Adversarial Data Augmentation via Deformation Statistics

Sahin Olut[1], Zhengyang Shen[1], Zhenlin Xu[1],
Samuel Gerber[2], and Marc Niethammer[1]

[1] UNC Chapel Hill {olut,zyshen,zhenlinx,mn}@cs.unc.edu
[2] Kitware Inc. samuel.gerber@kitware.com

Abstract. Deep learning models have been successful in computer vision and medical image analysis. However, training these models frequently requires large labeled image sets whose creation is often very time and labor intensive, for example, in the context of 3D segmentations. Approaches capable of training deep segmentation networks with a limited number of labeled samples are therefore highly desirable. Data augmentation or semi-supervised approaches are commonly used to cope with limited labeled training data. However, the augmentation strategies for many existing approaches are either hand-engineered or require computationally demanding searches. To that end, we explore an augmentation strategy which builds statistical deformation models from unlabeled data via principal component analysis and uses the resulting statistical deformation space to augment the labeled training samples. Specifically, we obtain transformations via deep registration models. This allows for an intuitive control over plausible deformation magnitudes via the statistical model and, if combined with an appropriate deformation model, yields spatially regular transformations. To optimally augment a dataset we use an adversarial strategy integrated into our statistical deformation model. We demonstrate the effectiveness of our approach for the segmentation of knee cartilage from 3D magnetic resonance images. We show favorable performance to state-of-the-art augmentation approaches.

Keywords: Data augmentation, image registration, and segmentation

Image segmentation is an important task in computer vision and medical image analysis, for example, to localize objects of interest or to plan treatments and surgeries. Deep neural networks (DNNs) achieve state-of-the art segmentation performance in these domains [20, 19, 44]. However, training DNNs typically relies on large datasets with labeled structures of interest. In many cases, and for medical segmentation problems in particular, labeled training data is scarce, as obtaining manual segmentations is costly and requires expertise [35]. To allow for training of well-performing DNNs from a limited number of segmentations, various data augmentation strategies have been proposed [24]. Augmentation strategies range from pre-defined random transformations [23, 25, 1, 27] to learning-based approaches [15, 42, 14]. Random transformations usually include intensity changes such as contrast enhancement, brightness adjustments, as well

as random deformations, e.g., affine transformations. These methods are often difficult to tune as they do not directly estimate image variations observable in real data [10, 43, 24]. Generative Adversarial Networks (GANs) [12] are also widely used to generate data augmentations [5, 18, 33]. However, the proposed methods do not explicitly model deformation spaces and hence only have indirect control over augmentation realism.

In fact, existing learning-based data augmentation techniques are generally not based on statistical deformation models as correspondences between random natural image pairs might not be meaningful. However, if such deformations can be established, as is frequently the case for medical images of the same anatomy within or across patients, they may be used to create plausible deformations for data augmentation [43, 18]. While statistical deformation models have a long history in computer vision [8, 7] and medical image analysis [28, 34] they have not been well explored in the context of data augmentation.

Our proposed data augmentation method (*AdvEigAug*) uses learned deformation statistics as a sensible constraint within an adversarial data augmentation setting. Specifically, we make the following contributions:

1) We explicitly model the deformation distribution via principal component analysis (PCA) to guide data augmentation. This allows us to estimate reasonable deformation ranges for augmentation.
2) We propose to efficiently estimate this PCA deformation space via deep image registration models. We explore PCA models on displacement and momentum fields, where the momentum fields assure spatial regularity via the integration of a partial differential equation model.
3) We integrate our PCA model into an adversarial formulation to select deformations for augmentation which are challenging for the segmentation.
4) We extensively evaluate our augmentation approach and show favorable performance with respect to state-of-the-art augmentation approaches.

The manuscript is organized as follows: Sec. 1 gives an overview of related work; Sec 2 describes our *AdvEigAug* technique; Sec. 3 describes our experimental setup; Sec. 4 presents and discusses the results of our method.

1 Related Work

We focus our discussion here on related data augmentation and semi-supervised learning approaches that use adversarial training or image registrations.

Adversarial Training It is well known that adversarial examples created by locally perturbing an input with imperceptible changes may drastically affect image classification results [30]. But it has also been shown that training DNNs with adversarial examples can improve DNN robustness and performance [13], which is our goal. Here, we focus on adversarial training via spatial transformations.

Engstrom *et al.* [11] use rigid transformations to generate adversarial examples for data augmentations, whereas Kanbak *et al.* [21] develop *ManiFool*, an adversarial training technique which is based on affine transformations. A more general adversarial deformation strategy is pursued by Xiao *et al.* [37] where a displacement field is optimized to cause misclassifications. Smoothness of the displacement field is encouraged via regularization.

All these approaches focus on classification instead of segmentation. Furthermore, transformation models are prescribed rather than inferred from observed transformations in the data and selecting a deformation magnitude requires an iterative approach. Our *AdvEigAug* approach instead explores using statistical deformation models obtained from image pairs by fluid- or displacement based registrations. The statistical model also results in a clear guidance for the perturbation magnitudes, thereby eliminating the need for iterations.

Data Augmentation and Semi-Supervised Learning via Registration Image registration is widely used for atlas based segmentation [17]. As it allows estimating deformations between unlabeled image pairs it has also been used to create plausible data deformations for data augmentation. Via *semi-supervised learning* this allows training deep segmentation networks with very limited labeled training data by exploiting large unlabeled image sets. Such approaches have successfully been used in medical image segmentation [38, 43, 5].

For example, Zhao *et al.* [43] train spatial and appearance transformation networks to synthesize labeled images which can then be used to train a segmentation network. Chaitanya *et al.* [5] model appearance variations and spatial transformations via GANs for few-shot image segmentation. Xu *et al.* [38] use a semi-supervised approach to jointly train a segmentation and a registration network. This allows the segmentation network to benefit from the transformations generated via the registration network while simultaneously improving registration results by including the obtained segmentations in the image similarity loss.

