City Research Online

City St George's, University of London

Citation: Chen, C., Qin, C., Ouyang, C., Li, Z., Wang, S., Qiu, H., Chen, L., Tarroni, G., Bai, W. & Rueckert, D. (2022). Enhancing MR image segmentation with realistic adversarial data augmentation. Medical Image Analysis, 82, 102597. doi: 10.1016/j.media.2022.102597

This is the published version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/28861/

Link to published version: https://doi.org/10.1016/j.media.2022.102597

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

Medical Image Analysis 82 (2022) 102597

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Enhancing MR image segmentation with realistic adversarial data augmentation

Chen Chen [a,*], Chen Qin [b,c], Cheng Ouyang [a], Zeju Li [a], Shuo Wang [d,e], Huaqi Qiu [a], Liang Chen [a], Giacomo Tarroni [a,f], Wenjia Bai [a,g,h], Daniel Rueckert [a,i]

[a] Department of Computing, Imperial College London, UK
[b] Institute for Digital Communications, School of Engineering, University of Edinburgh, UK
[c] Department of Electronics and Electrical Engineering, Imperial College London, UK
[d] Digital Medicine Research Centre, School of Basic Medical Sciences, Fudan University, China
[e] Shanghai Key Laboratory of MICCAI, Shanghai, China
[f] CitAI Research Centre, Department of Computer Science, City, University of London, UK
[g] Department of Brain Sciences, Imperial College London, UK
[h] Data Science Institute, Imperial College London, UK
[i] Klinikum rechts der Isar, Technical University of Munich, Germany

ARTICLE INFO

Keywords:
MR image segmentation
Adversarial training
Data augmentation
Adversarial data augmentation
Model generalization

ABSTRACT

The success of neural networks on medical image segmentation tasks typically relies on large labeled datasets for model training. However, acquiring and manually labeling a large medical image set is resource-intensive, expensive, and sometimes impractical due to data sharing and privacy issues. To address this challenge, we propose *AdvChain*, a generic adversarial data augmentation framework, aiming at improving both the *diversity* and *effectiveness* of training data for medical image segmentation tasks. *AdvChain* augments data with *dynamic data augmentation*, generating randomly chained photo-metric and geometric transformations to resemble realistic yet challenging imaging variations to expand training data. By jointly optimizing the data augmentation model and a segmentation network during training, challenging examples are generated to enhance network generalizability for the downstream task. The proposed adversarial data augmentation does not rely on generative networks and can be used as a plug-in module in general segmentation networks. It is computationally efficient and applicable for both low-shot supervised and semi-supervised learning. We analyze and evaluate the method on two MR image segmentation tasks: cardiac segmentation and prostate segmentation with limited labeled data. Results show that the proposed approach can alleviate the need for labeled data while improving model generalization ability, indicating its practical value in medical imaging applications.

1. Introduction

Medical image segmentation plays an essential role in healthcare applications, including disease diagnosis, treatment planning, and clinical research (Smistad et al., 2015). In recent years, many deep learning-based techniques have been developed for medical image segmentation, achieving high performance in terms of both speed and accuracy (Shen et al., 2017; Litjens et al., 2017). However, training a deep neural network generally requires a large amount of labeled data. In medical imaging, acquiring and manually labeling such a large dataset is extremely challenging for several reasons. First, labeling medical images is time-consuming and expensive as it requires experienced human observers with domain expertise. Second, collecting and sharing large datasets across clinical sites is difficult due to data privacy and ethical issues. As a result, it is typical that only a small number of labeled images are available for training a neural network, which hinders the deployment of effective deep learning solutions for healthcare applications.

To alleviate the data scarcity problem, data augmentation approaches have been proposed (Shorten and Khoshgoftaar, 2019), which aim to increase the diversity of the available training data without collecting and manually labeling new data. Conventional data augmentation methods mainly focus on applying simple *random* transformations to labeled images. These random transformations include intensity transformations (e.g. pixel-wise noise or image-wise brightness and

* Corresponding author.
E-mail address: chen.chen15@imperial.ac.uk (C. Chen).

https://doi.org/10.1016/j.media.2022.102597
Received 10 August 2021; Received in revised form 17 June 2022; Accepted 19 August 2022
Available online 28 August 2022
1361-8415/© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

contrast adjustment) and geometric transformations (e.g. affine or elastic transformations). Most of these transformations perform basic image manipulations without taking image contents into account or accounting for downstream tasks, which may introduce redundant data points that does not necessarily improve generalization (Miyato et al., 2018).

In this work, we introduce a generic adversarial data augmentation framework (*AdvChain*), aiming at improving both the *diversity* and *effectiveness* of training data for medical image segmentation tasks. *AdvChain* improves data diversity with *dynamic data augmentation*, generating randomly chained photo-metric and geometric transformations to resemble realistic complex data variation at training. Different from conventional random data augmentation approaches, *AdvChain* allows to optimize the underlying transformation parameters in arbitrarily chained transformations (providing they are differentiable). By taking both image information and the current network fragility into account to optimize the transformation parameters, *AdvChain* improves the 'hardness' of augmented images to better regularize the network training (Section 3.1).

The proposed framework *AdvChain* can accommodate a wide range of differentiable photometric and geometric transformations for the joint optimization of data augmentation and network in both supervised and semi-supervised learning. In this work, as a proof of concept, four different image transformation models are employed to resemble realistic imaging variations in MR imaging. They are: (a) an image noise augmentation model; (b) an intensity transformation model which amplifies intensity non-uniformity by simulating low-frequency intensity corruptions caused by inhomogeneities of the magnetic field; (c) a global image geometric transformation model based on affine transformation that simulates patient movement (e.g., rotation, translation) and imaging resolution variations (e.g., scaling) during scanning; (d) a diffeomorphic deformation model which simulates intra-subject morphological difference attributed to pathology, growth and motion, and inter-subject morphological difference. By generating realistic and various 'hard' examples for data augmentation, we force the network to learn robust semantic features against various imaging variations, leading to improved model generalization. Besides, *AdvChain* strengthens the consistency regularization for medical segmentation tasks by employing a composite loss function, which encourages both pixel-level consistency as well as contour-based consistency (Section 3.4). We demonstrate the efficacy of the proposed method on two public MR image datasets in challenging low-data supervised and semi-supervised settings (e.g., with only 1 labeled subject for training). Our method outperforms several strong consistency-regularized methods and strong composite data augmentation method (RandAugment (Cubuk et al., 2020)) in low-data regimes (e.g., with only 1 or 3 labeled subject for training), indicating its efficacy to improve the generalizability of the model on MR segmentation tasks when labeled data is limited.

This work is an extension to our previously presented work at MICCAI (Chen et al., 2020), where we introduced adversarial photometric data augmentation with a bias field intensity transformation model and demonstrated its effectiveness on a *binary* cardiac segmentation task. In this work, we substantially extend the framework by including both adversarial photometric and geometric transformations and composing these transformations in a flexible way to further improve image diversity and resemble data variations in magnetic resonance (MR) imaging. In particular, we present a novel adversarial diffeomorphic deformation model to generate challenging morphological variations, as a way to improve the segmentation model generalization ability. Finally, we extend the framework to multi-class segmentation problems and comprehensively evaluate our method on two public datasets, one consisting of cardiac MR images and the other of prostate MR images. Experiments on both datasets show the effectiveness of our method, which improves image segmentation performance and outperforms competitive consistency regularization-based methods.

2. Related work

We first review several advanced data augmentation techniques that have been developed recently (e.g. data mixing, adversarial data augmentation) and then introduce consistency-based semi-supervised learning methods which are closely related to this work.

2.1. Data mixing

Data mixing methods generate new data samples by mixing multiple samples together (Zhang et al., 2018; Hendrycks et al., 2020; Berthelot et al., 2019). A representative work is Mixup (Zhang et al., 2018), which creates new training samples by combining random pairs of images (x_a, x_b) and their labels (y_a, y_b) via linear interpolation: $x^{new} = \beta x_a + (1-\beta)x_b$, $y^{new} = \beta y_a + (1-\beta)y_b$, where β is a weighting parameter sampled from the beta distribution. Though originally proposed for image classification, Mixup has been successfully adapted to medical image segmentation tasks, including knee segmentation (Panfilov et al., 2019), brain segmentation (Li et al., 2019), and cardiac segmentation (Chaitanya et al., 2019). One problem with this technique is that the mixed images can be unrealistic and difficult to interpret. Also, the diversity of generated samples by data mixing is limited since the mixed samples still lie in the span of the training data (Wu et al., 2020).

2.2. Adversarial data augmentation

Adversarial data augmentation applies perturbations to original images to fool the model into making classification mistakes. These perturbed images (also known as adversarial images) are then used to optimize the network for improved robustness against particular perturbations. Recent studies have shown that adversarial data augmentation can be more effective than random data augmentation (Madry et al., 2017; Volpi et al., 2018; Suzuki and Sato, 2020). Most existing works are based on simple gradient-based noise attack, i.e. using the gradients of the neural network to generate additive adversarial noise to perturb images (Madry et al., 2017; Goodfellow et al., 2015; Carlini and Wagner, 2017; Tramèr and Boneh, 2019; Miyato et al., 2018; Paschali et al., 2018). However, researchers have found that neural networks can be fragile to other more complex forms of transformations that may occur in images, such as affine transformations (Kanbak et al., 2018; Engstrom et al., 2019; Zeng et al., 2019; Finlayson et al., 2019), illumination changes (Zeng et al., 2019) or small deformations (Alaifari et al., 2019). For medical image segmentation, the majority of related works focus on crafting effective adversarial examples and leverage them to evaluate model robustness. For example, Paschali et al. (2018) applied a targeted attack, specifically a dense adversary generation (DAG) attack (Xie et al., 2017), to generate effective pixel-wise noise, which fools a segmentation network into producing poor segmentation on brain images. Chen et al. (2019a) proposed to use conditional GANs to model spatial deformation and noises for adversarial image construction.

In contrast to existing adversarial data augmentation which augments images with a *single, fixed* type of image transformation (Miyato et al., 2018; Wang et al., 2021a; Chen et al., 2020; Xie et al., 2017; Alaifari et al., 2019; Kanbak et al., 2018; Zeng et al., 2019; Finlayson et al., 2019; Madry et al., 2017; Goodfellow et al., 2015; Carlini and Wagner, 2017; Tramèr and Boneh, 2019; Paschali et al., 2018), *AdvChain* is capable of directly optimizing the transformations parameters in *dynamic* data augmentations, e.g., arbitrarily chained image photometric and geometric transformations, better generating realistic and challenging image variations that may occur at medical imaging applications. Existing composite data augmentation optimization works such as generative adversarial network (GAN)-based data augmentation approach (Gao et al., 2021) are very computational intensive and suffer from the training instability problem, as they need to train different GANs to produce photometric and geometric transformation

parameters separately. Their approach cannot be used to optimize randomly chained transformations due to high training instability and memory costs, which involves the optimization of multiple stacked GANs. Our method, by contrast, can efficiently optimize all different transformations in a chain, even with only one forward pass and backward pass.

On top of *AdvChain*, we also present a *novel* adversarial data augmentation with diffeomorphic transformations based on stationary velocity fields, which could generate realistic, morphological variations to fool the network. At training, we directly optimize the underlying static velocity field and integrate them to generate diffeomorphic deformations. This is fundamentally different from existing adversarial deformation works based on GANs (Chen et al., 2019a; Gao et al., 2021) where a generative network is required to model additive displacement fields. The network has to be pre-trained with a carefully designed regularization loss on the deformation fields to restrict the realism of generated deformations. The produced deformations may not be invertible, thus restricting its use for computing the pixel-wise consistency regularization in the original input space.

2.3. Consistency regularization

Viewing data augmentation as a way of encoding invariances and equivalences into a neural network, consistency regularization methods apply data augmentation to unlabeled data for semi-supervised learning based on the assumption that the predictions of a data point and its augmented/perturbed example should be consistent (Sajjadi et al., 2016; Li et al., 2020; Miyato et al., 2018; Berthelot et al., 2019; Xie et al., 2020a; Sohn et al., 2020; Wang et al., 2021a). A consistency regularization term is generally introduced to the loss function to encourage a model to produce consistent predictions on similar inputs (e.g., unlabeled data and its augmented ones). On the basis of this mechanism, many works explored different data augmentation techniques, including random data augmentation (e.g., pixel-level noise, affine transformations) (Sajjadi et al., 2016; Liu et al., 2020; Li et al., 2020), data mixing (Berthelot et al., 2019; Hendrycks et al., 2020) and adversarial data augmentation techniques (Miyato et al., 2018; Volpi et al., 2018; Xie et al., 2020b; Suzuki and Sato, 2020).

For medical image segmentation tasks, several related works explored different types of data augmentation to enhance consistency regularization (Cui et al., 2019; Li et al., 2020; Chen et al., 2020). These works focused on utilizing weak, random augmentation methods such as random Gaussian noise (Cui et al., 2019), random affine transformations (Li et al., 2020), and adversarial bias fields (Chen et al., 2020). A major difference in our work is that we consider modeling more complex photometric and geometric transformations and propose adversarial training to optimize the transformation parameters to generate *more challenging augmented images*. We believe that, with more diverse and effective realistic data augmentation to regularize training, the proposed method can better enforce the model to learn high-level, robust representations for an improved generalization ability.

