Journal of Machine Learning for Biomedical Imaging. 2022:014. pp 1-27 Submitted 12/2021; Published 05/2022
Special Issue: MICCAI 2021 Workshops Ominibus – DART
Guest Editors: Ziyue Xu, Konstantinos Kamnitsas

Re-using Adversarial Mask Discriminators for Test-time Training under Distribution Shifts

Gabriele Valvano gabriele.valvano@imtlucca.it
IMT School for Advanced Studies Lucca, Lucca LU, Italy
School of Engineering, University of Edinburgh, Edinburgh, UK

Andrea Leo andrea.leo@unipi.it
University of Pisa, Pisa, Italy

Sotirios A. Tsaftaris S.Tsaftaris@ed.ac.uk
School of Engineering, University of Edinburgh, Edinburgh, UK

Abstract

Thanks to their ability to learn flexible data-driven losses, Generative Adversarial Networks (GANs) are an integral part of many semi- and weakly-supervised methods for medical image segmentation. GANs jointly optimise a generator and an adversarial discriminator on a set of training data. After training is complete, the discriminator is usually discarded, and only the generator is used for inference. But should we discard discriminators? In this work, we argue that training stable discriminators produces expressive loss functions that we can re-use at inference to detect and *correct* segmentation mistakes. First, we identify key challenges and suggest possible solutions to make discriminators re-usable at inference. Then, we show that we can combine discriminators with image reconstruction costs (via decoders) to endow a causal perspective to test-time training and further improve the model. Our method is simple and improves the test-time performance of pre-trained GANs. Moreover, we show that it is compatible with standard post-processing techniques and it has the potential to be used for Online Continual Learning. With our work, we open new research avenues for re-using adversarial discriminators at inference. Our code is available at `https://vios-s.github.io/adversarial-test-time-training`.

Keywords: GAN, Discriminator, Segmentation, Test-time training, Shape prior

1. Introduction

Generative Adversarial Networks (GANs, Goodfellow et al. (2014)) have recently shown astonishing performance in semi-supervised and weakly-supervised segmentation learning (Cheplygina et al., 2019; Tajbakhsh et al., 2020), providing training signals when labels are sparse or missing. GANs jointly optimise two networks in the adversarial setup: one to solve an image generation task (generator) and the other to distinguish between real images and the generated ones (discriminator or critic). In image segmentation, the generator is named segmentor and, conditioned on an input image, learns to predict correct segmentation masks. Meanwhile, the discriminator learns a data-driven shape prior and penalise the segmentor when it produces unrealistic predictions. Once training is complete, the discriminator is discarded, and the segmentor is used for inference.

Unfortunately, segmentors may decrease their test-time performance when the test data fall outside the training data distribution. Often, mistakes are easy to detect by the human

`https://www.melba-journal.org/papers/2022:014.html`.

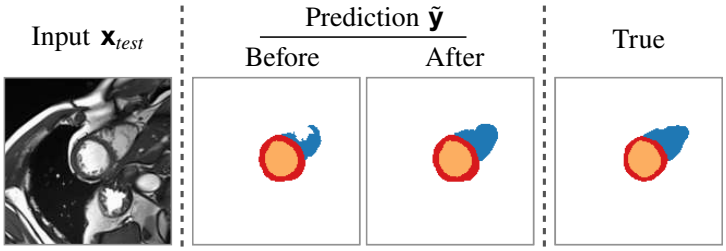

Figure 1: After training a GAN for semantic segmentation, whenever a test image falls outside the training data distribution, the segmentor may underperform and produce unrealistic predictions. Our work suggests re-using already optimised adversarial discriminators to tune the segmentor predictions on the individual test images until the predicted mask satisfies the learned shape prior.

eye as they appear as holes inside of the predicted masks or scattered false positives. Would a mask discriminator, trained to learn a shape prior, be able to spot such mistakes?