However, the approaches above do not employ adversarial samples for training segmentation networks and do not use statistical deformation models. Instead, our *AdvEigAug* approach captures deformation statistics via a PCA model which is efficiently estimated, *separately for each sample*, via deep registration networks and integrated into an adversarial training strategy.

2 Method

Our approach is an the adversarial training scheme. In particular, it builds upon, extends, and combines two different base strategies: (1) the estimation of a statistical deformation model and (2) the creation of adversarial examples that can fool a segmentation network and improve its training. Our proposed *AdvEigAug* approach builds upon the *ManiFool* augmentation idea [21]. *ManiFool* is limited to adversarial deformations for classification which are subject to an affine transformations model. The adversarial directions are based on the gradient with respect to the classification loss, but how far to move in this gradient

direction requires tuning. Our approach on the other hand estimates a statistical deformation space which can be much richer than an affine transformation. In fact, we propose two efficient ways of computing these spaces, both based on non-parametric registration approaches: an approach based on a deep registration network directly predicting displacement fields [3] and another deep network predicting the initial momentum of the large deformation diffeomorphic metric mapping model (LDDMM) [4, 31]. The benefit of the LDDMM model is that it can assure spatially regular (diffeomorphic) spatial transformations. Furthermore, as we estimate a statistical model we can sample from it and we also have a direct measure of the range of deformations which are consistent with deformations observed in the data. Note that these deformation networks can use labeled data [2, 38] to improve registration accuracies, but can also be trained directly on intensity images [40, 3, 31], which is the strategy we follow here.

As our approach is related to *ManiFool* and the augmentation approach in [11] we introduce our reformulation of their concepts for image segmentation and affine transformations in Sec. 2.1. Sec. 2.2 introduces our *AdvEigAug* approach.

2.1 Baseline Method: AdvAffine

Our baseline method is related to [11, 21] where rigid and affine transformations (in 2D) are generated with the goal of fooling a classifier. We extend the approach to 3D affine transformations and apply it to image segmentation instead of classification. While this is a minor change in terms of the loss function (see details below) it precludes simple approaches for step-size selection, i.e., to determine how strong of an affine transformation to apply. For example, while the *ManiFool* approach takes adversarial steps until the classification label changes such an approach is no longer appropriate when dealing with image segmentations as one is now dealing with a set of labels instead of one.

Specifically, we assume we have a deep neural segmentation network, NN, which, given an input image, I, results in a segmentation $\hat{y}$. We also assume we have a segmentation loss, $loss(y, \hat{y})$ (typically a binary cross-entropy loss averaged over the image volume; or a classification loss in [11]), where y is the target segmentation. Our goal is to determine how to spatially transform an input image I so that it is maximally detrimental (i.e., adversarial) for the loss to be minimized. We parameterize the affine transformation as

$$\Phi^{-1}(x;\theta) = (E + A)x + b,\ A, E \in \mathbb{R}^{d\times d},\ b, x \in \mathbb{R}^d, \tag{1}$$

where d denotes the spatial dimension ($d = 3$ in our experiments), x denotes the spatial coordinate, $\theta = \{A, b\}$ are the affine transformation parameters, and E is the identity matrix, which allows us to easily start from the identity transform ($A = 0, b = 0$). The transformed image is then $I \circ \Phi^{-1}$. To compute an adversarial direction we perform gradient ascent with respect to the loss

$$\mathcal{L}(\theta) = loss(y, \hat{y}),\ \text{s.t. } \hat{y} = NN(I \circ \Phi^{-1}(\cdot;\theta)). \tag{2}$$

It is unclear how far to step into the ascent direction for segmentation problems: we simply take t steps with a chosen learning rate, Δt, i.e., $\theta_{t+1} = \theta_t + \Delta t \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta}$.

This process is repeated for each sample in the training data set. The resulting affine transformations are then applied to the images and their segmentations (using nearest neighbor interpolation) to augment the training dataset.

2.2 Proposed Method: AdvEigAug

Fig. 1 gives an overview of our proposed approach. We use a statistical deformation model to capture plausible anatomical variations and efficiently estimate them via two different deep registration approaches [31, 9]. These statistical deformation models are integrated into an adversarial strategy which selects a deformation direction to challenge the segmentation loss.

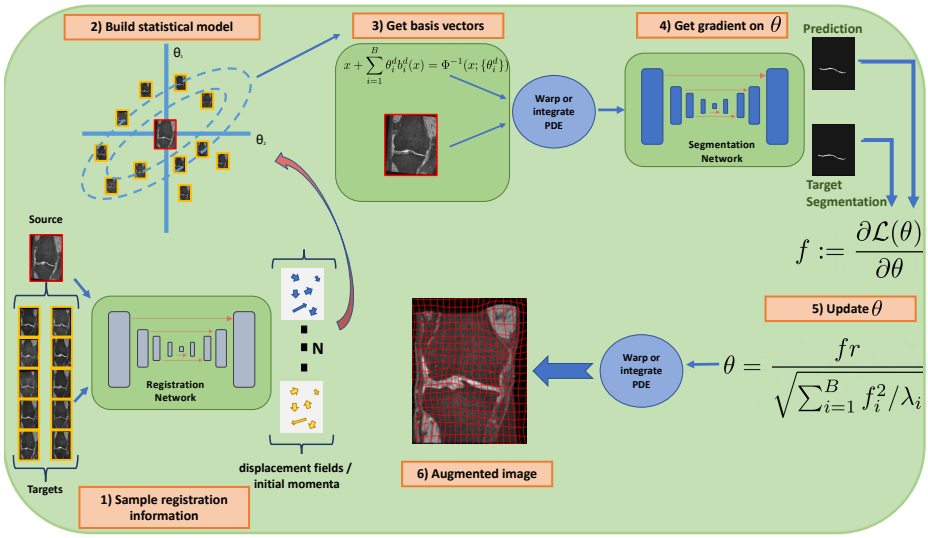

Fig. 1. Overview: We first build the deformation model (1-3) and then obtain an adversarial deformation from the segmentation network (4-6). Our method generates samples (warped images and warped segmentations) that are difficult to segment, but that help training the segmentation network.