3. Methods

The goal of medical image segmentation is to learn a mapping from an image space $\mathcal{X}$ to a label space $\mathcal{Y}$. In deep learning, the mapping is parameterized by a neural network, e.g., U-net (Ronneberger et al., 2015), f_θ: $\mathcal{X} \rightarrow \mathcal{Y}$, where θ denotes the network parameters, such as weights and biases in the convolutional layers. Assume we have a *small* labeled dataset $D_L : \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ $(N > 0)$ and an unlabeled dataset $D_U : \{(\mathbf{x}_i)\}_{i=1}^{M} (M \geq 0)$, where images $\mathbf{x}$ and pixel-wise labels $\mathbf{y}$ are drawn from the joint distribution $P(\mathcal{X}, \mathcal{Y})$. The learning goal is to train a network f_θ parameterized by θ to model the conditional probability distribution $P(\mathcal{Y}|\mathcal{X})$. As the network usually contains millions of parameters, it is important to regularize the network to alleviate the over-fitting problem, especially when there is *insufficient* training data.

Recent works on image classification have shown that consistency regularization with data augmentation can be an effective approach to regularizing the network and exploiting the value of unlabeled data (Xie et al., 2020a). Our method follows this learning paradigm. At a high level, the learning objective for the network can be formulated as follows:

$$\min_\theta \mathbb{E}_{\mathbf{x}\in D_l} \mathcal{L}_s \left(\mathbf{y}, f_\theta(\mathbf{x})\right) + \lambda \mathbb{E}_{\mathbf{x}\in D_l \cup D_u} \mathcal{R}(\mathbf{x}; f_\theta, \mathcal{T}_\mathbf{t}). \tag{1}$$

Here $\mathcal{L}_s$ denotes the supervised loss (e.g., cross-entropy loss) for labeled images in the training set; $\mathcal{R}$ is a consistency regularization term computed on both labeled and unlabeled data; λ is a weighting factor to balance the supervised and regularization loss terms. In particular, $\mathcal{R}$ measures the inconsistency between the prediction for the original image $f_\theta(\mathbf{x})$ and the prediction for the image under a single or a composite perturbation/transformation function: $f_\theta(\mathcal{T}_\mathbf{t}(\mathbf{x}))$. $\mathcal{T}_\mathbf{t}$ is short for $\mathcal{T}(\cdot, \mathbf{t})$ denoting the transformation function $\mathcal{T}$ parameterized by $\mathbf{t}$.

3.1. AdvChain: A generic adversarial data augmentation framework for effective consistency regularization

In this work, we employ an adversarial training approach to first optimize transformation parameters $\mathbf{t}$ so that augmented images can better regularize network training (Miyato et al., 2018). In other words, we would like first to find perturbations/transformations to which the current segmentation model is most sensitive and then utilize them for consistency regularization. Different from existing adversarial data augmentation methods (Chen et al., 2020; Miyato et al., 2018) where they only consider a *fixed* type of pixel-level perturbations, we propose *AdvChain*, a generic adversarial data augmentation framework which allows optimizing the parameters in a random chain of different image transformation functions (incl. photo-metric and geometric transformations) for improved data diversity, with the aim of better reflecting complex image variations in MR imaging. In this work, we employ four different image transformation functions as a proof of concept (will be introduced in Section 3.3), allowing to generate not only pixel-level perturbations but also geometric variations, e.g. morphological variations.

The whole learning procedure can be generally described as a two-step optimization, as shown in Fig. 1:

- With the segmentation parameters θ fixed, we update the image transformation parameters $\mathbf{t}$ in the search space to produce an adversarial image augmentation $\mathcal{T}$, so that it **maximizes** the disagreement (measured by $\mathcal{R}$) between the original prediction $\mathbf{p} = f_\theta(\mathbf{x})$ and the prediction on the perturbed image $\mathbf{p}' = f_\theta(\mathcal{T}_\mathbf{t}(\mathbf{x}))$. Here $\mathcal{T}$ can be a single transformation or a composite transformation with chained image transformation functions $\mathcal{T} : \mathcal{T}_1 \circ \mathcal{T}_2, \ldots, \circ \mathcal{T}_n$ $(n >= 1)$ with corresponding transformation parameters $\mathbf{t} : \mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n$;
- With the optimized transformation parameters $\mathbf{t}^*$ fixed, we obtain an *AdvChain* augmented image $\mathbf{x}^* = \mathcal{T}_{\mathbf{t}^*}(\mathbf{x})$ and feed it to the network to update the segmentation network parameters θ to **minimize** the supervised loss $\mathcal{L}_s$ and the consistency regularization loss $\mathcal{R}$.

Mathematically, the learning objective can then be formulated as follows:

$$\min_\theta \mathbb{E}_{\mathbf{x}\in D_l} \mathcal{L}_s \left(\mathbf{y}, f_\theta(\mathbf{x})\right) + \lambda \mathbb{E}_{\mathbf{x}\in D_l \cup D_u} \mathcal{R}(\mathbf{x}; f_\theta, \mathcal{T}_{\mathbf{t}^*}). \tag{2a}$$

$$s.t.\ \mathbf{t}^* = \underset{\mathbf{t}:C(\mathbf{t})}{\operatorname{argmax}} \ \mathcal{R}(\mathbf{x}; f_\theta, \mathcal{T}_\mathbf{t}). \tag{2b}$$

Here, $C(\mathbf{t})$ denotes a set of constraints that specify the search space of corresponding transformation parameters. These constraints are essential as they explicitly ensure that augmented or perturbed images remain meaningful and realistic. Since it is difficult to determine the optimum parameters $\mathbf{t}^*$ in practice, we relax the objective in Eq. (2)(b).

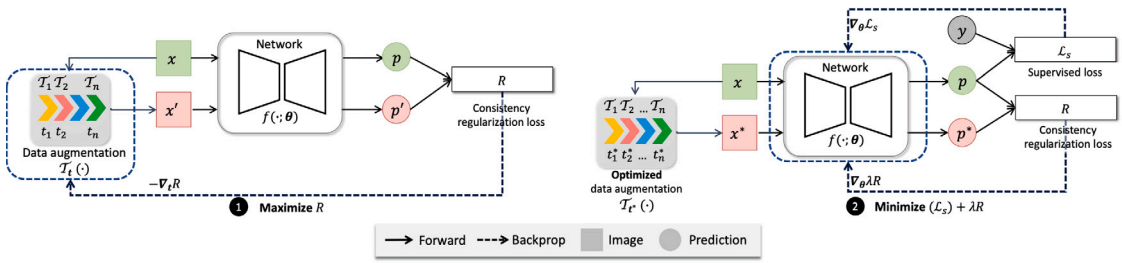

Fig. 1. ***AdvChain* overview.** *AdvChain* is a generic adversarial data augmentation framework for medical image segmentation, which allows optimizing the parameters in a *randomly sampled* augmentation chain (incl. photo-metric and geometric transformations) for enhanced consistency regularization. **Left**: Given a segmentation network f_θ, an input image $\mathbf{x}$ and a *randomly sampled* chain of transformation functions $\mathcal{T} : \mathcal{T}_1 \circ \mathcal{T}_2, \ldots, \circ \mathcal{T}_n$ $(n >= 1)$ with corresponding transformation parameters $\mathbf{t} : \mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n$, *AdvChain* first optimizes the underlying transformation parameters $\mathbf{t}$ in the direction of *maximizing* the inconsistency (measured by $\mathcal{R}$) between the network prediction for the original image $\mathbf{p} = f_\theta(\mathbf{x})$ and the prediction for the augmented image $\mathbf{p}' = f_\theta(\mathbf{x}')$. **Right**: The updated transformation parameters $\mathbf{t}^*$ are then used to generate an *AdvChain* augmented image $\mathbf{x}^*$ to train the network. Specifically, the network parameters θ are optimized in the direction of *minimizing* the consistency loss $\mathcal{R}$ computed between the original prediction $\mathbf{p}$ and the prediction for the *AdvChain* image $\mathbf{p}^*$, in together with the supervised loss $\mathcal{L}_s$ (if its ground-truth label $\mathbf{y}$ is available).

We instead try to find a relatively effective $\mathbf{t}^*$ that produces higher inconsistency loss $\mathcal{R}$ to strengthen the network regularization, compared to its random initialized counterpart. To achieve the goal, we employ the commonly used projected gradient descent (PGD) algorithm (Madry et al., 2017) to update the randomly initialized transformation parameters in a chain, which has been found effective to optimize the parameters with constraints across many applications (Xing et al., 2021):

$$\mathbf{t}_i \leftarrow \Pi_C \ \mathbf{t}_i + \alpha_i \nabla_{\mathbf{t}_i} \mathcal{R} / \|\nabla_{\mathbf{t}_i} \mathcal{R}\|_2. \tag{3}$$

Here, Π is the projection operation that projects the updated parameters onto the feasible set constrained by C, α_i specifies the step size when we update the parameters $\mathbf{t}_i$ for the transformation function $\mathcal{T}_i$ in a chain along the direction of the normalized gradient $\nabla_{\mathbf{t}_i} \mathcal{R} / \|\nabla_{\mathbf{t}_i} \mathcal{R}\|_2$. In practice, we use the same step size for different transformation for simplicity. We apply the chain rule to efficiently compute the gradients along the augmentation chain[1]. We use normalized gradients to update the parameters in each transformation function to avoid gradient explosion or vanishing problems when the length of chained transformations is long.

3.2. Increasing the data diversity of AdvChain with dynamic transformations

AdvChain allows to optimize *dynamic* transformations: e.g. single or composite transformations randomly generated at training. Such flexibility is highly adorable as the data diversity can be largely increased at a low cost. While it is possible that better performance can be achieved by employing the optimum combinations of transformation functions for a particular task, it often requires extraordinary high computational costs to search for improved data augmentation policies (Cubuk et al., 2019). Therefore, in *AdvChain* we simply randomly select and chain the transformations in an arbitrary order, allowing itself to explore all possible solutions as a trade-off between efficiency and effectiveness. Specifically, for each image, each transformation function is randomly selected with a probability of p and then chained in a random order to produce a high diversity of augmented images. We then apply adversarial training to this chain, which optimizes the underlying parameters in each transformation.

In Algorithm 1, we illustrate the detailed steps of the proposed adversarial data augmentation method with a random chain of transformations for consistency regularization. For ease of understanding, we use subscripts 1, 2, 3 to represent three different arbitrary image transformation functions.

Algorithm 1 *AdvChain*

Input: labelled or unlabelled training set: $L \cup U$, a segmentation network f_θ
Requires: a set of predefined transformations $\{\mathcal{T}\}$, number of update steps k, step size α
for $\mathbf{x} \in L \cup U$ **do**
  Randomly select transformation operations with a probability of p from a group of transformation functions $\{\mathcal{T}\}$, e.g., $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$.
  Chain them in a random order, e.g., $\mathcal{T}_{312} = \mathcal{T}_3 \circ \mathcal{T}_1 \circ \mathcal{T}_2$ and randomly initialize the transformation parameters $\mathbf{t}_{312}^{(0)} : \mathbf{t}_3^{(0)}, \mathbf{t}_1^{(0)}, \mathbf{t}_2^{(0)}$
  for j=0,..., k-1 **do**
    Compute consistency loss $\mathcal{R}(\mathbf{x}; f_\theta, \mathcal{T}_{312}(\cdot; \mathbf{t}_{312}^{(j)}))$
    Apply the chain rule to computing gradients $\nabla_{\mathbf{t}_3} \mathcal{R}$, $\nabla_{\mathbf{t}_1} \mathcal{R}$, $\nabla_{\mathbf{t}_2} \mathcal{R}$ and update $\mathbf{t}_3$, $\mathbf{t}_1$, $\mathbf{t}_2$, respectively using Eq.(3)
  end for
  Return the chain of data augmentation with optimized parameters to augment images: $\mathcal{T}_{312}^{adv}(\mathbf{x}) = \mathcal{T}_{312}(\mathbf{x}; \mathbf{t}_{312}^{adv})$.
  Compute the loss for network optimization using Eq. (2a).
end for

3.3. Realistic image transformation functions

AdvChain is an advanced, generic, lightweight adversarial data augmentation framework, which can be applied to optimize any types of differentiable photometric and geometric transformations on-the-fly. In this work, we consider four different image transformation functions as a proof of concept. The transformation functions are constructed to reflect common data variations that exist in MR images, including:

- an additive image noise model $\mathcal{T}_{\text{noise}}$, which adds pixel-wise noise to images (Section 3.3.1),
- an image intensity transformation model $\mathcal{T}_{\text{bias}}$, which generates bias fields to corrupt images. Bias field can introduce intensity inhomogeneities in images, which is a common artefact in MR imaging (Sled et al., 1998; Tustison et al., 2010; Ferreira et al., 2013) (Section 3.3.2),
- an image geometric transformation model $\mathcal{T}_{\text{affine}}$, which simulates image spatial variance caused by patient movement and the adjustment of acquisition parameters (e.g., image resolution, field-of-view) during scanning (Section 3.3.3),
- a diffeomorphic deformation model $\mathcal{T}_{\text{morph}}$, which simulates inter- and intra-subject shape variability attributed to age, pathology, and motion (Section 3.3.4).

[1] Applying the chain rule allows us to calculate the gradient of the loss function with respect to the parameters of each transformation function in a chain in an efficient way. The transformation functions are required to be differentiable. In our work, all transformations satisfy the criterion, where the geometric transformations are implemented using the differentiable spatial transformer module (Jaderberg et al., 2015).

