture the variations in our image dataset. We then use these models to synthesize a dataset of labeled examples. This synthesized dataset is used to train a supervised segmentation network.

Methods for few-shot semantic segmentation incorporate information from prototypical examples of the classes to be segmented [24, 69]. Few-shot video segmentation is frequently implemented by aligning objects in each frame to a labeled reference frame [37, 75]. Other approaches leverage large labeled datasets of supplementary information such as object appearances [11], or incorporate additional information such as human input [60]. Medical images present different challenges from natural images; for instance, the visual differences between tissue classes are very subtle compared to the differences between objects in natural images.

### 2.4. Data augmentation

In image-based supervised learning tasks, data augmentation is commonly performed using simple parameterized transformations such as rotation and scaling. For medical images, random smooth flow fields have been used to simulate anatomical variations [51, 63, 64]. These parameterized transformations can reduce overfitting and improve test performance [34, 43, 51, 63, 64]. However, the performance gains imparted by these transformations vary with the selection of transformation functions and parameter settings [25].

Recent works have proposed learning data augmentation transformations from data. Hauberg et al. [31] focus on data augmentation for classifying MNIST digits. They learn digit-specific spatial transformations, and sample training images and transformations to create new examples aimed at improving classification performance. We learn an appearance model in addition to a spatial model, and we focus on the problem of MRI segmentation. Other recent works focus on learning combinations of simple transformation functions (e.g., rotation and contrast enhancement) to perform data augmentation for natural images [18, 61]. Cubuk et al. [18] use a search algorithm to find augmentation policies that maximize classification accuracy. Ratner et al. [61] learn to create combinations of transformations by training a generative adversarial network on user input. These simple transformations are insufficient for capturing many of the subtle variations in MRI data.

### 3. Method

We propose to improve one-shot biomedical image segmentation by synthesizing realistic training examples in a semi-supervised learning framework.

Let \( \{ y^{(i)} \} \) be a set of biomedical image volumes, and let the pair \((x, l_x)\) represent a labeled reference volume, or at-
Each convolution uses $3 \times 3 \times 3$ kernels, and is followed by a LeakyReLU activation layer. The encoder uses max pooling layers to reduce spatial resolution, while the decoder uses upsampling layers.

Figure 3: We use a convolutional neural network based on the U-Net architecture [63] to learn each transform model. The application of the transformation is a spatial warp for the spatial model, and a voxel-wise addition for the appearance model. We assume a spatial transformation takes the form of a smooth voxel-wise displacement field $u$. Following the medical registration literature, we define the deformation function $\phi = id + u$, where id is the identity function.

To perform data augmentation, we apply transformations $\tau(k)$ to the labeled atlas $x$. We first learn separate spatial and appearance transform models to capture the distribution of anatomical and appearance differences between the labeled atlas and each unlabeled volume. Our method can be easily extended to leverage additional segmented volumes.

To perform data augmentation, we apply transformations $\tau(k)$ to the labeled atlas $x$. We first learn separate spatial and appearance transform models to capture the distribution of anatomical and appearance differences between the labeled atlas and each unlabeled volume. Using the two learned models, we synthesize labeled volumes $\{\tilde{y}(k), \hat{y}(k)\}$ by applying a spatial transformation and an appearance transformation to the atlas volume, and by warping the atlas label maps using the spatial transformation. Compared to single-atlas segmentation, which suffers from uncertainty or errors in the spatial transform model, we use the same spatial transformation to synthesize the volume and label map, ensuring that the newly synthesized volume is correctly labeled. These synthetic examples form a labeled dataset that characterizes the anatomical and appearance variations in the unlabeled dataset. Along with the atlas, this new training set enables us to train a supervised segmentation network. This process is outlined in Fig. 2.

3.1. Spatial and appearance transform models

We describe the differences between scans using a combination of spatial and intensity transformations. Specifically, we define a transformation $\tau(\cdot)$ from one volume to another as a composition of a spatial transformation $\tau_s(\cdot)$ and an intensity or appearance transformation $\tau_a(\cdot)$, i.e., $\tau(\cdot) = \tau_s(\tau_a(\cdot))$.