Deformation Models Following the approach in Sec. 2.1 we want to obtain adversarial samples subject to a given transformation model. However, instead of specifying this transformation model (e.g., an affine model as in *AdvAffine*) we want to learn plausible models from data.

Displacement Deformation Model We parameterize the transformation via a small number of learned displacement field basis elements:

$$\Phi^{-1}(x; \{\theta_i^d\}) = x + u(x) \approx x + \mu^d(x) + \sum_{i=1}^{B} \theta_i^d b_i^d(x), \tag{3}$$

where $u(x)$ is the displacement field which we approximate by B basis elements. Here, $\mu^d(x)$ is the mean displacement field, $\{b_i^d(x)\}$ is a set of basis displacement fields, and $\{\theta_i^d\}$ are the basis coefficients with respect to which we will compute the adversarial direction. These basis fields could be specified, but we estimate them from sample image pairs (see *Statistical Model* paragraph below).

Fluid Deformation Model Specifying the transformation via basis displacement fields, $\{b_i^d(x)\}$ as in Eq. 3 might result in transformations that are not spatially smooth or invertible. For example, for large values of the coefficients $\{\theta_i\}$ folds might occur. This can be avoided by transformation models based on fluid mechanics [16]. Foldings are avoided in such models by indirectly representing the transformation via integration of a velocity field. Essentially, this amounts to concatenating an infinite number of small deformation whose regularity is easier to control. We use the following LDDMM transformation model:

$$\phi_t^{-1} + D\phi^{-1}v = 0,\ \phi^{-1}(x,0) = x, \tag{4}$$

$$m_t + div(v)m + Dv^T(m) + Dm(v) = 0,\ v = K \star m, \tag{5}$$

$$m(x,0) = m_0(x) \approx \mu^m(x) + \sum_{i=1}^{B} \theta_i^m b_i^m(x),\ \Phi^{-1}(x;\{\theta_i^m\}) = \phi^{-1}(x,1), \tag{6}$$

where $(\cdot)_t$ denotes a partial derivative with respect to time and D the Jacobian. This model is based on the LDDMM shooting equations of [41, 36, 32] which allow specifying the spatial transformation $\phi^{-1}(x,t)$ for all times t via the initial momentum $m_0(x)$ and the solution of the Euler-Poincaré equation for diffeomorphisms [41] (Eq. 5). Our desired transformation $\Phi^{-1}(x;\{\theta_i^m\})$ is obtained after integration for unit time starting from the initial momentum which (similar to the displacement transformation model) we approximate and parameterize via a finite set of momentum basis vector fields $\{b_i^m(x)\}$; $\{\theta_i^m\}$ are the basis coefficients and $\mu^m(x)$ is the mean momentum field. The instantaneous velocity field, v, is obtained by smoothing the momentum field, m, via K. We use a multi-Gaussian smoother K for our experiments [26]. Note that for a sufficiently strong smoother K the resulting spatial transformations are assured to be diffeomorphic. We will see that this is a convenient property for our augmentation strategy.

Statistical Model The introduced deformation models require the specification of their mean displacement and momentum vector fields ($\mu^d(x)$, $\mu^m(x)$) as well as of their basis vector fields ($\{b_i^d(x)\}$, $\{b_i^m(x)\}$). We learn them from data. Specifically, given a source image I we register it to N target images $\{I_i\}$ based on the displacement or the fluid deformation model respectively (without the finite basis approximation). This results in a set of N displacement fields $\{u_i(x)\}$ or initial momentum fields $\{m_i(x)\}$ from which we can build the statistical model. Specifically, we obtain the set of basis vectors ($\{b_i^d(x)\}$,$\{b_i^m(x)\}$) via principal component analysis (PCA) based on the covariance matrix, C, of the displacement or initial momentum fields. As we estimate these spaces relative to a source image I (which defines the tangent space for the transformations)

we typically set $\mu^d(x) = 0$ and $\mu^m(x) = 0$. In the following we denote the set of (eigenvalue,eigenvector) pairs of the covariance matrix as $\{(\lambda_i, e_i(x))\}$, where eigenvalues are ordered in descending order, i.e., $\lambda_i \geq \lambda_{i+1} \geq 0$. The same analysis will hold for the displacement-based and the fluid transformation models. The only difference will be how to obtain the transformations from the basis representations. As computing registrations by numerical optimization is costly, we approximate their solution via deep-learning registration models [3, 31]. These can rapidly predict the displacement or initial momentum fields.