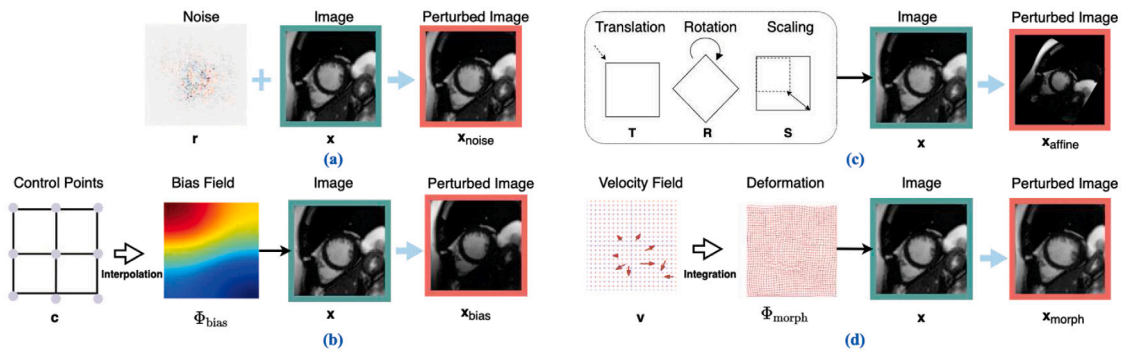

Fig. 2. Adversarial example construction with: (a) image noise model $\mathcal{T}_{\text{noise}}$: (b) image intensity transformation model with bias field $\mathcal{T}_{\text{bias}}$; (c) image geometric transformation model $\mathcal{T}_{\text{affine}}$; (d) image deformation model $\mathcal{T}_{\text{morph}}$.

Without loss of generality, we assume that all image transformations are defined in 2D space and performed on 2D images $\mathbf{x} \in \mathbb{R}^{H \times W}$. One should note that these transformations can be potentially extended to 3D space.

3.3.1. Image noise model $\mathcal{T}_{\text{noise}}$

We begin with the most commonly used image noise model, which applies additive noise to perturb images, as shown in Fig. 2(a). Following existing adversarial data augmentation works (Goodfellow et al., 2015; Madry et al., 2017; Miyato et al., 2018), the image noise model is formulated as

$$\mathbf{x}_{\text{noise}} = \mathcal{T}_{\text{noise}}(\mathbf{x}; \mathbf{r}) = \mathbf{x} + \mathbf{r}, \tag{4}$$

where the magnitude of the noise $\mathbf{r}$ is bounded by ϵ_{noise}. The magnitude constraint C_{noise} can be described as follows:

$$\|\mathbf{r}\|_2 \leq \epsilon_{\text{noise}}. \tag{5}$$

Here $\epsilon_{\text{noise}}(\epsilon_{\text{noise}} \geq 0)$ is a scalar controlling the level of noise.

3.3.2. Image intensity transformation with bias field $\mathcal{T}_{\text{bias}}$

Following previous related works on bias field correction (Sled et al., 1998; Tustison et al., 2010), a multiplicative intensity transformation is adopted here to introduce intensity non-uniformity to images. As shown in Fig. 2(b), the intensity of the image $\mathbf{x}$ is perturbed by multiplying with the bias field $\boldsymbol{\Phi}_{\text{bias}} \in \mathbb{R}^{H \times W}$:

$$\mathbf{x}_{\text{bias}} = \mathcal{T}_{\text{bias}}(\mathbf{x}; \mathbf{c}) = \mathbf{x} \odot \boldsymbol{\Phi}_{\text{bias}}(\mathbf{c}), \tag{6}$$

where $\odot$ denotes point-wise multiplication. Similar to the bias field model in N4ITK (Tustison et al., 2010), we use a set of uniformly distributed $b \times b$ control points $\mathbf{c} = \{c^{(i,j)}\}_{b \times b}$ for bias field construction, based on the fact that the bias field is smoothly varying across the image, see Fig. 2(b). Following Sled et al. (1998), Tustison et al. (2010), $\mathbf{c}$ is defined in the log-transformed bias field space for numerical stability during optimization. A smooth bias field is obtained by first interpolating a grid of regularly spaced control points $\mathbf{c}$ with a third-order B-spline smoothing (Gallier and Gallier, 2000) and then taking the exponential for value recovering: $\boldsymbol{\Phi}_{\text{bias}}(\mathbf{c}) = \exp(B(\mathbf{c}))$. Here, B represents the interpolation function with uniform B-splines for smoothing. When $\mathbf{c} = \mathbf{0}$, $\boldsymbol{\Phi}_{\text{bias}} = \mathbf{1}$ (identity field).

The magnitude constraint for the bias field perturbation C_{bias} is defined as:

$$\forall (x, y) \in \mathbb{R}^2, \|\boldsymbol{\Phi}_{\text{bias}}(\mathbf{c}) - \mathbf{1}\|_\infty \leq \epsilon_{\text{bias}}, \tag{7}$$

where $\epsilon_{\text{bias}}(\epsilon_{\text{bias}} \geq 0)$ is a scalar to control the maximum level of corruption caused by the bias field.

3.3.3. Image geometric transformation model $\mathcal{T}_{\text{affine}}$

To model image-level geometric variations we use an affine transformation to transform images. This process is defined as:

$$\mathbf{x}_{\text{affine}} = \mathcal{T}_{\text{affine}}(\mathbf{x}; \mathbf{a}) = \mathcal{T}_{\text{affine}}(\mathbf{x}; t_x, t_y, r, s_x, s_y), \tag{8}$$

where $\mathbf{a}$ contains five affine parameters $[t_x, t_y, r, s_x, s_y]$ to characterize translation T, rotation R and scaling S operations which are performed in succession, see Fig. 2(c). Given a 2D image $\mathbf{x}$ and the affine parameters $\mathbf{a}$, a pixel $\mathbf{x}(u, v)$ at position (u, v) in the original image $\mathbf{x}$ is transformed to a new position (u', v') via the following matrix multiplications:

$$\begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = (T \cdot R \cdot S) \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}, \tag{9}$$

where $T = \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix}$, $R = \begin{bmatrix} \cos r\pi & -\sin r\pi & 0 \\ \sin r\pi & \cos r\pi & 0 \\ 0 & 0 & 1 \end{bmatrix}$, $S = \begin{bmatrix} 1+s_x & 0 & 0 \\ 0 & 1+s_y & 0 \\ 0 & 0 & 1 \end{bmatrix}$. We use a normalized Cartesian coordinate system centered at (0,0) to specify each pixel's location (u, v). Each location is normalized by the input spatial dimensions so that its value lies in $[-1, 1]$: $-1 \leq u \leq 1, -1 \leq v \leq 1$. Each transformation parameter is restricted in a user-defined range to control the range of the spatial transformations. The constraint for the affine transformation model C_{affine} can be described as:

$$-\epsilon_{a^i} \leq a^i \leq \epsilon_{a^i}; \forall a^i \in [t_x, t_y, r, s_x, s_y]. \tag{10}$$

3.3.4. Image deformation model $\mathcal{T}_{\text{morph}}$

To introduce intra- and inter-subject anatomical variations, we would like to construct a generator that can produce a smooth spatial transformation to deform the image, while preserving its topology and spatial layout. To achieve the goal, we model these variations using invertible and differentiable diffeomorphic transformations (Vercauteren et al., 2009). Specifically, following previous works on diffeomorphic image registration, e.g., Demons (Vercauteren et al., 2009), we parameterize the deformation $\boldsymbol{\Phi}_{\text{morph}}$ using an underlying stationary velocity field $\mathbf{v}$ that $\frac{\partial \boldsymbol{\Phi}_{\text{morph}}(t)}{\partial t} = \mathbf{v}(\boldsymbol{\Phi}^{(t)}_{morph})$, where $\boldsymbol{\Phi}^{(t)}_{morph}$ represents the deformation at time t. The final deformation $\boldsymbol{\Phi}_{\text{morph}}$ is obtained by starting with an identity transform $\boldsymbol{\Phi}^{(0)}_{morph} = Id$ and integrating the stationary velocity field $\mathbf{v}$ over $t \in [0, 1]$.[2] As shown in Fig. 2(d), given

[2] We employ the *scaling and squaring* (Arsigny et al., 2006) to approximate the integration to accelerate the computation as a common practice (Balakrishnan et al., 2019; Arsigny et al., 2006; Vercauteren et al., 2009).

a 2D image $\mathbf{x}$ and a 2-dimensional velocity field $\mathbf{v}$, the deformed image is obtained using the following function:

$$\mathbf{x}_{\text{morph}} = \mathcal{T}_{\text{morph}}(\mathbf{x};\mathbf{v}) = \mathbf{x}\circ\Phi_{\text{morph}} = \mathbf{x}\circ\int_{t=0}^{1}\mathbf{v}(\Phi_{\text{morph}}^{(t)})\,dt. \tag{11}$$

Here $\mathbf{x}\circ\Phi_{\text{morph}}$ represents $\mathbf{x}$ warped by a deformation field $\Phi_{\text{morph}} \in \mathbb{R}^{H\times W\times 2}$. To initialize the velocity field $\mathbf{v}$, we sample a random, low-resolution tensor $\mathbf{v}' \in \mathbb{R}^{\frac{H}{ds}\times\frac{W}{ds}\times 2}$ $(ds \geq 1)$.[3] We impose a magnitude constraint C_{morph} to $\mathbf{v}'$ to control the level of deformation:

$$\|\mathbf{v}'\|_2 \leq \epsilon_{\text{morph}}. \tag{12}$$

This is achieved by directly applying L2 norm to $\mathbf{v}'$ and re-scaling it to find an approximate solution in the constrained space: $\mathbf{v}' \leftarrow \epsilon_{\text{morph}}\frac{\mathbf{v}'}{\|\mathbf{v}'\|_2}$ (Miyato et al., 2018). To further encourage the spatial smoothness of the deformation, following the related work on the diffeomorphic demons (Vercauteren et al., 2009), we apply Gaussian smoothing K_{smooth} to the velocity field: $\mathbf{v}' \leftarrow K_{\text{smooth}}(\mathbf{v}')$ as well as to the integrated deformation: $\Phi_{\text{smooth}} \leftarrow K_{\text{smooth}}(\Phi_{\text{morph}})$. In this way, we ensure the deformation is smooth and diffeomorphic without introducing additional smoothness regularization terms, simplifying the optimization procedure (Cachier et al., 2003). In the experiments, we used a small Gaussian kernel K_{smooth} with $\sigma_{smooth} = 1$, as suggested by Vercauteren et al. (2009).

3.4. Consistency loss function $\mathcal{R}$

3.4.1. Consistency loss function for photometric transformations

For photometric transformations, i.e. $\mathcal{T}_{\text{bias}}$, $\mathcal{T}_{\text{noise}}$, we directly use a composite distance loss function $\mathcal{D}$ to compute the consistency regularization term $\mathcal{R}$ computed on the original probabilistic prediction $f_\theta(\mathbf{x})$ and perturbed prediction $f_\theta(\mathcal{T}(\mathbf{x}))$:

$$\mathcal{R}_{\mathcal{D}}(\mathbf{x}; f_\theta, \mathcal{T}) = \mathcal{D}(f_\theta(\mathbf{x}), f_\theta(\mathcal{T}(\mathbf{x};\mathbf{t}))). \tag{13}$$

The composite distance function $\mathcal{D}$ measures two predictions $\mathbf{p}, \mathbf{p}'$ in the same image coordinates, which is defined as follows:

$$\begin{aligned} &\mathcal{D}(\mathbf{p},\mathbf{p}') = \mathcal{L}_{MSE}(\mathbf{p},\mathbf{p}') + w\mathcal{L}_{Contour}(\mathbf{p},\mathbf{p}'),\\ &\mathcal{L}_{MSE}(\mathbf{p},\mathbf{p}') = \|\mathbf{p}-\mathbf{p}'\|_2^2,\\ &\mathcal{L}_{Contour} = \sum_{c\neq BG}^{C}\sum_{S\in S_x,S_y}\|S(\mathbf{p}^{(c)}) - S(\mathbf{p}'^{(c)})\|_2^2. \end{aligned} \tag{14}$$

Here we adopt the mean-squared-error loss $\mathcal{L}_{MSE}$ to measure pixel-wise differences, as a common practice in consistency regularization related works (Tarvainen and Valpola, 2017; Li et al., 2020; Cui et al., 2019; Berthelot et al., 2019). In addition, we employ a contour-based loss function $\mathcal{L}_{Contour}$ to better capture the difference on the foreground objects' boundaries between two predictions (Chen et al., 2019b). S_x, S_y represent Sobel filters in the x- and y-directions, which are used to extract object boundaries from the model's probabilistic map for every class c except the background (BG) class. w is a weight that controls the relative importance of two terms. In our experiments, we empirically set it to 0.5. We believe that combining pixel-wise and contour-based loss terms can help the network to better capture the semantic dissimilarity between two predicted segmentation maps.