We assume a spatial transformation takes the form of a smooth voxel-wise displacement field $u$. Following the medical registration literature, we define the deformation function $\phi = id + u$, where id is the identity function. We use $x \circ \phi$ to denote the application of the deformation $\phi$ to $x$. To model the distribution of spatial transformations in our dataset, we compute the deformation that warps atlas $x$ to each volume $y(i)$ using $\phi^i = g_{\theta_s}(x, y^{(i)})$, where $g_{\theta_s}(\cdot, \cdot)$ is a parametric function that we describe later. We write approximate inverse deformation of $y^{(i)}$ to $x$ as $\phi^{-1}(i) = g_{\theta_s}(y^{(i)}, x)$.

We model the appearance transformation $\tau_a(\cdot)$ as per-voxel addition in the spatial frame of the atlas. We compute this per-voxel volume using the function $\psi^i = h_{\theta_a}(x, y^{(i)} \circ \phi^{-1}(i))$, where $y^{(i)} \circ \phi^{-1}(i)$ is a volume that has been registered to the atlas space using our learned spatial model. In summary, our spatial and appearance transformations are:

$$
\tau_s^i(x) = x \circ \phi^i, \quad \phi = g_{\theta_s}(x, y^{(i)}) \quad (1)
$$

$$
\tau_a^i(x) = x + \psi^i, \quad \psi^i = h_{\theta_a}(x, y^{(i)} \circ \phi^{-1}(i)) \quad (2)
$$

3.2. Learning

We aim to capture the distributions of the transformations $\tau_s$ and $\tau_a$ between the atlas and the unlabeled volumes. We estimate the functions $g_{\theta_s}(\cdot, \cdot)$ and $h_{\theta_a}(\cdot, \cdot)$ in Eqs. (1) and (2) using separate convolutional neural networks, with each network using the general architecture outlined in Fig. 3. Drawing on insights from Morphable Models [38] and Active Appearance Models [16, 17], we optimize the spatial and appearance models independently.

For our spatial model, we leverage VoxelMorph [8, 9, 20], a recent unsupervised learning-based approach with an open-source implementation. VoxelMorph learns to output...
a smooth displacement vector field that registers one image to another by jointly optimizing an image similarity loss and a displacement field smoothness term. We use a variant of VoxelMorph with normalized cross-correlation as the image similarity loss, enabling the estimation of \( g_{	heta_c} (\cdot, \cdot) \) with unnormalized input volumes.

We use a similar approach to learn the appearance model. Naively, one might define \( h_{\theta_a} (\cdot, \cdot) \) from Eq. (2) as a simple per-voxel subtraction of the volumes in the atlas space. While this transformation would perfectly reconstruct the target image, it would include extraneous details when the registration function \( \phi^{-1} \) is imperfect, resulting in image details in \( x + \psi \) that do not match the anatomical labels. We instead design \( h_{\theta_a} (\cdot, \cdot) \) as a neural network that produces a per-voxel intensity change in an anatomically consistent manner. Specifically, we use an image similarity loss as well as a semantically-aware smoothness regularization. Given the network output \( y^{(i)} = h_{\theta_a} (x, y^{(i)} \circ \phi^{-1}) \), we define a smoothness regularization function based on the atlas segmentation map:

\[
L_{\text{smooth}}(c_x, \psi) = (1 - c_x) \nabla \psi,
\]

where \( c_x \) is a binary image of anatomical boundaries computed from the atlas segmentation labels \( l_x \), and \( \nabla \) is the spatial gradient operator. Intuitively, this term discourages dramatic intensity changes within the same anatomical region.

In the total appearance transform model loss \( L_a \), we use mean squared error for the image similarity loss \( L_{\text{sim}}(\tilde{y}, y) = \| \tilde{y} - y \|^2 \). In our experiments, we found that computing the image similarity loss in the spatial frame of the subject was helpful. We balance the similarity loss with the regularization term \( L_{\text{smooth}} \):

\[
L_a (x, y^{(i)}, \phi^{(i)}, \psi^{(i)}; \tau^{(i)}_s, \tau^{(i)}_a, c_x) = L_{\text{sim}}(x + \psi^{(i)} \circ \phi^{(i)}, y^{(i)}) + \lambda_a L_{\text{smooth}}(c_x, \psi^{(i)}),
\]

where \( \lambda_a \) is a hyperparameter.