In this work, we introduce a simple mechanism to detect and correct such segmentation errors re-using components already developed during training. We adopt the emerging paradigm of Test-time Training: a term coined by Sun et al. (2020). Test-time Training fine-tunes pre-trained networks on the individual test samples without any additional supervision, resulting in Variable Decision Boundary models (Sun et al., 2020; Wang et al., 2021; Karani et al., 2021; Zhang et al., 2021; Bartler et al., 2021). We present strategies to allow for the recycling of adversarial mask discriminators to make them useful at inference. Inspired by Asano et al. (2020) showing that it is possible to train the early layers of a CNN with just a single image, we propose fine-tuning these layers on each test sample until the segmentor prediction satisfies the learned adversarial shape prior. We report an example of the benefits of our method when applied on initially erroneous predictions in Fig. 1, and summarise our **contributions** as follows:

- To the best of our knowledge, our work is the first attempt to re-use adversarial mask discriminators at inference to *detect* and *correct* segmentation mistakes.
- We define specific assumptions that make discriminators still useful after training, and present possible solutions to satisfy them.
- We investigate the possibility to further improve test-time segmentation through causal learning, where we complement the discriminator with an image reconstructor.
- We explore various learning scenarios and report consistent performance increase on multiple medical datasets.

This paper extends our previous publication (Valvano et al., 2021a) by: i) including a mechanism to set the number of iterations at test-time automatically; ii) evaluating our method on three additional datasets and in different training scenarios; iii) investigating if reconstruction losses can endow the model a causal perspective to improve model performance and inference speed; iv) analysing the method's compatibility with post-processing

techniques; v) including experiments in continual learning settings; and vi) including additional ablations.

2. Related Work

2.1 Learning from Test Samples

In our work, we use a discriminator to tune a segmentor on the individual *test* images until it predicts realistic masks. The idea of unsupervised fine-tuning of a model on the test data has been recently introduced by Sun et al. (2020) and termed Test-time Training (TTT). To optimise a model on the training set, TTT suggests jointly minimising a supervised and an auxiliary self-supervised loss, such as predicting the rotation angle of a given image. After training, TTT uses the auxiliary task to fine-tune the model on the individual test samples and adapts to potential distribution shifts without the need for supervision. Unfortunately, Sun et al. only test their method "simulating" domain shifts through hand-crafted image corruptions, such as noise and blurring, and do not investigate if TTT can also improve semantic segmentation. Moreover, despite the success of TTT in image classification tasks, designing a well-suited auxiliary task is non-trivial (Sun et al., 2020). For instance, predicting rotation angles may be non-optimal in medical imaging, where images have different acquisition geometries.

After this seminal work, Wang et al. (2021) proposed tuning an adaptor network to minimise the prediction entropy on a test set. Unfortunately, this method needs access to the *entire* test set for fine-tuning. Hence, Zhang et al. (2021) proposed to focus on a single test point, minimising the marginal entropy of the model predictions under a set of data augmentations. Unfortunately, neural networks are well known for making low-entropy overly-confident predictions (Guo et al., 2017), and minimising segmentation entropy could be sub-optimal. Moreover, the quality of the results also depends on making a good choice of augmentations Zhang et al. (2021).

Recently, Karani et al. (2021) extended TTT to semantic segmentation by proposing Test-time Adaptable Neural Networks (TTANN). At first, TTANN learns a data-driven shape prior by pre-training a mask Denoising Autoencoder (DAE). At inference, the DAE auto-encodes the masks generated by a segmentor producing a reconstruction error. This error drives the fine-tuning (i.e. TTT) of a small adaptor CNN in front of the segmentor, ultimately mapping the individual test images onto a normalised space that overcomes domain shifts problems for the segmentor. Our work achieves the same goal but uses GANs. Unlike TTANN, which separately pre-trains the mask DAE, our model is end-to-end because it can learn the shape prior while optimising the segmentor. Moreover, it reduces computational requirements as discriminators need less GPU memory and computation than autoencoders.

Concurrent with our work, He et al. (2021) propose to use auto-encoders for TTT. They propose a bespoke model that can be fine-tuned at test time to adapt to local distributions using the decoder. However, our contribution is orthogonal to their work as He et al. (2021) do not introduce any shape prior during adaptation.

Herein, we open new research directions towards learning re-usable discriminators that can improve segmentation performance at inference. We also show that we can further

enhance test-time performance by combining the discriminator with reconstruction costs (via decoders), thus endowing a causal perspective to TTT.