Algorithm 1 *AdvEigAug*

Inputs: I_0, a set of (zero-centered) displacement fields for X registering I_0 to a set of target images, y segmentation mask
Outputs: augmented image I_{aug}, y_{aug} mask for augmented image
 # Compute the Gramian matrix
$\mathbf{G} = \mathbf{X}\mathbf{X}^\mathbf{T}$
$\{\lambda_i\}$,$\{e_i\}$ = eigendecompose($\mathbf{G}$)
 # Compute deformation & warp image
$\{\theta_\mathbf{i}\}_1^B \leftarrow 0$
$\phi^{-1}(x) \leftarrow x + \mu^d(x) + \sum_{i=1}^B \theta_i e_i(x)$
$\hat{\mathbf{I}}_\mathbf{0} = \mathbf{I_0} \circ \phi^{-\mathbf{1}}(\mathbf{x})$
 # Get the gradient wrt. θ
$\hat{\mathbf{y}}$ = predict_segmentation($\hat{\mathbf{I}}_\mathbf{0}$)
f = ∇_θ(segmentation_loss($y, \hat{y}$))
 # Update θ
$\theta \leftarrow \frac{f|r|}{\sqrt{\sum_{i=1}^B f_i^2/\lambda_i}}$.
$\phi^{-1}(x) \leftarrow x + \mu^d(x) + \sum_{i=1}^B \theta_i e_i(x)$
 # Warp image and mask
$\mathbf{I}_{aug} = \mathbf{I_0} \circ \phi^{-\mathbf{1}}(\mathbf{x})$
$\mathbf{y}_{aug} = \mathbf{y} \circ \phi^{-\mathbf{1}}(\mathbf{x})$

Low-rank Approximation Given a set of N centered displacement or initial momentum vector fields (linearized into column vectors) we can write the covariance matrix, C, and its low-rank approximation C_l as

$$C = \sum_{i=1}^{N} \lambda_i e_i e_i^T \approx \sum_{i=1}^{B} \lambda_i e_i e_i^T = C_B. \quad (7)$$

We use the first B eigenvectors to define the basis vectors for our deformation models. As for the *AdvAffine* approach of Sec. 2.1 we can then compute the gradient of the segmentation loss with respect to the transformation parameters

$$f := \frac{\partial \mathcal{L}(\theta)}{\partial \theta}. \quad (8)$$

The only change is in the deformation model where θ is now either $\{\theta_i^d\}$ or $\{\theta_i^m\}$ to parameterize the displacement or the initial momentum field respectively. Everything else stays unchanged. A key benefit of our approach, however, is that the low-rank covariance matrix, C_B induces a B-dimensional Gaussian distribution on the basis coefficient $\{\theta_i\}$, which are by construction decorrelated and have variances $\{\lambda_i\}$, i.e., they are normally distributed according to $N(0, C_m)$, where $C_m = diag(\{\lambda_i\})$. As we will see next, this is beneficial to define step-sizes in the adversarial direction which are consistent with the observed data.

Adversarial Direction So far our transformation models allow more flexible, data-driven transformations than the affine transformation model of Sec. 2.1, but it is still unclear how far one should move in the adversarial gradient direction, f (Eq. 8). However, now that we have a statistical deformation model we can use it to obtain deformation parameters θ which are consistent with the range

of values that should be expected based on the statistical model. Specifically, we want to specify how far we move away from the mean in terms of standard deviations of the distribution, while moving in the adversarial direction. To move $r \geq 0$ standard deviations away we scale the adversarial gradient as follows

$$\theta = \frac{fr}{\sqrt{\sum_{i=1}^{B} f_i^2/\lambda_i}}. \tag{9}$$

It is easy to see why this is the case.

Proof. We first transform the gradient direction (with respect to the basis coefficients, θ), f, to Mahalanobis space via $C_m^{-\frac{1}{2}}$ in which the *coefficients*, θ, are distributed according to $N(0, I)$. We are only interested in the direction and not the magnitude of the adversarial gradient, f. Hence, we normalize it to obtain

$$\overline{f} = \frac{C_m^{-\frac{1}{2}} f}{\|C_m^{-\frac{1}{2}} f\|_2}. \tag{10}$$

In this space, we scale the transformed coefficients by r to move r standard deviations out and subsequently transform back to the original space by multiplying with the inverse transform $C_m^{\frac{1}{2}}$ resulting in

$$\theta = C_m^{\frac{1}{2}} \overline{f} r = C_m^{\frac{1}{2}} \frac{C_m^{-\frac{1}{2}} fr}{\|C_m^{-\frac{1}{2}} f\|_2} = \frac{fr}{\|C_m^{-\frac{1}{2}} f\|_2} = \frac{fr}{\sqrt{\sum_{i=1}^{B} f_i^2/\lambda_i}}. \tag{11}$$

Given an adversarial direction, f, this allows us to sample a set of transformation coefficients, θ, which are consistent with this adversarial direction *and* which have a desired magnitude, r, as estimated via the statistical deformation model. Hence, in contrast to the *AdvAffine* approach of Sec. 2.1 there is now an intuitive way for specifying how large the adversarial deformations should be so that they remain consistent with the observed deformations between image pairs. Fig. 1 shows an example illustration of determining such a scaled adversarial direction. Pseudo-code for the overall augmentation algorithm for the displacement-based deformation model is given in Alg. 2.2. The fluid-flow-based approach follows similarly, but requires integration and backpropagation through Eqs. 4-6. The resulting deformations are then applied to the training image and its segmentation (using nearest neighbor interpolation) to augment the training dataset.

3 Experiments

We investigate the performance of different augmentation strategies for the segmentation of knee cartilage from the 3D magnetic resonance images (MRIs) of the Osteoarthritis Initiative[3]. All images are affinely registered to a knee

[3] https://nda.nih.gov/oai/

atlas. The original images are of size $384 \times 384 \times 160$ with a voxel size of $0.36 \times 0.36 \times 0.7\ mm^3$. We normalize the intensities of each image such that the 0.1 percentile and the 99.9 percentile are mapped to $\{0, 1\}$ and clamp values that are smaller to 0 and larger to 1 to avoid outliers. We do not bias-field correct these images as it has been shown in [39] that it is not required for this dataset. This also justifies our choice to test the *Brainstorm* approach without its appearance-normalization part. To be able to store the 3D data in the 11 GB of our NVIDIA GTX 2080Ti GPU we re-sample the input images and their segmentations to $190 \times 190 \times 152$. Note that this resampling might introduce slight evaluation bias with respect to the manual segmentations drawn on the original full-resolution images. However, our objective is to compare augmentation approaches which would all be equally affected, hence, the relative comparisons between methods remain fair. To assure that we can compare between methods we use the same hyperparameters that we use in the *NoAug* experiments across all experiments. All segmentation networks are trained with a learning rate of 0.001 using Adam [22] where $\beta_1 = 0.5$ and $\beta_2 = 0.99$. For the displacement field network, we use a learning rate of 0.0002. For the momentum generation network, we use the same settings as in [32].