3.4.2. Consistency loss function for geometric transformations

For geometric transformations, i.e., $\mathcal{T}_{\text{morph}}$, $\mathcal{T}_{\text{affine}}$, Eq. (13) is not directly applicable as the position and/or structural information of target objects also changes accordingly. We, therefore, transform the perturbed prediction back to the coordinates of the original image accordingly before computing the consistency loss. The regularization loss is defined as:

$$\mathcal{R}_{\mathcal{D}}(\mathbf{x}; f_\theta, \mathcal{T}) = \mathcal{D}(f_\theta(\mathbf{x}), \mathcal{T}_{\mathbf{t}}^{-1}(f_\theta(\mathcal{T}(\mathbf{x};\mathbf{t})))). \tag{15}$$

Here $\mathcal{T}_{\mathbf{t}}^{-1}$ denote the inverse transformation for $\mathcal{T}(\cdot;\mathbf{t})$. The inverse transformations for the two types of geometric transformations are easy to compute:

- **Inverse affine transformation** $\mathcal{T}_{\text{affine}}^{-1}$: Since the affine transformation is parameterized by a homogeneous transformation matrix, its inverse transformation matrix can be directly computed via: $\mathcal{T}_{\text{affine}}^{-1} = S^{-1}R^{-1}T^{-1}$;
- **Inverse deformation** $\mathcal{T}_{\text{morph}}^{-1}$: The inverse deformation Φ_{morph}^{-1} is obtained by integrating the *negative* velocity field $(-\mathbf{v})$ backward: $\mathcal{T}_{\text{morph}}^{-1} = \int_{t=0}^{-1}(-\mathbf{v})(\Phi^{(t)})dt$ (Ashburner, 2007).

3.4.3. Consistency loss function for a chained transformation

For a chained transformation $\mathcal{T}_{1\circ 2\circ\ldots\circ m} : \mathcal{T}_1\circ\mathcal{T}_2\circ\ldots\mathcal{T}_m$ including both photometric and geometric transformations, we employ Eqs. (13) and (15) to compute the consistency loss between the original prediction and the perturbed prediction:

$$\mathcal{R}_{\mathcal{D}}^{chain} = \mathcal{R}_{\mathcal{D}}(\mathbf{x}; f_\theta, \mathcal{T}_{1\circ 2\circ\ldots\circ m}). \tag{16}$$

This means one needs to transform the perturbed prediction back to the coordinates of the original image if there is any geometric transformation involved. For instance, given a chain of transformation functions $\mathcal{T}_{\text{affine}\circ\text{noise}} : \mathcal{T}_{\text{affine}}\circ\mathcal{T}_{\text{noise}}$, the loss function is defined as follows:

$$\mathcal{R}_{\mathcal{D}}^{chain}(\mathbf{x}; f_\theta, \mathcal{T}_{\text{affine}\circ\text{noise}}) = \mathcal{D}(\mathbf{p}, \mathcal{T}_{\text{affine}}^{-1}(\mathbf{p}')) \tag{17}$$

where $\mathbf{p} = f_\theta(\mathbf{x}), \mathbf{p}' = f_\theta(\mathcal{T}_{\text{affine}\circ\text{noise}}(\mathbf{x}))$.

4. Experiments settings

4.1. Datasets

4.1.1. Cardiac MR dataset

The cardiac dataset is provided by The Automated Cardiac Diagnosis Challenge (ACDC) (Bernard et al., 2018),[4] which is a public dataset for cardiac MR image segmentation. The left ventricular cavity (LV), the left ventricular myocardium (MYO), and the right ventricular cavity (RV) in end-diastolic and end-systolic frames were manually labeled by experts. The original in-plane pixel spacing ranges from 1.37×1.37 mm^2 to 1.68×1.68 mm^2.

We preprocessed images to have the same in-plane pixel spacing: 1.37×1.37 mm^2, following Chaitanya et al. (2019). After that, all images were centrally cropped to 192×192 in order to save computational cost. We used the same data setting as in Chaitanya et al. (2019), splitting the dataset (100 subjects in total) into 4 subsets: an unlabeled set for semi-supervised learning (M=25), a validation set (5 subjects), and a test set (20 subjects). The rest 50 subjects were used as the labeled training pool. We selected N subjects from the rest to form a labeled set for training, simulating a low-data learning regime. Specifically, we evaluated one-shot learning (N=1) and three-shot learning (N=3) in both supervised (using the labeled set only) and semi-supervised (using both labeled and unlabeled sets) settings. We also trained the segmentation with different numbers of labeled subjects from the pool (N=10, N=25) to test the performance improvements against different settings. In all settings, we trained the network five times, each time with a different, randomly selected labeled set to alleviate the dataset selection bias, and reported the mean performance.

[3] We apply bilinear upsampling to $\mathbf{v}'$ to obtain $\mathbf{v}$.

[4] https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html

4.1.2. Prostate MR dataset

The prostate dataset is provided by the Medical Segmentation Decathlon Challenge (Antonelli et al., 2022),[5] which consists of 32 subjects. The peripheral zone (PZ) and the central zone (CZ) of the prostate have been manually labeled and verified by an expert human rater. We performed segmentation on T2 images, where all images have been resampled to have the same pixel spacing (0.625×0.625 mm^2, the median value of pixel spacings in this dataset) and then centrally cropped to 224×224 to reduce computational cost.

To train and evaluate the proposed method, we split the dataset into 22/4/6 for training/validation/testing. The training set was further divided into two subsets (11 subjects each). We randomly selected N (with $N \leq 11$) subjects from the first set to form a small labeled set while all subjects in the second one were used to construct the unlabeled set ($M = 11$) for semi-supervised learning. We trained the network three times, each time with a different, randomly selected labeled set and reported the mean performance.

4.2. Implementation details

4.2.1. Default data augmentation

For all experiments, we applied a random data augmentation pipeline as a default setting. This augmentation pipeline includes random affine transformation (i.e. scaling, rotation, translation), image flipping, random global intensity transformation (brightness and contrast), and elastic transformation. Detailed configurations of these random transformations can be found in Chaitanya et al. (2019). After random data augmentation, the image intensity was rescaled to $[0, 1]$.

4.2.2. Training details

The proposed method is independent of the network structures. For ease of comparison, we adopted the commonly-used 2D U-net (Ronneberger et al., 2015) as our segmentation network, which has been demonstrated its superiority across various medical image segmentation datasets (Isensee et al., 2021). The Adam optimizer was used to update network parameters with a batch size of 20. To accelerate training, we first trained the network with the default data augmentation for 1,000 epochs (learning rate=10^{-3}) to get a pre-trained segmentation network. We then fine-tuned the network with the proposed data augmentation method. We applied a small learning rate (10^{-5}) and trained the network for 600 epochs. Following previous work on semi-supervised learning (Berthelot et al., 2019; Tarvainen and Valpola, 2017; Li et al., 2020), we evaluated the segmentation model using an exponential moving average of model parameters with a decay rate of 0.999.

Loss configuration: For the supervised loss $\mathcal{L}_s$, we used a combination of a weighted cross entropy loss function and a soft Dice loss function (Baumgartner et al., 2018), to alleviate the class imbalance problem in data. Empirically, for cardiac segmentation, class weights for background (BG), LV, MYO, RV were set to 0.01:0.33:0.33:0.33 respectively to give equal weights to foreground classes; for prostate segmentation, weights for BG, PZ, CZ were set to 0.01:0.66:0.33. We set a higher weight to the PZ class, as it has significantly fewer number of pixels in each image, compared to the CZ class. Regarding the consistency regularization loss $\mathcal{R}$, the weighting parameters for the contour loss term w was empirically set to 0.5. Since the prediction for images $f_\theta(\mathbf{x})$ can be very noisy at the beginning of the training, it can produce incorrect supervision signals to misguide the training. This is a common issue in consistency-regularized methods Li et al. (2020), Berthelot et al. (2019). Similar to Li et al. (2020), Berthelot et al. (2019), we started the training with a small weight λ for $\mathcal{R}$ and slowly increased it in the first e_{ramp} epochs until it reached to its maximum λ^{max}. The value of λ was linearly increased with the number of training epochs: $\lambda = \min(\lambda^{max} \times e/e_{ramp}, \lambda^{max})$. e is the number of the current epoch and $e_{ramp} = 200$. We empirically set $\lambda^{max} = 1.0$ so that in the later stage the supervised loss $\mathcal{L}_s$ and the consistency regularization loss $\mathcal{R}$ share the same weight to balance the training. Ideally, the two losses are expected to be zero when the network reaches its optimum, suggesting that the network not only produces correct segmentations on original images but also produces consistent predictions on adversarially augmented images.

Table 1
Configurations for image transformations.

Transformation	Parameter constraints
$\mathcal{T}_{\text{noise}}$	noise $\parallel \mathbf{r} \parallel_2 \leq \epsilon_{\text{noise}} = 1$
$\mathcal{T}_{\text{bias}}$	Control points $\mathbf{c} \in \mathbb{R}^{b \times b}$; $b = 4, \epsilon_{\text{bias}} = 0.3$
$\mathcal{T}_{\text{affine}}$	Translation: $-0.1 \leq t_x, t_y \leq 0.1$ Rotation: $-\frac{30^\circ}{180^\circ} \leq r \leq \frac{30^\circ}{180^\circ}$ Scaling: $-0.2 \leq s_x, s_y \leq 0.2$
$\mathcal{T}_{\text{morph}}$	$\mathbf{v}' \in \mathbb{R}^{\frac{H}{ds} \times \frac{W}{ds} \times 2}$, $ds = 16$, $\lVert \mathbf{v}' \rVert_2 \leq \epsilon_{\text{morph}} = 1.5$

Adversarial data augmentation configuration: For bias field construction, we adopted the B-spline convolution kernel provided by the AirLab library to interpolate the control points (Sandkühler et al., 2018). To ensure the generated variations are realistic, one has to specify the magnitude constraints for each transformation. For simplicity, one can reuse the magnitude constraints specified in existing hand-crafted random data augmentation works or empirically set up the range based on visual inspection, which is a common practice in most data augmentation frameworks. Specifically, in this work, the constraints for noise, bias field, rotation, and translation were directly taken from our previous work (Chen et al., 2020) and RandAugment (Cubuk et al., 2020), whereas for our proposed novel image deformation model, the velocity magnitude constraint was set based on visual inspection. We provide an interactive Jupyter notebook for readers' interest to visualize the augmented images with different transformation configurations in our code repository.[6] Table 1 lists the detailed configurations for the employed transformations in our work. The probability of selecting each transformation p is set to 0.5 for simplicity. For the optimization of the underlying transformation parameters, we first randomly sampled the parameters from the specified range for initialization and then applied projected gradient descent to ensure the updated transformations are still within the search space. Specifically, for noise $\mathbf{r}$ and velocity fields $\mathbf{v}$ with l^2 norm constraints, similar to (Miyato et al., 2018), we normalized and re-scaled the updated parameters to meet the magnitude constraints specified in Eq. (5) and Eq. (12), respectively. For the bias field, we clipped the values of generated bias field to meet the criterion specified in Eq. (7). For affine transformation, we applied the element-wise HardTanh activation function to the transformation parameters (rotation, translation, scaling) and re-scaled them to meet the criterion specified in Eq. (10). We used the same step size ($\alpha_i = 1$) and performed only a one-step ($k = 1$) search for simplicity and training efficiency, similar to Miyato et al. (2018). Detailed implementation can be found in our code repository.[7] We used the same configuration for both cardiac and prostate segmentation tasks to test the generality. Results show that it can yield substantial improvements for both applications. The full code implementation for *AdvChain* is based on PyTorch and is available at GitHub.[8] All experiments were performed on an Nvidia® GeForce® 2080 Ti.

[5] http://medicaldecathlon.com/

Table 2
Comparison between the proposed method (*AdvChain*) against high-performing consistency regularized semi-supervised learning methods on the cardiac, and prostate segmentation datasets.

Task	Cardiac								Prostate					
Dataset setting	N = 1, M = 25				N = 3, M = 25				N = 3, M = 11			N = 11, M = 11		
Method	LV	MYO	RV	AVG	LV	MYO	RV	AVG	PZ	CZ	AVG	PZ	CZ	AVG
Pretrained	0.5155	0.4290	0.2201	0.3882 (0.2353)	0.8269	0.7905	0.6288	0.7487 (0.1154)	0.3897	0.7075	0.5486 (0.1034)	0.5077	0.8019	0.6548 (0.0934)
MixMatch (Berthelot et al., 2019)	0.6611	0.5415	0.3547	0.5191 (0.1913)	0.8406	0.8203	0.6849	0.7819 (0.1050)	0.4770	0.7505	0.6137 (0.0701)	0.5893	0.8220	0.7057 (0.0566)
FixMatch (Sohn et al., 2020)	0.6437	0.5496	0.3666	0.5200 (0.1675)	0.8370	0.8119	0.6461	0.7650 (0.1029)	0.4243	0.7327	0.5785 (0.0756)	0.5439	0.8107	0.6773 (0.0726)
TCSM (Li et al., 2020)	0.6391	0.5491	0.3369	0.5084 (0.2283)	0.8442	0.8179	0.6816	0.7812 (0.1047)	0.4740	0.7512	0.6126 (0.0753)	0.5986	0.8248	0.7117 (0.0613)
VAT (Miyato et al., 2018)	0.6729	0.5228	0.3400	0.5119 (0.2218)	0.8598	0.8353	0.6646	0.7866 (0.0999)	0.4571	0.7565	0.6068 (0.0763)	0.5436	0.8106	0.6771 (0.0648)
FixMatch+VAT (Wang et al., 2021a)	0.6675	0.5778	0.3715	0.5389 (0.1666)	0.8396	0.8121	0.6504	0.7674 (0.1006)	0.4254	0.7175	0.5715 (0.0886)	0.5576	0.8226	0.6901 (0.0642)
AdvChain (proposed)	**0.7151**	**0.6369**	**0.4064**	**0.5861 (0.1939)**	**0.8708**	**0.8469**	**0.7072**	**0.8083 (0.0849)**	**0.5243**	**0.7742**	**0.6492 (0.0789)**	**0.6245**	**0.8405**	**0.7325 (0.0474)**
Upperbound[a]	0.8963	0.8553	0.7419	0.8312 (0.0730)	0.8951	0.8627	0.7654	0.8411 (0.0600)	0.5930	0.7960	0.6945 (0.0666)	0.6298	0.8280	0.7288 (0.0648)

Reported values are mean Dice scores for each class. We also report the mean and standard deviation of average Dice scores over all foreground classes (AVG) for each task. N: # of labeled images, M: # of unlabeled images. LV: left ventricle; MYO: left ventricular myocardium; RV: right ventricle; PZ: peripheral zone; CZ: central zone.