2.2 Tackling Distribution Shifts

In recent years, improving model robustness under distribution shifts has attracted considerable attention in medical imaging, where images vary among scanners, patients, and acquisition protocols (Castro et al., 2020). In this context, domain adaptation and generalisation have become relevant research areas. Several methods attempt to learn domain invariant representations by anticipating the distribution difference between training and test (Joyce et al., 2017; Li et al., 2018; Dou et al., 2019; Guan and Liu, 2021; Zhou et al., 2021). However, these approaches usually require prior knowledge about the test data, such as a small subset of (possibly labelled) images from the test distribution. Unfortunately, these data can be expensive or even impossible to acquire for every target domain, and distribution shifts might be not easily identifiable (Recht et al., 2018).

An alternative approach is adapting the network parameters directly to the test samples (Sun et al., 2020; Karani et al., 2021). Similarly, our method does not need to simulate test distribution shifts, as it automatically adapts the segmentor to the individual test instances. Thus, our approach can be assumed to perform *one-sample unsupervised domain adaptation* on the fly. Notice also that, compared to standard domain adaptation techniques, Test-time Training has the advantage that it does not become ill-defined when there is only one sample from the target domain and does not require the source data during test-time training.

2.3 Shape Priors in Deep Learning for Medical Segmentation

In the past, several methods have introduced shape priors into segmentation models (Nosrati and Hamarneh, 2016; Jurdi et al., 2020) in the form of penalties (Kervadec et al., 2019; Clough et al., 2020; Jurdi et al., 2021), atlases (Dalca et al., 2019), autoencoders (Oktay et al., 2017; Dalca et al., 2018), post-processing operations (Painchaud et al., 2019; Larrazabal et al., 2020), and adversarial learning (Yi et al., 2019; Valvano et al., 2021b). Thanks to their flexibility, GANs are a popular way of introducing data-driven shape priors (Yi et al., 2019), with the advantage of learning the prior while also optimising the segmentor.

2.4 Re-using Adversarial Discriminators

Pre-trained discriminators have been re-used to navigate the generator's latent space Liu et al. (2020), as anomaly detectors (Zenati et al., 2018; Ngo et al., 2019), or as features extractors for transfer learning (Radford et al., 2015; Donahue et al., 2017; Mao et al., 2019). To the best of our knowledge, no prior work uses them to detect test time segmentor mistakes or to fine-tune pre-trained segmentors at inference.

3. Proposed Method

Below, we first provide an overview of our method. Then, we describe the challenges of re-using adversarial discriminators at inference, suggesting possible solutions to address them. Lastly, we detail model architecture, training, and the re-use of discriminators at test time.

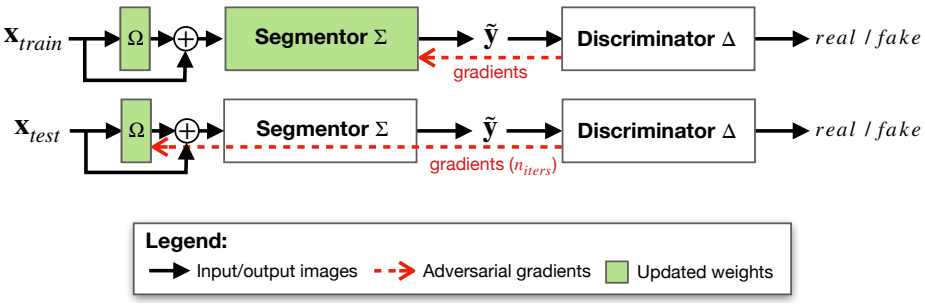

Figure 2: We re-use adversarial discriminators to correct segmentation mistakes at inference. As thoroughly discussed in Section 3.2, crucial to the method success is training stable and re-usable discriminators. At inference, we tune a shallow adaptor Ω on each test sample $\mathbf{x}$ independently, until predictions $\tilde{\mathbf{y}}$ satisfy the adversarial shape prior. We only need one sample for fine-tuning.

Notation. We use capital Greek letters to denote functions Φ, italic lowercase for scalars s, and bold lowercase for 2D images $\mathbf{x} \in \mathbb{R}^{h \times w}$, being $h, w \in \mathbb{N}$ image height and width.

3.1 Method Overview

As we summarise in Fig. 2, to segment an image we process it through a small adaptor Ω and a subsequent segmentor Σ. When annotations are available, we train both the adaptor and segmentor to minimise a supervised cost. For unlabelled images, we instead optimise them based on an adversarial cost. Meanwhile, we train an adversarial discriminator to discern real and predicted segmentation masks. At inference, we fine-tune the adaptor Ω on each test sample, leveraging only the (unsupervised) adversarial loss to increase performance.