### 3.3. Synthesizing new examples

The models described in Eqs. (1) and (2) enable us to sample spatial and appearance transformations \( \tau_s^{(i)}, \tau_a^{(i)} \) by sampling target volumes \( y^{(i)}, y^{(j)} \) from an unlabeled dataset. Since the spatial and appearance targets can be different subjects, our method can combine the spatial variations of one subject with the intensities of another into a single synthetic volume \( \tilde{y} \). We create a labeled synthetic example by applying the transformations computed from the target volumes to the labeled atlas:

\[
\tilde{y}^{(i,j)} = \tau_s^{(i)} (\tau_a^{(j)} (x)),
\]

\[
\tilde{y}^{(i,j)} = \tau_a^{(i)} (l_x).
\]

This process is visualized in steps 3 and 4 in Fig. 2. These new labeled training examples are then included in the labeled training set for a supervised segmentation network.

### 3.4. Segmentation network

The newly synthesized examples are useful for improving the performance of a supervised segmentation network. We demonstrate this using a network based on the state-of-the-art architecture described in [66]. To account for GPU memory constraints, the network is designed to segment one slice at a time. We train the network on random slices from the augmented training set. We select the number of training epochs using early stopping on a validation set. We emphasize that the exact segmentation network architecture is not the focus of this work, since our method can be used in conjunction with any supervised segmentation network.

### 3.5. Implementation

We implemented all models using Keras [12] and Tensorflow [1]. The application of a spatial transformation to an image is implemented using a differentiable 3D spatial transformer layer [8]; a similar layer that uses nearest neighbor interpolation is used to transform segmentation maps. For simplicity, we capture the forward and inverse spatial transformations described in Section 3.1 using two identical neural networks. For the appearance transform model, we use the hyperparameter setting \( \lambda_a = 0.02 \). We train our transform models with a single pair of volumes in each batch, and train the segmentation model with a batch size of 16 slices. All models are trained with a learning rate of \( 5e^{-4} \). Our code is available at https://github.com/xamyzhao/brainstorm.

### 4. Experiments

We demonstrate that our automatic augmentation method can be used to improve brain MRI segmentation. We focus on one-shot segmentation of unnormalized scans – a challenging but practical scenario. Intensity normalization methods such as bias field correction [27, 71, 74] can work poorly in realistic situations (e.g., clinical-quality scans, or scans with stroke [73] or traumatic brain injury).

#### 4.1. Data

We use the publicly available dataset of T1-weighted MRI brain scans described in [8]. The scans are compiled from eight databases: ADNI [54], OASIS [46], ABIDE [48], ADHD200 [49], MCIC [30], PPMI [47], HABS [19], and Harvard GSP [33]; the segmentation labels are computed using FreeSurfer [27]. As in [8], we resample the brains to \( 256 \times 256 \times 256 \) with 1mm isotropic voxels, and affinely align and crop the images to \( 160 \times 192 \times 224 \). We do not apply any intensity corrections, and we perform...
skull-stripping by zeroing out voxels with no anatomical label. For evaluation, we use segmentation maps of the 30 anatomical labels described in [8].

We focus on the task of segmentation using a single labeled example. We randomly select 101 brain scans to be available at training time. In practice, the atlas is usually selected to be close to the anatomical average of the population. We select the most similar training example to the anatomical average computed in [8]. This atlas is the single labeled example that is used to train our transform models; the segmentation labels of the other 100 training brains are not used. We use an additional 50 scans as a validation set, and an additional 100 scans as a held-out test set.

4.2. Segmentation baselines

Single-atlas segmentation (SAS): We use the same state-of-the-art registration model [8] that we trained for our method’s spatial transform model in a single-atlas segmentation framework. We register the atlas to each test volume, and warp the atlas labels using the computed deformation field [6, 13, 22, 32, 41]. That is, for each test image \( y^{(i)} \), we compute \( \phi^{(i)} = g_{\theta_{S}}(x, y^{(i)}) \) and predict labels \( \hat{y}^{(i)} = l_{\tau} \circ \phi^{(i)} \).