3.1 Registration and Segmentation Networks

We use a 3D-UNet [6] with 5 encoder and decoder layers. The number of feature channels in the encoder layers are $[16, 32, 32, 32, 32]$ and the number of feature channels in the decoder layers are $[32, 32, 32, 8, 8]$. Each convolutional block is followed by batch normalization and ReLU activation except for the last one.

Segmentation Network We use binary cross-entropy loss for simplicity and jointly segment the femoral and tibial cartilage. However, our approach generalizes to any loss function, e.g., to multiple class labels.

Registration Networks We train two registration networks: to predict (1) displacement fields and (2) the initial momentum of EPDiff (Eq. 5).

Displacement field network For the displacement field network we follow the VoxelMorph architecture [3], but use a modified loss of the form

$$\mathcal{L}_{disp}(u(x)) = \mathcal{L}^d_{reg}(u(x)) + \lambda^d \mathcal{L}_{sim}(I_T, I \circ (x + u(x))). \tag{12}$$

where u is the displacement field predicted by the network for a given source image, I, and target image I_T; $\mathcal{L}_{sim}$ is the image similarity measure (we use normalized cross correlation (NCC)) and $\mathcal{L}_{reg}$ is the regularizer for which we choose the bending energy [29] to keep transformations smooth:

$$\mathcal{L}^d_{reg}(u(x)) = \frac{1}{N} \sum_x \sum_{i=1}^{d} \|H(u_i(x)\|^2_F, \tag{13}$$

where $\|\cdot\|_F$ is the Frobenius norm, $H(u_i(x))$ is the Hessian of the i-th component of $u(x)$, N the number of voxels, and d is the spatial dimension ($d = 3$). We set $\lambda^d = 200$.

Initial momentum network We use the LDDMM network of [32], which predicts the initial momentum as an intermediate output. The network loss is

$$\mathcal{L}_m(m_0(x)) = \mathcal{L}^m_{reg}(m_0(x)) + \lambda^m \mathcal{L}_{sim}(I_T, I \circ \Phi^{-1}), \text{ s.t. } Eqs.\ 4-6 \text{ hold}, \quad (14)$$

where $m_0(x)$ is the initial momentum (not yet approximated via a finite basis for the network loss), $\mathcal{L}_{reg} = \langle m_0, K * m_0 \rangle$, Φ^{-1} is the spatial transform induced by $m_0(x)$ and we use a Localized NCC similarity loss [31]. See [32] for details.

3.2 Experimental Design

Our goal is to demonstrate that (1) building a statistical deformation model is useful, that (2) using adversarial samples based on the statistical deformation model improves segmentation results, and that (3) *AdvEigAug* outperforms other augmentation approaches, in particular, approaches that do not attempt to learn deformation models from data. We focus on a few-shot learning setting.

Methods we Compare to and Rationale for Comparisons

NoAug uses no augmentation. It only uses the segmented training samples. All augmentation approaches are expected to outperform this method.

RandDeform is our representative of an augmentation strategy using various spatial transformations which are *not* inferred from data and hence is expected to show inferior performance to approaches that do. Specifically, we randomly rotate images between ± 15 degrees in all axes and apply random translations by up to 15 voxels in each space direction. We simultaneously apply random displacement fields. Translations and rotations are drawn from uniform distributions. The displacement fields are obtained by Gaussian smoothing of random displacement fields drawn from a unit normal distribution.

AdvAffine is described in detail in Sec. 2.1. It is our baseline adversarial augmentation approach. It lacks our ability to determine a desirable deformation magnitude, uses a simple affine transformation, and does not build a statistical deformation model. Our various *AdvEigAug* approaches directly compete with *AdvAffine* and favorable performance indicates that using statistical deformation models can be beneficial for augmentation. We choose $t = 20$ and $\Delta t = 0.25$ to select the deformation magnitude for *AdvAffine*.

Brainstorm [43] relates to our *AdvEigAug* approach in the sense that it uses unlabeled images via registrations (we use the displacement field network of Sec. 3.1 for its implementation) to create realistic deformations for augmentation. In contrast to *AdvEigAug* it is not an adversarial approach nor does it build a statistical deformation model and hence it relies on the set of deformations that can be observed between images. We will show that these design choices matter by comparing to *AdvEigAug* directly and to a non-adversarial variant *EigAug* (explained next). *Brainstorm* also uses an intensity transfer network. However, as we are working on one specific dataset without strong appearance differences we only use the deformation component of *Brainstorm* for a more direct comparison.

EigAug is our *AdvEigAug* approach when replacing the adversarial gradient direction, f by a random direction. We compare against this approach to (1) show that modeling the deformation space is beneficial (e.g., with respect to *Brainstorm*) and (2) that using the adversarial direction of *AdvEigAug* yields improvements.

AdvEigAug is our proposed approach and described in detail in Sec. 2.2.

Upper bound We use the segmentations for the entire training set ($n = 200$) to train the segmentation network. We regard the resulting segmentation accuracy as the quasi upper-bound for achievable segmentation accuracy.