Note that developing novel segmentors and adaptors is not our scope. Thus, we use previously developed architectures that have already shown success in segmentation tasks. On the contrary, a major contribution of this work is identifying the crucial challenges behind re-using adversarial discriminators at inference and suggesting possible solutions to overcome them.

3.2 Re-usable Discriminators: Challenges and Solutions

3.2.1 Challenge 1: Avoid Overfitting and Catastrophic Forgetting

Re-usable discriminators Δ must not *overfit* nor *catastrophically forget*. Otherwise, their predictions at test time will not be reliable. Ensuring this condition holds is challenging because GANs can easily memorise data if trained for too long (Nagarajan et al., 2018).[1] Moreover, once the segmentor's training has converged (Shrivastava et al., 2017), the discriminator may forget how unrealistic segmentation masks look like. In these cases, although Δ may work well at training, it would not generalise on test data, as we explain below.

1. Notice that memorisation can also happen just in the discriminator. In fact, contrarily to the segmentors, there is no supervised cost to regularise the discriminator's training. We show how to detect memorisation from the losses in Appendix A.

At convergence, a properly trained segmentor Σ predicts *realistic* segmentation masks. Thus, we stop training of standard GANs while optimising Δ to tell apart *real* from more and more *real-looking* masks. During the latest stages of training, this produces ambiguous training signals for the discriminator, which consistently receives real-looking inputs but is trained to label them as *real* half the time, as *fake* the other half. Since the discriminator loss will encourage Δ to classify realistic masks as fake (even if it was the segmentor to generate them), the training gradients will become unreliable. At that point, the discriminator training becomes unstable, and Δ collapses to one of these cases: **i)** always predicting the equilibrium point (which in vanilla GANs is the number 0.5, equidistant from the labels *real*: 1, *fake*: 0) but still able of detecting unrealistic images; **ii)** predicting the equilibrium point independently of the input image, forgetting what *fake* samples look like (Shrivastava et al., 2017; Kim et al., 2018); or **iii)** memorising the real masks (which, differently from the generated ones, appear unchanged since the beginning of training) and always classifying them as *real*, while classifying *any other input* as *fake*. It is crucial to prevent the last two behaviours (**ii)** and **iii)**) to have a re-usable discriminator. Thus, we use:

- *Fake anchors*: We want to expose the discriminator to unrealistic masks (labelled as *fake*) until the end of training. In particular, we train Δ using real masks $\mathbf{y}$, predicted masks $\tilde{\mathbf{y}}$, and corrupted masks $\mathbf{y}_{corr}$. We obtain $\mathbf{y}_{corr}$ by randomly swapping squared patches within the image[2] and adding binary noise to the real masks, as this proved to be a fast and effective strategy to learn robust shape priors in autoencoders (Karani et al., 2021). While, towards the end of the training, the discriminator may not distinguish $\mathbf{y}$ from the real-looking $\tilde{\mathbf{y}}$, the exposure to $\mathbf{y}_{corr}$ will prevent forgetting how unrealistic masks look like, ensuring informative training gradients.[3]

3.2.2 Challenge 2: Ensuring Stability

An additional challenge is to train *stable* discriminators, which do not change much during the last training epochs. In other words, we want to make oscillations of the discriminator loss as small as possible. This is necessary because we typically stop training using early stopping criteria on the segmentor loss. Therefore, we want to promote the optimisation of Lipschitz smooth discriminators, avoiding suddenly big gradient updates (which make the training loss oscillate). [4]

Hence, we limit the risk of having discriminators residing in sharp local minima, which are well known to generalise worse (Keskar et al., 2017), using:

2. We use patches having size equal to 10% of the image size.
3. Concurrent to our work, Sinha et al. (2021) recently introduced a similar idea, named Negative Data Augmentation, which improved the training of GAN generators. However, differently from Sinha et al. (2021), we highlight that our scope is to build a stable discriminator, which can be re-used at inference.
4. Lipschitz smooth discriminators have better stability (Chu et al., 2020) and, thus, their training gradients are bounded. Hence: i) abrupt changes in their weight values are unlikely to happen, and ii) the discriminator behaviour does not change abruptly across epochs, which is crucial at convergence. In fact, by comparing different discriminators obtained between subsequent epochs (i.e. different statistical realization of the adversarial discriminator, having different weights), we find similar behaviours, reducing the influence of eschewing discriminator-based early stopping criteria.