[a]Upperbound performance of the segmentation network (U-net) when trained using labeled images (N + M) from both the labeled set and the unlabeled set.

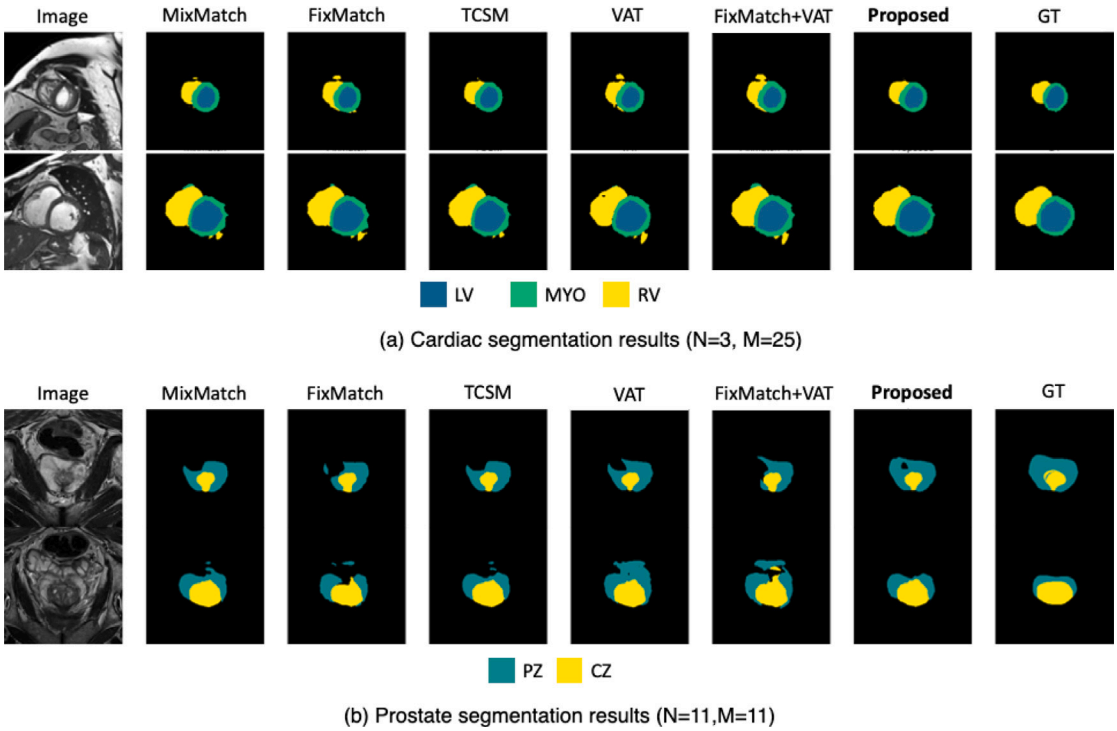

Fig. 3. Visualization of the segmentation results on the (a) cardiac and (b) prostate test data with U-net trained using different consistency regularization-based semi-supervised methods. The proposed method (*AdvChain*) consistently outperforms the other competing methods in both tasks, producing more anatomically correct segmentation results. GT: manual labels.

5. Results

5.1. Comparison study

We compared our method (*AdvChain*) to several high-performing consistency-regularization-based semi-supervised methods powered by different data augmentation techniques, which are mostly related to ours:

- MixMatch (Berthelot et al., 2019)[9] is a semi-supervised learning method based on Mixup (Zhang et al., 2018) Mixmatch performs linear interpolation to mix both labeled examples and unlabeled examples to get augmented image-label pairs;
- FixMatch (Sohn et al., 2020)[10] enhances regularization by enforcing the prediction consistency between weakly augmented images (i.e., flip, shifts) and strongly augmented images with RandAugment (Cubuk et al., 2020);
- TCSM (Li et al., 2020)[11] is an enhanced Mean-Teacher based semi-supervised learning method (Cui et al., 2019) It enhances the consistency regularization by extending the noise perturbation with *random geometric transformations* including scaling and rotation;
- VAT (Miyato et al., 2018)[12] is an adversarial noise-based semi-supervised learning method. Unlike the proposed method, it injects only adversarial noise to clean data for consistency regularization and use confidence thresholding to obtain pseudo labels for reliable regularization.
- FixMatch+VAT (Wang et al., 2021a) is a semi-supervised learning method combining FixMatch (Sohn et al., 2020) and VAT (Miyato et al., 2018), which has achieved state-of-the-art performance in some large-scale medical image classification tasks.

[6] https://github.com/cherise215/advchain/tree/master/example
[7] https://github.com/cherise215/advchain/tree/master/advchain/augmentor
[8] https://github.com/cherise215/advchain
[9] https://github.com/google-research/mixmatch
[10] https://github.com/google-research/fixmatch
[11] https://github.com/xmengli999/TCSM
[12] https://github.com/takerum/vat_tf

For all methods, we adopted their official implementation[13] and trained the same network with the same training setup (e.g. using the same pre-trained models) for fair comparison. Quantitative results in Table 2 and qualitative results in Fig. 3 show that the proposed approach achieves the highest performance on the two segmentation tasks. Surprisingly, one interesting finding from Table 2 is that when we have the same number of labeled and unlabeled images (i.e. N=11, M=11) to train the prostate segmentation network, the proposed method even exceeds the upper-bound performance, e.g., 0.7325 vs 0.7288 in terms of average Dice score. This may be due to the presence of noisy labels in the prostate dataset (see the top-right block in Fig. 3 for reference), which can affect the learning in a fully supervised setting. As semi-supervised learning does not fully rely on manual labels on the training dataset, it can therefore be more robust against noisy labels.

We further compared our data augmentation with the state-of-the-art random composite data augmentation method: RandAugment (Cubuk et al., 2020) adopted in FixMatch (Sohn et al., 2020), which employs a wide collection of image transformations including color inversion, translation, contrast adjustment.[14] As shown in Fig. 4 when we replaced our proposed data augmentation with RandAugment in our consistency-regularized method, the segmentation performance declines with lower average Dice scores on the cardiac segmentation task. The segmentation performance on the prostate segmentation tasks does not significantly outperform ours although RandAugment employs a larger number of image processing functions (*autoContrast, equalize, solarize, color, posterize, contrast, brightness, sharpness, rotation, translation and shearing* (Cubuk et al., 2020)).

We found that compared to *AdvChain*, RandAugment focuses more on modifying the style of images. The geometric variations are quite limited compared to ours. In fact, RandAugment only considers basic spatial augmentation operations (e.g., rotation, translation) without applying any local deformations (Cubuk et al., 2020). Such a limitation is also shared in the other semi-supervised learning frameworks compared in our study, such as TCSM (Li et al., 2020). By contrast, our proposed method supports to generate diffeomorphic transformations to account for realistic morphological variations. The generated transformations are reversible, which allows to measure the prediction inconsistency in the original image space for ease of optimization. With adversarial training, *AdvChain* observes the image content and takes the segmentation model's prediction to identify and deform the local structures of interest with increased variations (see Fig. 5), which helps the segmentation model to better generalize across different populations with varied morphology. Of note, different from *AdvChain*, RandAugment does not support adversarial training to optimize the transformation parameters as the underlying image transformation functions in Python Image Library (PIL)[15] do not support automatic differentiation.

We attribute the efficacy of our method against other baseline methods mainly to (1) the increased data diversity and effectiveness with adversarial training applied to dynamic image transformations, see Sections 5.2.1 and 5.2.2 and (2) the enhanced consistency regularization loss with a contour-based loss term, see Section 5.2.3.

5.2. Ablation studies

5.2.1. Effects of adversarial training w/ individual augmentation and dynamically chained composite data augmentation

To understand the effects of individual data augmentations and the importance of the chain strategy for improved data diversity used in *AdvChain*, we investigate the performance of our framework when applying augmentations individually with a fixed type of transformation or with dynamic augmentation chains. Specifically, we trained the same network with each of the four transformations separately in the semi-supervised setting, and compared the results to their random counterparts (i.e. without adversarial training). Table 3 shows the obtained results.

In Table 3, we observe that individual data augmentations with adversarial training consistently outperform those corresponding ones without adversarial training, see row 1 vs. row 2, row 3 vs. row 4. Another finding is that adversarial training with morphological transformations always outperforms the other three individual data augmentations, highlighting the importance of introducing local anatomical variations to enhance the data variety.

In both cardiac and prostate segmentation tasks, the proposed composite adversarial augmentation (*AdvChain*) achieves the highest Dice scores on both tasks. By contrast, the random-based composite data augmentation: *chain* w/o adversarial training does not always outperform other random individual data augmentations. For example, on the cardiac segmentation performance, the average Dice score slightly drops from 0.7857 to 0.7802, compared to the one with random bias field augmentation. This highlights the importance of applying adversarial training to optimizing dynamically chained transformations, which increases both the *diversity* and *effectiveness* of augmented data points to improve the network generalization for the downstream tasks.

5.2.2. The generality of AdvChain with different chained transformations of varied lengths and different orders

To verify the generality of *AdvChain* with different types of chained transformations, we apply *AdvChain* to optimizing chained transformations of different lengths, i.e. the maximum number of sampled transformations in a chain are fixed to a certain number (1/2/3/4) during the course of training in each experiment. Fig. 6 plots the results on the cardiac test set using the same data setting (N=3, M=25) and prostate test set using the similar setting (N=3, M=11), respectively. We also plot the results with its downgraded variant (*RandChain*), i.e., *AdvChain* without applying the adversarial optimization on the transformation parameters for comparison. Results show that *AdvChain* consistently provides segmentation performance improvements regardless the change of chained lengths.

We also applied *AdvChain* to optimizing the same set of transformations but chained in different orders. Since there are 24 different arrangements with the four transformations, we selected the two most common ones as a proof of concept: (a) $T_{affine \circ morph \circ bias \circ noise}$: first apply photo-metric transformations and then geometric transformations from local perturbations to global perturbations, similar to Chen et al. (2019a), Zhao et al. (2019); (b) $T_{noise \circ bias \circ morph \circ affine}$: the one chained in the opposite direction. Results shown in Fig. 7 confirm that *AdvChain* provides consistent improvements with composite transformations chained in different orders, indicating the generality of *AdvChain* with composite transformations chained in random orders for improved segmentation model performance.

Fig. 8 visualizes the data augmentation optimization process with the chain $T_{affine \circ morph \circ bias \circ noise}$ and network predictions before and after augmentations. From Fig. 8, we can clearly see that after applying adversarial optimization, the optimized data augmentations (see the bottom row in each block) are more effective at perturbing network predictions compared to those with random initialization (the top row in each block). This is because adversarial data augmentation takes both model information and image content into account to augment images, which produces more informative, challenging samples to regularize the network. It is particularly evident when we compare adversarial noise ($\mathbf{r}^{adv}$) and adversarial deformation ($\boldsymbol{\Phi}^{adv}_{\text{morph}}$) to their random initialized counterparts ($\mathbf{r}^{rand}, \boldsymbol{\Phi}^{rand}_{\text{morph}}$). We can see that adversarial data augmentation can identify and focus more on attacking/deforming local target structures in images to fool the network to

[13] For VAT, MixMatch, FixMatch, we re-implemented them in PyTorch as the original code repositories are based on Tensorflow.

[14] We adopted the implementation of RandAugment provided in the official PyTorch website with its recommended set-up: https://pytorch.org/vision/stable/generated/torchvision.transforms.RandAugment.html.

[15] Python Image Library: https://pillow.readthedocs.io/en/stable/.

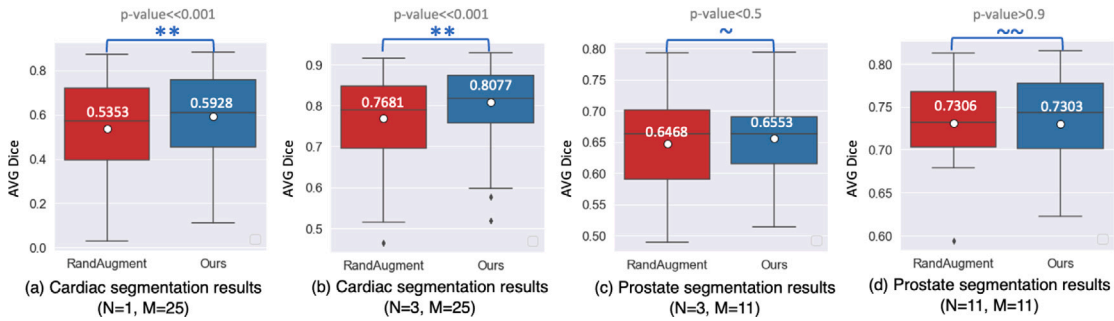

Fig. 4. Boxplots of average Dice scores evaluated on the (a, b) cardiac and (c, d) prostate test sets using different composite data augmentation methods for consistency regularization: RandAugment (Cubuk et al., 2020) and ours: *AdvChain*. For ease of comparison, in *AdvChain*, we limit the maximum number of selected transformations to 2, making it aligned with the recommended set-up in RandAugment. We trained the networks using different numbers of labeled (N) and unlabeled images (M). Compared to RandAugment, our proposed method *AdvChain* achieves higher average Dice scores (see white numbers) in most cases even with a smaller set of transformations, especially when labeled data is extremely limited.