Data augmentation using single-atlas segmentation (SAS-aug): We use SAS results as labels for the unannotated training brains, which we then include as training examples for supervised segmentation. This adds 100 new training examples to the segmenter training set.

Hand-tuned random data augmentation (rand-aug): Similarly to [51, 63, 64], we create random smooth deformation fields by sampling random vectors on a sparse grid, and then applying bilinear interpolation and spatial blurring. We evaluated several settings for the amplitude and smoothness of the deformation field, including the ones described in [63], and selected the settings that resulted in the best segmentation performance on a validation set. We synthesize variations in imaging intensity using a global intensity multiplicative factor sampled uniformly from the range [0.5, 1.5], similarly to [35, 40]. We selected the range to match the intensity variations in the dataset; this is representative of how augmentation parameters are tuned in practice. This augmentation method synthesizes a new randomly transformed brain in each training iteration.

Supervised: We train a fully-supervised segmentation network that uses ground truth labels for all 101 examples in our training dataset. Apart from the atlas labels, these labels are not available for any of the other methods. This method serves as an upper bound.

Table 1: Segmentation performance in terms of Dice score [23], evaluated on a held-out test set of 100 scans. We report the mean Dice score (and standard deviation in parentheses) across all 30 anatomical labels and 100 test subjects. We also report the mean pairwise improvement of each method over the SAS baseline.

<table>
<thead>
<tr>
<th>Method</th>
<th>Dice score</th>
<th>Pairwise Dice improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS</td>
<td>0.759 (0.137)</td>
<td>-</td>
</tr>
<tr>
<td>SAS-aug</td>
<td>0.775 (0.147)</td>
<td>0.016 (0.041)</td>
</tr>
<tr>
<td>Rand-aug</td>
<td>0.765 (0.143)</td>
<td>0.006 (0.088)</td>
</tr>
<tr>
<td>Ours-coupled</td>
<td>0.795 (0.133)</td>
<td>0.036 (0.036)</td>
</tr>
<tr>
<td>Ours-indep</td>
<td>0.804 (0.130)</td>
<td>0.045 (0.038)</td>
</tr>
<tr>
<td>Ours-indep + rand-aug</td>
<td>0.815 (0.123)</td>
<td>0.056 (0.044)</td>
</tr>
<tr>
<td>Supervised (upper bound)</td>
<td>0.849 (0.092)</td>
<td>0.089 (0.072)</td>
</tr>
</tbody>
</table>

4.3. Variants of our method

Independent sampling (ours-indep): As described in Section 3.3, we sample spatial and appearance target images independently to compute \( r_{s}^{(i)}, r_{a}^{(i)} \). With 100 unlabeled targets, we obtain 100 spatial and 100 appearance transformations, enabling the synthesis of 10,000 different labeled examples. Due to memory constraints, we synthesize a random labeled example in each training iteration, rather than adding all 10,000 new examples to the training set.

Coupled sampling (ours-coupled): To highlight the efficacy of our independent transform models, we compare ours-indep to a variant of our method where we sample each of the spatial and appearance transformations from the same target image. This results in 100 possible synthetic examples. As in ours-indep, we synthesize a random example in each training iteration.

Ours-indep + rand-aug: When training the segmenter, we alternate between examples synthesized using ours-indep, and examples synthesized using rand-aug. The addition of hand-tuned augmentation to our synthetic augmentation could introduce additional variance that is unseen even in the unlabeled set, improving the robustness of the segmenter.

4.4. Evaluation metrics

We evaluate the accuracy of each segmentation method in terms of Dice score [23], which quantifies the overlap between two anatomical regions. A Dice score of 1 indicates perfectly overlapping regions, while 0 indicates no overlap. The predicted segmentation labels are evaluated relative to anatomical labels generated using FreeSurfer [27].

4.5. Results

4.5.1 Segmentation performance

Table 1 shows the segmentation accuracy attained by each method. Our methods outperform all baselines in mean
Figure 4: Pairwise improvement in mean Dice score (with the mean computed across all 30 anatomical labels) compared to the SAS baseline, shown across all test subjects.

Figure 5: Pairwise improvement in mean Dice score (with the mean computed across all 30 anatomical labels) compared to the SAS baseline, shown for each test subject. Subjects are sorted by the Dice improvement of ours-indep+rand-aug over SAS.