- *Smoothness constraints*: We use Spectral Normalisation (Miyato et al., 2018), *tanh* activations, and Gradient Penalty (Gulrajani et al., 2017) to encourage discriminator smoothness (Chu et al., 2020).
- *Discriminator data augmentation*: We use Instance Noise (Sønderby et al., 2017; Müller et al., 2019) and random roto-translations, to map similar inputs to the same prediction label. We generate noise sampling from a Normal distribution having zero mean and 0.1 standard deviation. We rotate images between $0 \div \pi/2$ and translate them up to 10% of image pixels on both vertical and horizontal axes.

3.3 Architectures and Training Objectives

Given an input image $\mathbf{x}$, we first pass it through the adaptor Ω and obtain $\mathbf{x}' = \Omega(\mathbf{x})$. Then, we use the segmentor Σ to predict a segmentation mask $\tilde{\mathbf{y}} = \Sigma(\mathbf{x}') = \Sigma \circ \Omega(\mathbf{x})$.

We parametrised Ω as the adaptor introduced by Karani et al. (2021). Such an adaptor consists in 3 convolutional layers with 16 3×3 kernels and activation $\Phi(\mathrm{T}) = e^{-\mathrm{T}^2/s^2}$, where T is an input tensor and s a trainable scaling parameter, randomly initialised and optimised at test-time. Instead, for the segmentor we use a UNet (Ronneberger et al., 2015) with batch normalisation (Ioffe and Szegedy, 2015).

For the pairs of annotated data ($\mathbf{x}$, $\mathbf{y}$) in the training set, we minimise the weighted cross-entropy loss between the real mask $\mathbf{y}$ and the predicted one $\tilde{\mathbf{y}}$:

$$\mathcal{L}(\Omega, \Sigma) = -\sum_{i=1}^{c} w_i \cdot \mathbf{y}_i \log(\tilde{\mathbf{y}}_i), \tag{1}$$

where, given the number of classes c and the class index i, we address the class imbalance problem with the scaling factor w_i. We compute $w_i = 1 - n_i/n_{tot}$ as the ratio between the number of pixels n_i having label i and the total number of pixels n_{tot}.

The discriminator Δ is a convolutional encoder, processing an input mask through 5 convolutional layers with 4×4 filters. The number of filters follows the series: 32, 64, 128, 256, 512. After the first two layers, we use a stride of 2 to downsample the extracted features maps. As discussed in Section 3.2, we increase discriminator smoothness through spectral normalisation layers and *tanh* activations. Lastly, we use a fully-connected layer to integrate high-level representation and produce a scalar linear output, used to compute the adversarial loss adapted from Mao et al. (2018):

$$\begin{aligned} &\min_{\Delta} \left\{ \mathcal{V}_{LS}(\Delta) = \frac{1}{2} E_{\mathbf{y} \sim p(\mathbf{y})}[(\Delta(\mathbf{y}) - 1)^2] + \frac{1}{2} E_{\mathbf{x} \sim p(\mathbf{x})}[(\Delta(\Sigma \circ \Omega(\mathbf{x})) + 1)^2] \right\} \\ &\min_{\Omega, \Sigma} \left\{ \mathcal{V}_{LS}(\Omega, \Sigma) = \frac{1}{2} E_{\mathbf{x} \sim p(\mathbf{x})}[(\Delta(\Sigma \circ \Omega(\mathbf{x})))^2] \right\}, \end{aligned} \tag{2}$$

where $+1$ and -1 are the labels for *real* and *fake* masks, respectively. For training, we alternately minimise Eq. 1 on a batch of labelled images and Eq. 2 on a batch of unpaired images and unpaired masks. To avoid the adversarial loss from prevailing over the supervised cost, we rescale $\mathcal{V}_{LS}(\Omega, \Sigma)$ by multiplying it by a dynamic weighting value $a = 0.1 \cdot \frac{\|\mathcal{L}(\Omega,\Sigma)\|}{\|\mathcal{V}_{LS}(\Omega,\Sigma)\|}$, as in Valvano et al. (2021b). Hence, we ensure to expose the segmentor to a supervised cost that is always one order of magnitude larger than the adversarial cost, which can judge predictions only qualitatively. We minimise losses with Adam (Kingma and Ba, 2015),

learning rate: 10^{-4}, and batch size: 12. We end training when the supervised loss stops decreasing on a validation set.