Randomization and Augmentation Strategies

Randomization *AdvEigAug* and *EigAug* require the selection of, r, which specifies how many standard deviations away from the mean the parameterized deformation is. In our experiments we explore fixing $r \in \{1, 2, 3\}$ as well as randomly selecting it: $r \sim U(1.5, 3)$. We hypothesize that randomly selecting r results in a richer set of deformations and hence in better segmentation performance.

Covariance and basis vector computations *AdvEigAug* and *EigAug* also require the computation of the covariance matrix based on displacement or momentum fields. We recompute this covariance matrix for *every* augmented sample we create to assure sufficient variability of deformation directions. Specifically, for a given training sample we randomly select N images (we do not use or need their segmentations) and register the training sample to these images using one of our registration networks to obtain the displacement or momentum fields respectively. We choose the top two eigendirections ($B = 2$) as our basis vectors. Clearly, using more samples results in more stable results, but it also limits the space being explored. In the extreme case one would use all available image pairs to estimate the covariance matrix which would then limit the variability. We therefore use only $N = 10$ samples to estimate these directions.

Offline augmentation For almost all our experiments we employ an *offline* augmentation strategy. We train the segmentation network with the original training data for 700 epochs. We then create a set of adversarial examples. Specifically, we create 5 adversarial samples for each training sample. We then continue training with the augmented training set for another 1,300 epochs. To assure that training is not biased too strongly to the adversarial samples [24] we down-weight the adversarial examples by 1/5 in the loss.

Online augmentation We also explore an online augmentation approach for *AdvEigAug* only to determine if the successive introduction of adversarial samples is beneficial. As for the offline approach, we first train the segmentation network for 700 epochs. We then add 1 adversarial example per original training sample and continue training for 150 epochs. We repeat this augmentation step 4 more times and finally train for another 550 epochs. As for the offline variant adversarial samples are down-weighted. We down-weight by $\frac{1}{k}$, where k is the k^{th} addition of adversarial samples. This assures that at all times the adversarial samples are balanced with the original training data.

Data We split our dataset into a test set with segmentations (n=100), a validation set with segmentations (n=50), as well as a training set (n=200) of which we use n=4 or n=8 as the original training set including their segmentations. This training set (excluding their segmentations) is also used to obtain deformations for *Brainstorm* and the deformation spaces for *AdvEigAug* and *EigAug*.

4 Results and Discussion

Tab. 1 shows Dice scores and surface distance measures between the obtained segmentations and the manual knee cartilage segmentations on the test set ($n = 100$). As we focus on training models with small numbers of manual segmentations we show results for training with 4 and 8 segmentations respectively. Images without segmentations were also used during training of the *Brainstorm*, *EigAug*, and *AdvEigAug* models. All results are averages over 5 random runs; standard deviations are reported in parentheses.

Overall Performance Our proposed *L-AdvEigAug*$^{\text{online}}_{\text{rand\_r}}$ approach performs best in terms of Dice scores and surface distances demonstrating the effectiveness of using statistical deformation models to capture plausible anatomical variations in low sample settings in combination with a randomized deformation approach and an LDDMM-based deformation model which can assure spatially well behaved deformations. In particular, our *AdvEigAug* approaches outperform all competing baseline approaches we tested: no augmentation (*NoAug*); augmentations with random deformations (*RandDeform*); *Brainstorm* which also uses unsegmented images to obtain deformations via registrations; and a competing adversarial approach using affine transformations only (*AdvAffine*). Note that our images are all already affinely aligned to an atlas, hence an improvement by

AdvAffine was not expected in this setting. However, it illustrates that the more flexible deformation models of *AdvEigAug* are indeed beneficial.

Table 1. Reported mean Dice scores and surface distance measures in *mm* (average surface distance (ASD), 50th, and 95th-percentile) over 5 training runs. Surface distances are measured based on the binarized predictions. Standard deviations are in parentheses. Prefix L denotes training with the LDDMM model, prefix Adv indicates using adversarial samples. Dice scores for the approaches were tested for statistical difference with respect to the L-AdvEigAug$^{online}_{rand_r}$ approach (bold) using a paired t-test. At a significance level $\alpha = 0.05$ with 14 different tests per measure results in statistical significance based on Bonferroni correction for $p < \alpha/14 \approx 0.0035$; † shows statistical significance where $p < 1e^{-3}$ and * where $p < 1e^{-6}$.

	4 training samples				8 training samples			
Experiment	Dice in %	ASD	50%	95%	Dice in %	ASD	50%	95%
NoAug	75.2(0.31)*	1.79	0.134	20.17	77.1(0.28)*	1.17	0.093	9.14
RandDeform	76.3(0.34)*	1.80	0.171	14.72	78.1(0.25)*	0.87	0.072	4.12
Brainstorm [43]	77.7(0.20)*	1.17	0.087	8.80	79.4(0.17)*	0.71	0.031	**3.17**
AdvAffine	75.1(0.49)*	1.62	0.102	13.78	77.8(0.37)*	1.19	0.042	5.13
EigAug$_{\text{rand_r}}$	73.3(0.81)*	1.76	0.352	9.37	76.2(0.13)*	0.86	0.139	5.15
EigAug$_{\text{fix_r=1}}$	73.4(0.45)*	1.45	0.241	12.04	77.8(0.12)*	0.92	0.086	5.36
EigAug$_{\text{fix_r=2}}$	73.6(0.40)*	1.46	0.217	10.32	77.7(0.10)*	0.85	**0.011**	5.17
EigAug$_{\text{fix_r=3}}$	73.0(0.40)*	1.44	0.232	11.05	77.3(0.30)*	0.98	0.147	5.24
AdvEigAug$_{\text{fix_r=1}}$	75.6(0.31)*	1.77	0.231	10.12	77.8(0.23)*	0.97	0.074	6.78
AdvEigAug$_{\text{fix_r=2}}$	76.1(0.20)*	1.24	0.128	9.28	78.7(0.21)*	0.88	0.031	4.76
AdvEigAug$_{\text{fix_r=3}}$	74.9(0.15)*	1.75	0.287	11.18	76.9(0.27)*	1.22	0.113	9.99
AdvEigAug$_{\text{rand_r}}$	78.4(0.25)†	0.85	0.091	7.14	81.0(0.15)†	0.65	0.031	3.77
AdvEigAug$^{\text{online}}_{\text{rand_r}}$	78.2(0.23)†	0.92	0.071	6.44	81.1(0.16)†	0.58	0.025	3.51
L-AdvEigAug$_{\text{rand_r}}$	78.5(0.17)	0.79	0.083	4.68	81.1(0.11)	0.55	0.021	3.85
L-AdvEigAug$^{\text{online}}_{\text{rand_r}}$	**79.1(0.14)**	**0.72**	**0.020**	**3.72**	**81.2(0.12)**	**0.51**	0.023	3.46
Upper bound	82.5(0.2)	0.47	0.019	2.42				