Table 3
Random vs adversarial data augmentation with individual image transformations and chained transformations for consistency regularization.

Cardiac					
Adversarial training	Noise	Bias	Affine	Morph	Chain
✗	0.7706 (0.1066)	0.7857 (0.0994)	0.7704 (0.1123)	0.7836 (0.0940)	0.7802 (0.1035)
✓	0.7864 (0.0976)	0.7955 (0.1013)	0.7885 (0.1033)	0.8014 (0.0861)	**0.8083 (0.0849)**
Prostate					
Adversarial training	Noise	Bias	Affine	Morph	Chain
✗	0.5880 (0.1016)	0.6093 (0.0817)	0.6268 (0.0912)	0.6104 (0.0813)	0.6270 (0.0872)
✓	0.6211 (0.0951)	0.6123 (0.0811)	0.6294 (0.0897)	0.6408 (0.0784)	**0.6492 (0.0789)**

Experiments were performed on cardiac segmentation and prostate segmentation datasets in the semi-supervised setting. For both tasks, we use only three labeled subjects. Reported values are mean (std) of average Dice scores across foreground classes over multiple runs (5 runs for cardiac, 3 runs for prostate).

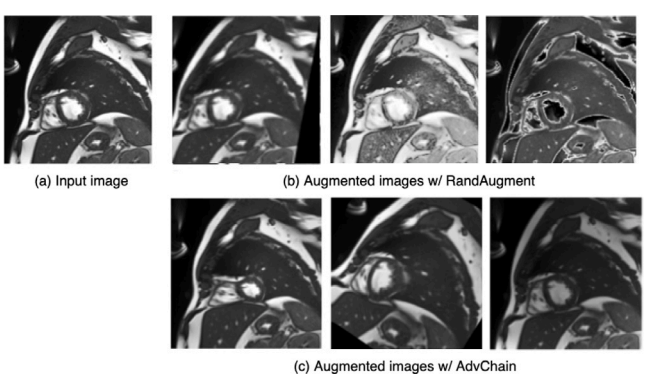

Fig. 5. An input image (a) and (b) augmented images generated using RandAugment (Cubuk et al., 2020) and (c) augmented images using our proposed adversarial data augmentation *AdvChain*.

make inconsistent predictions. Augmenting images with these adversarial transformations contribute to stronger consistency regularization to enforce the network to be invariant under photometric transformations and equivariant under geometric transformations.

In Fig. 7, it is interesting to notice that *AdvChain* with $T_{noise \circ bias \circ morpho affine}$ yields slightly better performance compared to the one with $T_{affine \circ morpho bias \circ noise}$ on the cardiac segmentation task. For the prostate segmentation, *AdvChain* w/ $T_{affine \circ morpho bias}$ achieves higher segmentation. Similarity, *AdvChain* achieves slightly better performance when the maximum chained length is fixed to 1 for cardiac segmentation and 2 for prostate segmentation, respectively, as shown in Fig. 6. We also found that the optimum maximum chain length depends on not only the task but also the selection of the training set, see Fig. A.1 in the appendix. It is possible that better performance

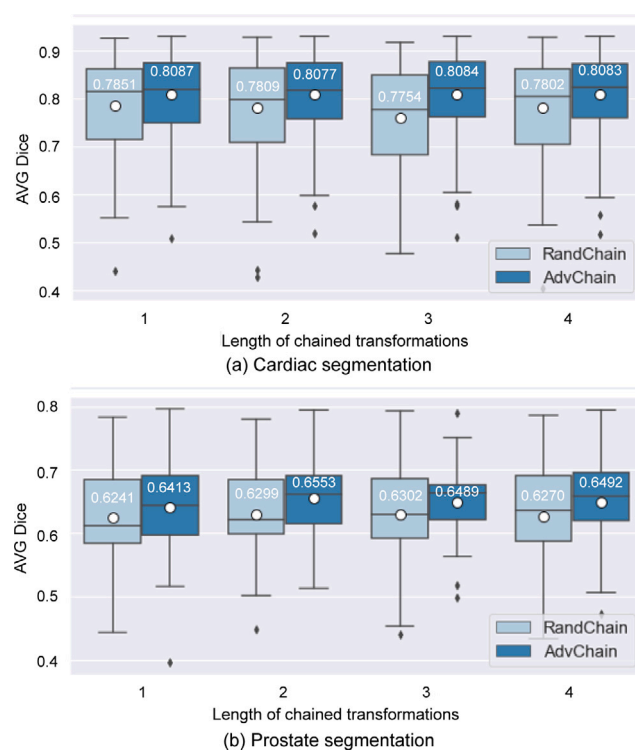

Fig. 6. Boxplots of average Dice scores between the results of *RandChain* and *AdvChain* for cardiac segmentation (a) and prostate segmentation (b). *AdvChain* provides consistent improvements with chained transformations of different maximum lengths (1–4). White circles with numbers show the mean value of Dice scores across the test tests.

Table 4
Effect of different distance functions for consistency loss function $\mathcal{R}$.

Consistency loss functions	Cardiac (N = 3, M = 25)				Prostate (N = 3, M = 11)		
	LV	MYO	RV	AVG	PZ	CZ	AVG
$\mathcal{R}_{\mathcal{D}_{KL}}$	0.8635	0.8429	0.6747	0.7937	0.5053	0.7663	0.6358
$\mathcal{R}_{\mathcal{D}_{KL+Contour}}$	0.8655	0.8445	0.6744	0.7948	0.4989	0.7729	0.6359
$\mathcal{R}_{\mathcal{D}_{MSE}}$	0.8660	0.8450	0.7043	0.8051	0.5156	**0.7744**	0.6450
$\mathcal{R}_{\mathcal{D}_{MSE+Contour}}$ (proposed)	**0.8708**	**0.8469**	**0.7072**	**0.8083**	**0.5243**	0.7742	**0.6492**

Reported scores are average Dice scores over segmented structures. We also report the mean Dice scores over all foreground classes (AVG) for each task. N: number of labeled subjects, M: number of unlabeled subjects.

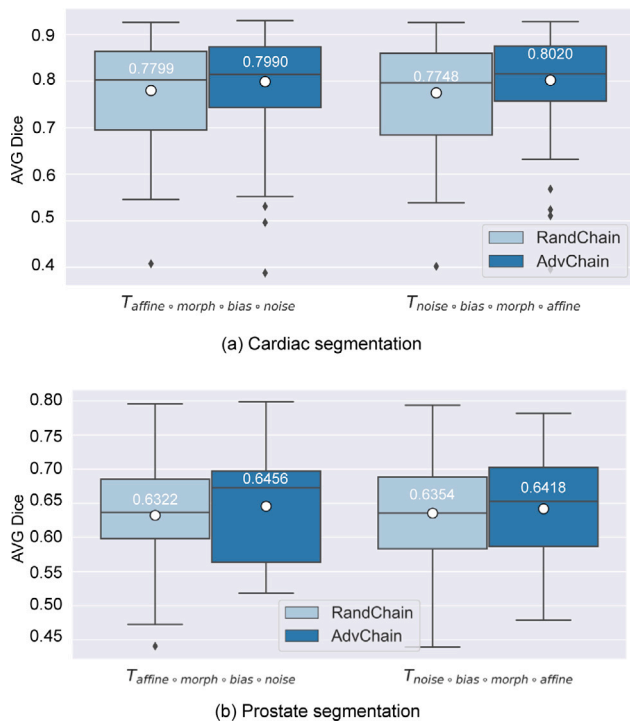

Fig. 7. Boxplots of average Dice scores between the results of *RandChain* and *AdvChain* with transformations chained in different orders. Compared to *RandChain*, *AdvChain* boosts the segmentation performance with improved mean Dice scores and reduced outliers. White circles with numbers show the mean value of Dice scores across the test tests.

can be further achieved by identifying the optimum maximum length and the optimum arrangement (taking the validation set performance into account) to improve segmentation performance for a *specific* task. Yet since the search space can be extremely large and the policy optimization requires extraordinary high computational costs (Cubuk et al., 2019), we randomly generate arbitrarily chained transformations to explore all different kinds of possibilities as a trade-off between efficiency and effectiveness for general segmentation tasks.

5.2.3. Effect of different consistency loss functions

We further compared the proposed with the other three different distance functions to highlight the superiority of the proposed inconsistency regularization $\mathcal{R}$. The three different distance functions have been commonly used in the literature for semi-supervised learning, which are:

- $\mathcal{R}_{\mathcal{D}_{KL}}$, the regularization loss used in VAT (Miyato et al., 2018) and FixMatch (Sohn et al., 2020), FixMatch+VAT (Wang et al., 2021a), where $\mathcal{D}_{KL}$ is Kullback–Leibler divergence (KL) loss: $\mathcal{R}_{\mathcal{D}_{KL}} = \mathcal{D}_{KL}(\mathbf{p}, \mathbf{p}') = 1/n \sum_{i=1}^{n} \sum_{c=1}^{C} \mathbf{p}(i)^{(c)} \log \frac{\mathbf{p}(i)^{(c)}}{\mathbf{p}'(i)^{(c)}}$ where n is the number of pixels in the image;
- $\mathcal{R}_{\mathcal{D}_{MSE}}$, the regularization loss used in TCSM (Li et al., 2020) and MixMatch (Berthelot et al., 2019), where $\mathcal{D}_{MSE}$ is the mean squared loss;
- $\mathcal{R}_{\mathcal{D}_{KL+Contour}}$, where $\mathcal{D}_{KL+Contour}$ consists of the $\mathcal{D}_{KL}$ loss and the contour-based loss $\mathcal{D}_{Contour}$ used in our previous work (Chen et al., 2020).

And the proposed one is denoted as $\mathcal{R}_{\mathcal{D}_{MSE+Contour}}$ for clarity. We ran experiments on the two tasks. Results are shown in Table 4. Compared to the other three loss functions, the proposed one $\mathcal{R}_{\mathcal{D}_{MSE+Contour}}$ outperforms the other three in most cases and achieves the highest average Dice scores in both segmentation tasks. Adding contour-based loss ($\mathcal{R}_{\mathcal{D}_{KL+Contour}}$, $\mathcal{R}_{\mathcal{D}_{MSE+Contour}}$) in general provides better performance than their corresponding standalone counterpart (i.e. $\mathcal{R}_{\mathcal{D}_{MSE}}, \mathcal{R}_{\mathcal{D}_{KL}}$), highlighting the benefits of taking additional boundary information for consistency regularization.

5.2.4. Effect of number of labeled images

In Fig. 9, we report the segmentation performance of our methods under different data settings on the cardiac segmentation task. Networks were trained with a different number of labeled subjects and the same unlabeled dataset ($M = 25$). As expected, the performance of semi-supervised learning improves when more labeled training images are included. Compared to standard training (trained without consistency regularization), the proposed method (*AdvChain*) consistently provides significant performance improvement across all settings. The performance gain is particularly evident in the extremely one-shot setting (N=1). The performance gap between the standard training and ours narrows when adding more labeled images into training, which is consistent with the finding reported in other semi-supervised learning frameworks (Li et al., 2020).

5.3. Supervised learning with extremely low data settings

We evaluate the performance of the proposed method under extremely low data settings, where there is no unlabeled data available ($M = 0$). Results are shown in Table 5. It is clear that the proposed *AdvChain* consistently outperforms the competitive baseline method *RandChain* on the two tasks by a large margin. The results confirm that in the scenario where training data is limited, the proposed method still enhances model training significantly. This indicates the great potential of the proposed method to alleviate the data scarcity problem.