Dice score across all 30 evaluation labels, showing significant improvements over the next best baselines rand-aug \((p < 1e-15 \text{ using a paired t-test})\) and SAS-aug \((p < 1e-20)\).

In Figs. 4 and 5, we compare each method to the single-atlas segmentation baseline. Fig. 4 shows that our methods attain the most improvement on average, and are more consistent than hand-tuned random augmentation. Fig. 5 shows that ours-indep + rand-aug is consistently better than each baseline on every test subject. Ours-indep alone is always better than SAS-aug and SAS, and is better than rand-aug on 95 of the 100 test scans.

Fig. 6 shows that rand-aug improves Dice over SAS on large anatomical structures, but is detrimental for smaller ones. In contrast, our methods produce consistent improvements over SAS and SAS-aug across all structures. We show several examples of segmented hippocampi in Fig. 7.

4.5.2 Synthesized images

Our independent spatial and appearance models enable the synthesis of a wide variety of brain appearances. Fig. 8 shows some examples where combining transformations produces realistic results with accurate labels.

5. Discussion

Why do we outperform single-atlas segmentation? Our methods rely on the same spatial registration model that is used for SAS and SAS-aug. Both ours-coupled and SAS-aug augment the segmenter training set with 100 new images.

To understand why our method produces better segmentations, we examine the augmented images. Our method warps the image in the same way as the labels, ensuring that the warped labels match the transformed image. On the other hand, SAS-aug applies the warped labels to the original image, so any errors or noise in the registration results in a mis-labeled new training example for the segmenter. Fig. 9 highlights examples where our method synthesizes image texture within the hippocampus label that is more consistent with the texture of the ground truth hippocampus, resulting in a more useful synthetic training example.

Extensions Our framework lends itself to several plausible future extensions. In Section 3.1, we discussed the use of an approximate inverse deformation function for learning the appearance transformation in the reference frame of the atlas. Rather than learning a separate inverse spatial transform model, in the future we will leverage existing work in diffeomorphic registration \([3, 5, 10, 20, 81]\).

We sample transformations from a discrete set of spatial and appearance transformations. This could be extended to span the space of transformations more richly, e.g., through interpolation between transformations, or using compositions of transformations.

We demonstrated our approach on brain MRIs. Since the method uses no brain- or MRI-specific information, it is feasible to extend it to other anatomy or imaging modalities, such as CT.

6. Conclusion

We presented a learning-based method for data augmentation, and demonstrated it on one-shot medical image segmentation.

We start with one labeled image and a set of unlabeled examples. Using learning-based registration methods, we model the set of spatial and appearance transformations between the labeled and unlabeled examples. These transformations capture effects such as non-linear deformations and variations in imaging intensity. We synthesize new labeled examples by sampling transformations and applying them...
Figure 6: Segmentation accuracy of each method across various brain structures. Labels are sorted by the volume occupied by each structure in the atlas (shown in parentheses), and labels consisting of left and right structures (e.g., Hippocampus) are combined. We abbreviate the labels: white matter (WM), cortex (CX), ventricle (vent), and cerebrospinal fluid (CSF).

Figure 7: Hippocampus segmentation predictions for two test subjects (rows). Our method (column 2) produces more accurate segmentations than the baselines (columns 3 and 4).

In summary, this work shows that:

- learning independent models of spatial and appearance transformations from unlabeled images enables the synthesis of diverse and realistic labeled examples, and

- these synthesized examples can be used to train a segmentation model that out-performs existing methods in a one-shot scenario.

References


Figure 8: Since we model spatial and appearance transformations independently, we are able to synthesize a variety of combined effects. We show some examples synthesized using transformations learned from the training set; these transformations form the bases of our augmentation model. The top row shows a synthetic image where the appearance transformation produced a darkening effect, and the spatial transformation shrunk the ventricles and widened the whole brain. In the second row, the atlas is brightened and the ventricles are enlarged.

Figure 9: Synthetic training examples produced by SAS-aug (column 2) and ours-coupled (column 3). When the spatial model (used by both methods) produces imperfect warped labels, SAS-aug pairs the warped label with incorrect image textures. Our method still produces a useful training example by matching the synthesized image texture to the label.