Adversarial Sample Effect Comparing the adversarial *AdvEigAug* to the corresponding non-adversarial *EigAug* results directly shows the impact of the adversarial samples. In general, the adversarial examples result in improved Dice scores for the 4 training sample cases for $r \in \{1, 2, 3\}$ and in similar Dice scores for the 8 training sample case. Surface distances appear comparable. The exception is the randomized $AdvEigAug_{rand\_r}$ strategy which performs uniformly better on all measures than $EigAug_{rand\_r}$ and all its fixed r value variants.

***Randomization of Deformation Magnitude,* r** Tab. 1 shows that randomizing $r \sim U(1.5, 3)$ improves performance over using a fixed r in terms of Dice scores and surface distances. As we draw 5 augmented samples per training sample, randomization occurs for fixed r only via the computation of the covariance matrices based on the randomly drawn images. The random r strategy, however, introduces additional randomness through the random deformation magnitude and can therefore explore the deformation space more fully.

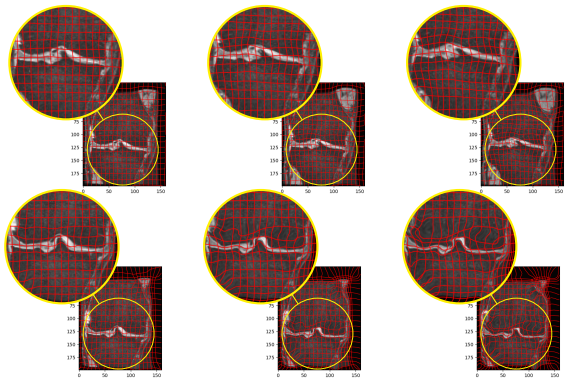

Fig. 2. Augmentation with LDDMM (top row) and displacement field network (bottom row). Left to right: augmentations with $r = 1$, $r = 2$, $r = 3$. LDDMM produces well behaved transformations even for large values of r.

Online vs Offline The online and offline variants of our approach show comparable performance, though we observe a slight improvement for the 95-th percentile surface distance for the online approach. This might be because the online approach allows improving upon earlier adversarial examples during training.

LDDMM vs Displacement Field Networks The LDDMM parameterization assures a well-behaved diffeomorphic transformation regardless of the chosen deformation magnitude, r. In contrast the displacement field approach might create very strong deformations even to the point of creating foldings. Fig. 2 shows that augmented samples created via the LDDMM model tend to indeed be better behaved than the ones created via the displacement field model. While slight, this benefit appears to be born out by the validation results in Tab. 1 which generally show higher Dice scores and lower surface distance measures for the LDDMM models, in particular for the lowest sample size ($n = 4$).

5 Conclusion

We introduced an end-to-end data augmentation framework guided by a statistical deformation model. To the best of our knowledge, this is the first work that studies statistical deformation modelling in conjunction with data augmentation. We proposed two variants of our method, one using a fluid- and the other a displacement-based deformation model. We showed that such statistical deformation models allow an intuitive control over deformation magnitudes and that a combination of randomizing the deformation magnitude, online training, and a fluid-based deformation model performs best for our segmentation task.

Acknowledgements: Research reported in this publication was supported by the National Institutes of Health (NIH) under award numbers NIH 1R41MH118845 and NIH 1R01AR072013. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D.L., Erickson, B.J.: Deep learning for brain MRI segmentation: state of the art and future directions. Journal of digital imaging **30**(4), 449–459 (2017)
2. Balakrishnan, G., Zhao, A., Sabuncu, M., Guttag, J., Dalca, A.V.: VoxelMorph: A learning framework for deformable medical image registration. IEEE TMI: Transactions on Medical Imaging **38**, 1788–1800 (2019)
3. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: VoxelMorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging **38**(8), 1788–1800 (2019)
4. Beg, M.F., Miller, M.I., Trouvé, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. International journal of computer vision **61**(2), 139–157 (2005)
5. Chaitanya, K., Karani, N., Baumgartner, C.F., Becker, A., Donati, O., Konukoglu, E.: Semi-supervised and task-driven data augmentation. In: International Conference on Information Processing in Medical Imaging. pp. 29–41. Springer (2019)
6. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
7. Cootes, T.F., Taylor, C.J.: Statistical models of appearance for medical image analysis and computer vision. In: Medical Imaging 2001: Image Processing. vol. 4322, pp. 236–248. International Society for Optics and Photonics (2001)
8. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Computer vision and image understanding **61**(1), 38–59 (1995)
9. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning for fast probabilistic diffeomorphic registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 729–738. Springer (2018)
10. Eaton-Rosen, Z., Bragman, F., Ourselin, S., Cardoso, M.J.: Improving data augmentation for medical image segmentation (2018)
11. Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., Madry, A.: A rotation and a translation suffice: Fooling CNNs with simple transformations. arXiv preprint arXiv:1712.02779 (2017)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
13. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
14. Hataya, R., Zdenek, J., Yoshizoe, K., Nakayama, H.: Faster AutoAugment: Learning augmentation strategies using backpropagation. arXiv preprint arXiv:1911.06987 (2019)
15. Ho, D., Liang, E., Stoica, I., Abbeel, P., Chen, X.: Population based augmentation: Efficient learning of augmentation policy schedules. arXiv preprint arXiv:1905.05393 (2019)
16. Holden, M.: A review of geometric transformations for nonrigid body registration. IEEE transactions on medical imaging **27**(1), 111–128 (2007)