6. Discussion

In this work, we have presented a novel adversarial data augmentation method, which is capable of introducing both realistic photometric and geometric transformations to improve the generalization capability for neural network-based medical image segmentation of MR images. The proposed method enhances several aspects of previous data augmentation and regularization schemes. Compared to VAT (Miyato et al., 2018), RandAugment (Cubuk et al., 2020) and the data augmentation in TCSM (Li et al., 2020), the proposed method provides counterpart samples with more realistic variations in medical imaging, including challenging local intensity variations (bias fields) and morphological

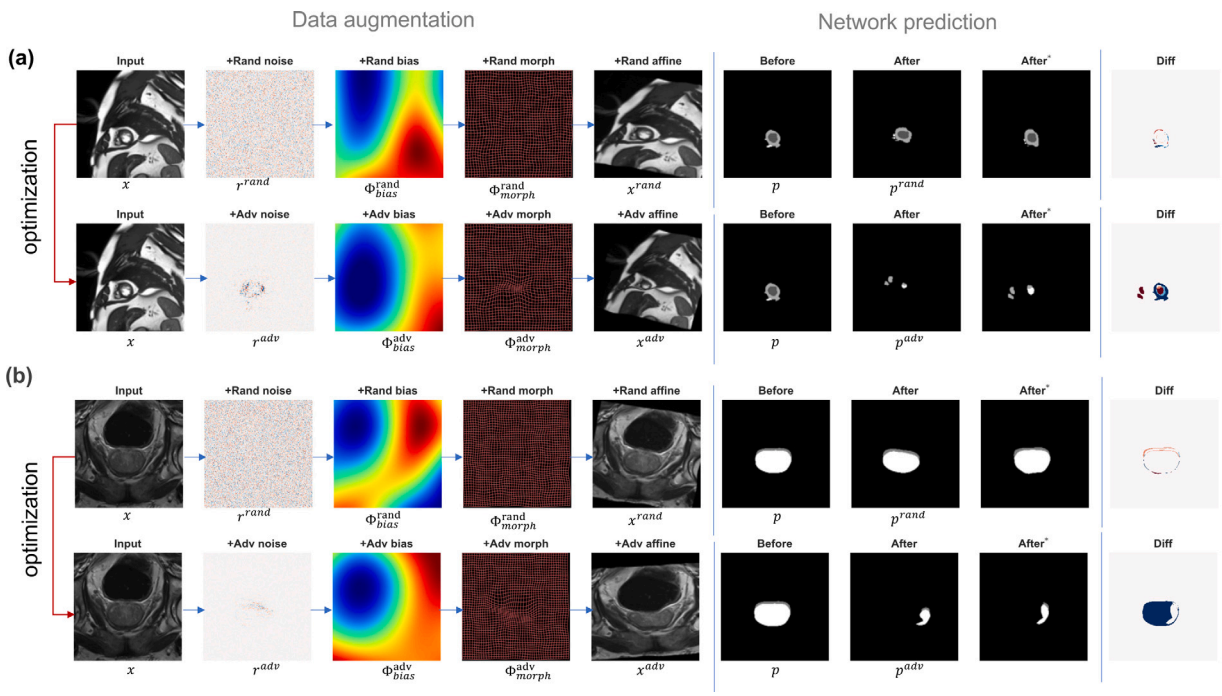

Fig. 8. Optimizing a chain of data transformation parameters produces effective adversarial samples, which greatly alter the network's predictions for (a) the cardiac segmentation task and (b) the prostate segmentation task, respectively. Here, the order of chained transformations is: $\mathcal{T}_{\text{noise}} \rightarrow \mathcal{T}_{\text{bias}} \rightarrow \mathcal{T}_{\text{morph}} \rightarrow \mathcal{T}_{\text{affine}}$ ($\mathcal{T}_{\text{affine}\circ\text{morph}\circ\text{bias}\circ\text{noise}}$). Before/After: predictions before/after data augmentation. After*: perturbed predictions which have been transformed back to the original image coordinates for consistency measurement.

Table 5
Results of low-shot supervised learning on the cardiac and prostate datasets.

	Cardiac						Prostate			
	N = 1, M = 0			N = 3, M = 0			N = 3, M = 0		N = 11, M = 0	
	LV	MYO	RV	LV	MYO	RV	PZ	CZ	PZ	CZ
Standard training	0.5155	0.4290	0.2201	0.8269	0.7905	0.6288	0.3017	0.6278	0.5057	0.8091
RandChain	0.5581	0.4570	0.2699	0.8183	0.7857	0.6123	0.3962	0.7570	0.5641	0.8385
AdvChain	**0.6093**	**0.5022**	**0.3079**	**0.8435**	**0.8122**	**0.6473**	**0.4192**	**0.7600**	**0.5720**	**0.8450**

Reported values are average Dice scores.

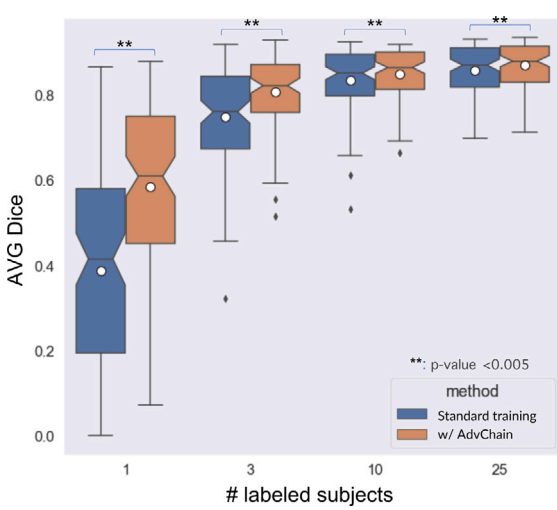

Fig. 9. Semi-supervised learning results on the cardiac test set with networks trained using different numbers of labeled subjects and the same unlabeled set (25 subjects). Standard training: supervised training using only labeled images.

changes (diffeomorphic deformations). Compared to data-mixing-based methods such as Mixmatch (Berthelot et al., 2019), which generates unrealistic mixed images with linear interpolation to ensure the 'linearity' of the network, the proposed method applies physics-based transformation models to generate a diverse set of extrapolated data points around each input, which can be viewed as a way to encourage the 'local smoothness' under various local perturbations. We believe that local smoothness is a better regularization for segmentation tasks, as it encourages the network to incorporate human perception, clustering perceptually similar images for decision making. In particular, it *strengthens the network's invariance against photometric transformations and equivariance under geometric transformations*, where the two properties are highly desirable for model generalization. On the segmentation tasks for cardiac and prostate MR images, we demonstrated that the proposed method has great potential to reduce the annotation effort, outperforming competitive baseline methods in both low-shot supervised settings and semi-supervised settings.

We notice that there are concurrent works using GAN (Chaitanya et al., 2019) and adversarial training (Gao et al., 2021) to find effective photometric and geometric transformations for data augmentation. These methods cannot be directly compared to ours as they require training additional neural networks. Since GANs are essentially large neural networks, they still require a large number of training images to avoid over-fitting. And their methods cannot be applied to optimize dynamically chained transformations as it suffers from the training instability problem with a set of randomly stacked GANs. Our method, by contrast, is more *flexible* and *data efficient*, as it only employs a small set of explainable and controllable parameters and can be used even in

extremely low data settings (e.g., only 1 or 3 labeled subjects, no access to unlabeled subjects). Also, training GAN requires considerable computational resources and expertise to tune hyperparameters and can be very unstable (Gulrajani et al., 2017). The proposed adversarial data augmentation by contrast can be directly used as a plug-in lightweight module to support training segmentation pipelines.

Limitations: One limitation of the proposed method is that it still requires expertise to explicitly specify the magnitude constraints for the employed parameters to ensure the naturalism of augmented images. Yet, how to automatically find optimal data augmentation policy (data augmentation operations and associated probabilities, magnitudes, the order of chained transformations, the optimum chain length) is still an active research area (Cubuk et al., 2019; Zhang et al., 2020; Shorten and Khoshgoftaar, 2019). In recent years, there has been an emerging research topic focusing on automatic data augmentation (Auto DA), which in general requires an external RNN controller to find optimal probabilities and magnitudes for a group of image transformations for a particular dataset (Cubuk et al., 2019; Shorten and Khoshgoftaar, 2019). Combining Auto DA with the proposed method may further automate the process with higher accuracy.

To further enhance the effectiveness of the proposed method for medical image segmentation tasks, one can also consider employing more advanced segmentation network architectures to increase its representation learning capacity for improved segmentation accuracy and integrating *AdvChain* with other advanced techniques to solve potential additional challenges. For example, medical image segmentation often suffers from class imbalance (long-tail problem), which may skew the performance of the segmentation model. In this work, we applied weighted supervised loss, as a common practice. It is worthwhile to explore more advanced class-imbalance invariant techniques, e.g., suppressed consistency loss (Hyun et al., 2020) for further improvements. On the other hand, medical images are typically grayscale images with poor image contrast where the anatomical structures may have very blurry contours. This increases the burden of producing reliable predictions (pseudo labels) on unlabeled images before applying perturbations for consistency regularization. To improve the reliability of pseudo labels, one can adopt an iterative training procedure, which distills previously learned knowledge into a neural network with an equal or larger capacity to boost model performance on label estimation (Xie et al., 2020c; Zoph et al., 2020). Also, it is interesting to introduce a pseudo label assessment module to select high quality pseudo labels for more effective uncertainty-aware consistency regularization (Xia et al., 2020a,b; Liu and Tan, 2021; Wang et al., 2021b; Yu et al., 2019). We will explore these extensions in future work.

7. Conclusion

This work tackles the challenging task of multi-class segmentation on MR images, given very limited number of labeled subjects. A novel adversarial data augmentation method has been presented, which jointly optimizes a *dynamic data augmentation* module and the segmentation network to better leverage labeled and unlabeled data for improved model generalization. The proposed data augmentation method is capable of improving both data effectiveness and diversity with challenging complex data variations based on photometric and geometric transformations (Section 5.2.1, Section 5.2.2), simulating realistic image appearance and anatomical variations that could exist in MR imaging. Our work also highlights the importance of (1) introducing adversarial diffeomorphic deformations for improved data diversity and effectiveness (Section 5.2.1), which has not been explored in prior consistency regularization-based methods. We have also demonstrated the effectiveness of adding the contour-based consistency loss for a more comprehensive inconsistency measurement to inform network training, see Section 5.2.3.

The whole framework can be used as a plug-in module to facilitate supervised and semi-supervised learning and is generic for MR image segmentation tasks. With only four types of photo-metric and geometric transformations, we have demonstrated its great data efficiency on two different tasks in challenging low-shot semi-supervised settings, outperforming several strong consistency-regularized methods in different scenarios. The proposed method even outperforms the state-of-the-art composite data augmentation method (RandAugment (Cubuk et al., 2020)) in most cases. The flexibility and the generic nature of *AdvChain* open the door to incorporate more image transformations to better reflect the imaging variations in the real world and thus *AdvChain* has the potential to be applied to different imaging modalities and different data-driven medical imaging applications, such as image registration (Uzunova et al., 2017) and image reconstruction (Cheng et al., 2020). We leave that for future work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by two EPSRC Grants (EP/P001009/1, EP/R005982/1) and the ERC Grant (884622). W. Bai was supported by EPSRC DeepGeM Grant (EP/W01842X/1). S. Wang was supported by the Shanghai Sailing Programs of Shanghai Municipal Science and Technology Committee, China (22YF1409300).

Appendix

See Fig. A.1.

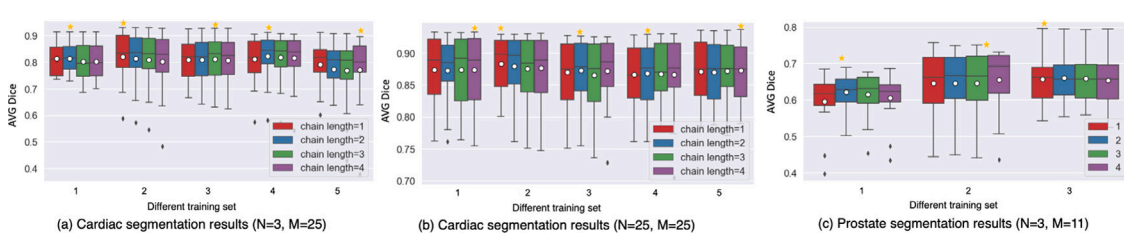

Fig. A.1. Optimal maximum chain length of *AdvChain* depends on the selection of training set, training set size, and the segmentation task. Here, we plot boxplots of average Dice scores with the segmentation network trained with *AdvChain* using different selections of labeled sets across different tasks. We varied the maximum chain lengths (1–4) to search for the optimum setting that achieves the highest average Dice scores on the test set. A yellow star in each group indicates the optimum chain length for a specific setting. We find that the optimum value varies across different tasks and different labeled set selections, which can be observed on the cardiac and prostate segmentation tasks, see (a), (c), respectively. This phenomenon still exists even when we increased the number of labeled subjects from 3 to 25 for the cardiac segmentation task (b). There is no consensus on the optimum maximum chain length across different selections of labeled sets. N: number of labeled images, M: number of unlabeled images.

References

Alaifari, R., Alberti, G.S., Gauksson, T., 2019. ADef: an iterative algorithm to construct adversarial deformations. In: 7th "International Conference on Learning Representations". ICLR 2019, New Orleans, la, USA, May 6-9, 2019, pp. 1–25, URL https://openreview.net/forum?id=Hk4dFjR5K7.

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., van Ginneken, B., Bilello, M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M.J., Heckers, S.H., Huisman, H., Jarnagin, W.R., McHugo, M.K., Napel, S., Pernicka, J.S.G., Rhode, K., Tobon-Gomez, C., Vorontsov, E., Meakin, J.A., Ourselin, S., Wiesenfarth, M., Arbeláez, P., Bae, B., Chen, S., Daza, L., Feng, J., He, B., Isensee, F., Ji, Y., Jia, F., Kim, I., Maier-Hein, K., Merhof, D., Pai, A., Park, B., Perslev, M., Rezaiifar, R., Rippel, O., Sarasua, I., Shen, W., Son, J., Wachinger, C., Wang, L., Wang, Y., Xia, Y., Xu, D., Xu, Z., Zheng, Y., Simpson, A.L., Maier-Hein, L., Cardoso, M.J., 2022. The medical segmentation decathlon. Nature Commun. 13 (1), 1–13.

Arsigny, V., Commowick, O., Pennec, X., Ayache, N., 2006. A log-euclidean framework for statistics on diffeomorphisms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 924–931.

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. NeuroImage 38 (1), 95–113.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. VoxelMorph: A learning framework for deformable medical image registration. IEEE Trans. Med. Imaging http://dx.doi.org/10.1109/TMI.2019.2897538.

Baumgartner, C.F., Koch, L.M., Pollefeys, M., Konukoglu, E., 2018. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. In: Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges. Springer International Publishing, pp. 111–119. http://dx.doi.org/10.1007/978-3-319-75541-0_12.

Bernard, O., Lalande, A., et al., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Trans. Med. Imaging 0062 (11), 2514–2525. http://dx.doi.org/10.1109/TMI.2018.2837502.