17. Iglesias, J.E., Sabuncu, M.R.: Multi-atlas segmentation of biomedical images: a survey. Medical image analysis **24**(1), 205–219 (2015)
18. Jendele, L., Skopek, O., Becker, A.S., Konukoglu, E.: Adversarial augmentation for enhancing classification of mammography images. arXiv preprint arXiv:1902.07762 (2019)
19. Kamnitsas, K., Ferrante, E., Parisot, S., Ledig, C., Nori, A.V., Criminisi, A., Rueckert, D., Glocker, B.: DeepMedic for brain tumor segmentation. In: International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries. pp. 138–149. Springer (2016)
20. Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical image analysis **36**, 61–78 (2017)
21. Kanbak, C., Moosavi-Dezfooli, S.M., Frossard, P.: Geometric robustness of deep networks: analysis and improvement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4441–4449 (2018)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Moeskops, P., Viergever, M.A., Mendrik, A.M., De Vries, L.S., Benders, M.J., Išgum, I.: Automatic segmentation of MR brain images with a convolutional neural network. IEEE transactions on medical imaging **35**(5), 1252–1261 (2016)
24. Paschali, M., Simson, W., Roy, A.G., Naeem, M.F., Göbl, R., Wachinger, C., Navab, N.: Data augmentation with manifold exploring geometric transformations for increased performance and robustness. arXiv preprint arXiv:1901.04420 (2019)
25. Pereira, S., Pinto, A., Alves, V., Silva, C.A.: Brain tumor segmentation using convolutional neural networks in MRI images. IEEE transactions on medical imaging **35**(5), 1240–1251 (2016)
26. Risser, L., Vialard, F.X., Wolz, R., Murgasova, M., Holm, D.D., Rueckert, D.: Simultaneous multi-scale registration using large deformation diffeomorphic metric mapping. IEEE transactions on medical imaging **30**(10), 1746–1759 (2011)
27. Roth, H.R., Lu, L., Farag, A., Shin, H.C., Liu, J., Turkbey, E.B., Summers, R.M.: Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 556–564. Springer (2015)
28. Rueckert, D., Frangi, A.F., Schnabel, J.A.: Automatic construction of 3-D statistical deformation models of the brain using nonrigid registration. IEEE transactions on medical imaging **22**(8), 1014–1025 (2003)
29. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J.: Non-rigid registration using free-form deformations: application to breast MR images. IEEE transactions on medical imaging **18**(8), 712–721 (1999)
30. Shaham, U., Yamada, Y., Negahban, S.: Understanding adversarial training: Increasing local stability of supervised models through robust optimization. Neurocomputing **307**, 195–204 (2018)
31. Shen, Z., Han, X., Xu, Z., Niethammer, M.: Networks for joint affine and non-parametric image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4224–4233 (2019)
32. Shen, Z., Vialard, F.X., Niethammer, M.: Region-specific diffeomorphic metric mapping. arXiv preprint arXiv:1906.00139 (2019)
33. Shin, H.C., Tenenholtz, N.A., Rogers, J.K., Schwarz, C.G., Senjem, M.L., Gunter, J.L., Andriole, K.P., Michalski, M.: Medical image synthesis for data augmentation

and anonymization using generative adversarial networks. In: International workshop on simulation and synthesis in medical imaging. pp. 1–11. Springer (2018)
34. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: A survey. IEEE transactions on medical imaging **32**(7), 1153–1190 (2013)
35. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J., Wu, Z., Ding, X.: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. arXiv preprint arXiv:1908.10454 (2019)
36. Vialard, F.X., Risser, L., Rueckert, D., Cotter, C.J.: Diffeomorphic 3D image registration via geodesic shooting using an efficient adjoint calculation. International Journal of Computer Vision **97**(2), 229–241 (2012)
37. Xiao, C., Zhu, J.Y., Li, B., He, W., Liu, M., Song, D.: Spatially transformed adversarial examples. arXiv preprint arXiv:1801.02612 (2018)
38. Xu, Z., Niethammer, M.: Deepatlas: Joint semi-supervised learning of image registration and segmentation. arXiv preprint arXiv:1904.08465 (2019)
39. Xu, Z., Shen, Z., Niethammer, M.: Contextual additive networks to efficiently boost 3D image segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 92–100. Springer (2018)
40. Yang, X., Kwitt, R., Styner, M., Niethammer, M.: Quicksilver: Fast predictive image registration–a deep learning approach. NeuroImage **158**, 378–396 (2017)
41. Younes, L., Arrate, F., Miller, M.I.: Evolutions equations in computational anatomy. NeuroImage **45**(1), S40–S50 (2009)
42. Zhang, X., Wang, Q., Zhang, J., Zhong, Z.: Adversarial AutoAugment. arXiv preprint arXiv:1912.11188 (2019)
43. Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V.: Data augmentation using learned transformations for one-shot medical image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8543–8553 (2019)
44. Zhou, X.Y., Guo, Y., Shen, M., Yang, G.Z.: Artificial intelligence in surgery. arXiv preprint arXiv:2001.00627 (2019)