Berthelot, D., Carlini, N., Goodfellow, I.J., Papernot, N., Oliver, A., Raffel, C., 2019. MixMatch: A holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019. NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada, pp. 5050–5060, URL http://papers.nips.cc/paper/8749-mixmatch-a-holistic-approach-to-semi-supervised-learning.

Cachier, P., Bardinet, E., Dormont, D., Pennec, X., Ayache, N., 2003. Iconic feature based nonrigid registration: the PASHA algorithm. Comput. Vis. Image Underst. 89 (2–3), 272–298.

Carlini, N., Wagner, D.A., 2017. Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy. SP 2017, San Jose, CA, USA, May 22-26, 2017, pp. 39–57. http://dx.doi.org/10.1109/SP.2017.49.

Chaitanya, K., Karani, N., Baumgartner, C.F., Becker, A., Donati, O., Konukoglu, E., 2019. Semi-supervised and task-driven data augmentation. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 29–41.

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D., 2019a. Intelligent image synthesis to attack a segmentation CNN using adversarial learning. In: Simulation and Synthesis in Medical Imaging - 4th International Workshop. SASHIMI 2019, Held in Conjunction with MICCAI 2019, pp. 90–99. http://dx.doi.org/10.1007/978-3-030-32778-1_10.

Chen, C., Ouyang, C., Tarroni, G., Schlemper, J., Qiu, H., Bai, W., Rueckert, D., 2019b. Unsupervised multi-modal style transfer for cardiac MR segmentation. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 209–219.

Chen, C., Qin, C., Qiu, H., Ouyang, C., Wang, S., Chen, L., Tarroni, G., Bai, W., Rueckert, D., 2020. Realistic adversarial data augmentation for MR image segmentation. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part I, In: Lecture Notes in Computer Science, vol.12261, Springer, pp. 667–677. http://dx.doi.org/10.1007/978-3-030-59710-8_65.

Cheng, K., Calivá, F., Shah, R., Han, M., Majumdar, S., Pedoia, V., 2020. Addressing the false negative problem of deep learning MRI reconstruction models by adversarial attacks and robust training. Proc. Mach. Learn. Res. 1, 15.

Cubuk, E.D., Zoph, B., Mané, D., Vasudevan, V., Le, Q.V., 2019. AutoAugment: Learning augmentation policies from data. In: IEEE Conference on Computer Vision and Pattern Recognition. arXiv:1805.09501.

Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V., 2020. Randaugment: Practical automated data augmentation with a reduced search space. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020, Computer Vision Foundation / IEEE, pp. 3008–3017. http://dx.doi.org/10.1109/CVPRW50498.2020.00359.

Cui, W., Liu, Y., Li, Y., Guo, M., Li, Y., Li, X., Wang, T., Zeng, X., Ye, C., 2019. Semi-supervised brain lesion segmentation with an adapted mean teacher model. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 554–565.

Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., Madry, A., 2019. Exploring the landscape of spatial robustness. In: Chaudhuri, K., Salakhutdinov, R. (Eds.), ICML. In: Proc. Mach. Learn. Res., vol.97, PMLR, Long Beach, California, USA, pp. 1802–1811, URL http://proceedings.mlr.press/v97/engstrom19a.html.

Ferreira, P.F., Gatehouse, P.D., Mohiaddin, R.H., Firmin, D.N., 2013. Cardiovascular magnetic resonance artefacts. J. Cardiovasc. Magn. Reson. Off. J. Soc. Cardiovasc. Magn. Res. 15, 41.

Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S., 2019. Adversarial attacks on medical machine learning. Science 363 (6433), 1287–1289. http://dx.doi.org/10.1126/science.aaw4399.

Gallier, J., Gallier, J.H., 2000. Curves and Surfaces in Geometric Modeling: Theory and Algorithms. Morgan Kaufmann.

Gao, Y., Tang, Z., Zhou, M., Metaxas, D.N., 2021. Enabling data diversity: Efficient automatic augmentation via regularized adversarial training. In: Feragen, A., Sommer, S., Schnabel, J.A., Nielsen, M. (Eds.), Information Processing in Medical Imaging - 27th International Conference. IPMI 2021, Virtual Event, June 28-June 30, 2021, Proceedings, In: Lecture Notes in Computer Science, vol.12729, Springer, pp. 85–97. http://dx.doi.org/10.1007/978-3-030-78191-0_7.

Goodfellow, I.J., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples. In: International Conference on Learning Representations. pp. 1–11, URL http://arxiv.org/abs/1412.6572.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C., 2017. Improved training of wasserstein GANs. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. December 4-9, 2017, Long Beach, CA, USA, pp. 5767–5777, URL https://proceedings.neurips.cc/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html.

Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B., 2020. AugMix: A simple data processing method to improve robustness and uncertainty. In: 8th International Conference on Learning Representations. ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, pp. 1–15, URL https://openreview.net/forum?id=S1gmrxHFvB.

Hyun, M., Jeong, J., Kwak, N., 2020. Class-imbalanced semi-supervised learning. In: International Conference on Machine Learning. pp. 1–16, arXiv:2002.06815.

Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18 (2), 203–211.

Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K., 2015. Spatial transformer networks. In: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015. December 7-12, 2015, Montreal, Quebec, Canada, pp. 2017–2025, URL https://proceedings.neurips.cc/paper/2015/hash/33ceb07bf4eeb3da587e268d663aba1a-Abstract.html.

Kanbak, C., Moosavi-Dezfooli, S., Frossard, P., 2018. Geometric robustness of deep networks: Analysis and improvement. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4441–4449. http://dx.doi.org/10.1109/CVPR.2018.00467, URL http://openaccess.thecvf.com/content_cvpr_2018/html/Kanbak_Geometric_Robustness_of_CVPR_2018_paper.html.

Li, Z., Kamnitsas, K., Glocker, B., 2019. Overfitting of neural nets under class imbalance: Analysis and improvements for segmentation. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P., Khan, A.R. (Eds.), Medical Image Computing and Computer Assisted Intervention. In: Lecture Notes in Computer Science, vol.11766, Springer, pp. 402–410. http://dx.doi.org/10.1007/978-3-030-32248-9_45.

Li, X., Yu, L., Chen, H., Fu, C.-W., Xing, L., Heng, P.-A., 2020. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. IEEE Trans. Neural Netw. Learn. Syst..

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Liu, L., Tan, R.T., 2021. Certainty driven consistency loss on multi-teacher networks for semi-supervised learning. J. Pattern Recognit..

Liu, Q., Yu, L., Luo, L., Dou, Q., Heng, P.A., 2020. Semi-supervised medical image classification with relation-driven self-ensembling model. IEEE Trans. Med. Imaging arXiv:2005.07377.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A., 2017. Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations. pp. 1–23, arXiv:1706.06083.

Miyato, T., Maeda, S.-I., Koyama, M., Ishii, S., 2018. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. URL http://arxiv.org/abs/1704.03976.

Panfilov, E., Tiulpin, A., Klein, S., Nieminen, M.T., Saarakkala, S., 2019. Improving robustness of deep learning based knee MRI segmentation: Mixup and adversarial domain adaptation. In: 2019 IEEE/CVF International Conference on Computer Vision Workshops. ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019, IEEE, pp. 450–459. http://dx.doi.org/10.1109/ICCVW.2019.00057.

Paschali, M., Conjeti, S., Navarro, F., Navab, N., 2018. Generalizability vs. Robustness: Investigating medical imaging networks using adversarial examples. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2018, Springer International Publishing, Cham, pp. 493–501.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Sajjadi, M., Javanmardi, M., Tasdizen, T., 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In: Advances in Neural Information Processing Systems. pp. 1163–1171.

Sandkühler, R., Jud, C., Andermatt, S., Cattin, P.C., 2018. AirLab: Autograd image registration laboratory. Arxiv arXiv:1806.09907.

Shen, D., Wu, G., Suk, H.-I., 2017. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19, 221–248. http://dx.doi.org/10.1146/annurev-bioeng-071516-044442, URL http://dx.doi.org/10.1146/annurev-bioeng-071516-044442.

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6 (1), 60. http://dx.doi.org/10.1186/s40537-019-0197-0.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17 (1), 87–97.

Smistad, E., Falch, T.L., Bozorgi, M., Elster, A.C., Lindseth, F., 2015. Medical image segmentation on GPUs–a comprehensive review. Med. Image Anal. 20 (1), 1–18.

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C., 2020. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In: Annual Conference on Neural Information Processing Systems.

Suzuki, T., Sato, I., 2020. Adversarial transformations for semi-supervised learning. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. AAAI Press, pp. 5916–5923.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems. pp. 1195–1204.

Tramèr, F., Boneh, D., 2019. Adversarial training and robustness for multiple perturbations. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019. NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pp. 5858–5868.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29 (6), 1310–1320. http://dx.doi.org/10.1109/TMI.2010.2046908.

Uzunova, H., Wilms, M., Handels, H., Ehrhardt, J., 2017. Training CNNs for image registration from few samples with model-based data augmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 223–231.

Vercauteren, T., Pennec, X., Perchant, A., Ayache, N., 2009. Diffeomorphic demons: efficient non-parametric image registration. NeuroImage 45 (1 Suppl), S61–72. http://dx.doi.org/10.1016/j.neuroimage.2008.10.040.

Volpi, R., Namkoong, H., Sener, O., Duchi, J.C., Murino, V., Savarese, S., 2018. Generalizing to unseen domains via adversarial data augmentation. In: Annual Conference on Neural Information Processing Systems. pp. 5339–5349, URL http://papers.nips.cc/paper/7779-generalizing-to-unseen-domains-via-adversarial-data-augmentation.

Wang, X., Chen, H., Xiang, H., Lin, H., Lin, X., Heng, P.-A., 2021a. Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification. Med. Image Anal. 70, 102010.

Wang, P., Peng, J., Pedersoli, M., Zhou, Y., Zhang, C., Desrosiers, C., 2021b. Self-paced and self-consistent co-training for semi-supervised image segmentation. Med. Image Anal. 73, 102146. http://dx.doi.org/10.1016/j.media.2021.102146, URL https://www.sciencedirect.com/science/article/pii/S1361841521001924.

Wu, S., Zhang, H.R., Valiant, G., Ré, C., 2020. On the generalization effects of linear transformations in data augmentation. In: 37th International Conference on Machine Learning. ICML 2020, 13-18 July 2020, Virtual Event, In: Proc. Mach. Learn. Res., vol.119, PMLR, pp. 10410–10420, URL http://proceedings.mlr.press/v119/wu20g.html.

Xia, Y., Liu, F., Yang, D., Cai, J., Yu, L., et al., 2020a. 3D semi-supervised learning with uncertainty-aware multi-view co-training. IEEE Winter URL http://openaccess.thecvf.com/content_WACV_2020/html/Xia_3D_semi-supervised_learning_with_uncertainty-aware_multi-view_co-training_WACV_2020_paper.html.

Xia, Y., Yang, D., Yu, Z., Liu, F., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., Roth, H., 2020b. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. Med. Image Anal. 65, 101766. http://dx.doi.org/10.1016/j.media.2020.101766.

Xie, Q., Dai, Z., Hovy, E.H., Luong, T., Le, Q., 2020a. Unsupervised data augmentation for consistency training. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.-F., Lin, H.-T. (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020. NeurIPS 2020, December 6-12, 2020, Virtual.

Xie, Q., Dai, Z., Hovy, E.H., Luong, T., Le, Q., 2020b. Unsupervised data augmentation for consistency training. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020. NeurIPS 2020, December 6-12, 2020, Virtual, URL https://proceedings.neurips.cc/paper/2020/hash/44feb0096faa8326192570788b38c1d1-Abstract.html.

Xie, Q., Luong, M., Hovy, E.H., Le, Q.V., 2020c. Self-training with noisy student improves ImageNet classification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, pp. 10684–10695.

Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.L., 2017. Adversarial examples for semantic segmentation and object detection. In: IEEE International Conference on Computer Vision. ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, pp. 1378–1387. http://dx.doi.org/10.1109/ICCV.2017.153.

Xing, Y., Song, Q., Cheng, G., 2021. On the algorithmic stability of adversarial training. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (Eds.), Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 26523–26535.

Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A., 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: Medical Image Computing and Computer Assisted Intervention. MICCAI 2019, Springer International Publishing, pp. 605–613.

Zeng, X., Liu, C., Wang, Y., Qiu, W., Xie, L., Tai, Y., Tang, C., Yuille, A.L., 2019. Adversarial attacks beyond the image space. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4302–4311, URL http://openaccess.thecvf.com/content_CVPR_2019/html/Zeng_Adversarial_Attacks_Beyond_the_Image_Space_CVPR_2019_paper.html.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2018. Mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations. pp. 1–13, URL http://arxiv.org/abs/1710.09412.

Zhang, X., Wang, Q., Zhang, J., Zhong, Z., 2020. Adversarial AutoAugment. In: 8th International Conference on Learning Representations. ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, pp. 1–13, URL https://openreview.net/forum?id=ByxdUySKvS.

Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V., 2019. Data augmentation using learned transformations for one-shot medical image segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 8543–8553.

Zoph, B., Ghiasi, G., Lin, T., Cui, Y., Liu, H., Cubuk, E.D., Le, Q., 2020. Rethinking pre-training and self-training. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020. NeurIPS 2020, December 6-12, 2020, Virtual, pp. 1–13, URL https://proceedings.neurips.cc/paper/2020/hash/27e9661e033a73a6ad8cefcde965c54d-Abstract.html.