3.4 Adversarial Test-Time Training

3.4.1 Adapting Ω

At test time, we only fine-tune the first few convolutional layers of the whole model: i.e. the three layers of the adaptor Ω described in Section 3.3. Our choice is motivated by Asano et al. (2020) observations that the early layers are the most suited for one-shot learning, and is similar to that of Karani et al. (2021). Leaving the deeper layers of Σ unchanged, we let the model adapt only to changes at lower abstraction levels, limiting its flexibility and preventing trivial solutions. Thus, given a test sample $\mathbf{x}$, we tune a shallow convolutional residual block (the adaptor Ω) in front of the segmentor by minimising $\mathcal{V}_{LS}(\Omega|\Sigma, \mathbf{x})$ for n_{iter} iterations. The number of iterations n_{iter} has an upper bound and is determined on each specific test sample independently. After tuning Ω, the input to the segmentor becomes an augmented version of $\mathbf{x}$, which can be more easily classified.

3.4.2 Setting the Number of Test-time Iterations

At inference, our method needs n_{iter} forward and backward passes to correct a segmentation.[5] We compute an optimal n_{iter} for each test sample, stopping TTT when the adversarial loss on the predicted mask has not decreased for the last 200 steps, or when TTT exceeds a maximum iteration number $n_{iter}^{max} = 1000$. After stopping TTT, we consider the prediction associated with the minimum adversarial loss as the best one.

4. Experimental Setup

4.1 Data

We consider four medical datasets acquired using a variety of MRI scanners and acquisition protocols: cardiac data of ACDC (Bernard et al., 2018), M&Ms (Campello et al., 2021), and LVSC (Suinesiaputra et al., 2014); and the abdominal organ data of CHAOS (Kavur et al., 2021). ACDC and M&Ms contain annotations for right ventricle, left ventricle and left myocardium, while LVSC only has myocardium masks. CHAOS contains labels for the two kidneys, the liver, and the spleen. We use specific datasets based on two different learning scenarios, which we describe below:

- **Identifiable Distribution Shift**: We use ACDC and M&Ms data to model test-time distribution shifts that we can identify as changes in the MRI acquisition scanner. For ACDC, we build the training and validation set using only data acquired from 1.5T scanners; then, we test the model on 3T MRI scans. In the following, we refer to this dataset setup as $\text{ACDC}_{1.5\rightarrow 3\text{T}}$. For M&Ms, we consider training and validation sets containing data from 3 out of the four available MRI vendors and a test set constructed using data from the held-out vendor. As a result, we can be sure that there is a distribution shift between training and test data in both $\text{ACDC}_{1.5\rightarrow 3\text{T}}$ and

5. Despite being slower than standard inference, where each image only requires one forward pass, we observe just a small temporal overhead in the model: $\sim$10$\div$20s/patient on a TITAN Xp GPU.

M&Ms. In both cases, we maintain a 2:1 ratio between the number of samples in the training and validation sets.

- **Non-identifiable Distribution Shift**: We consider randomly sampled data from ACDC, LVSC, and CHAOS, where we cannot say in advance if there is a change in distribution between train and test data. We consider a semi-supervised learning scenario, where only a portion of training data is annotated. This setup is of particular interest when collecting large-scale annotated datasets is not possible, and small labelled training sets do not accurately represent the test distribution. To prevent information leakage, we divide datasets by patients and use groups of 40%-20%-40% of patients for training, validation, and test set, respectively. In ACDC and LVSC training sets, we only consider annotations for one fourth of the training subjects (10 patients); in CHAOS, only one half (4 patients). We treat the remaining data as unpaired and use them for adversarial learning (Eq. 2). Despite being drawn from the same distribution (i.e. the entire dataset), the small amount of training data may not fully represent the data distribution. Hence, although we cannot identify distribution shifts a priori, they may still exist and lead to performance drop (Recht et al., 2018).

After defining train, validation and test sets, we pre-process data as follows:

- **ACDC$_{\mathbf{1.5\rightarrow 3T}}$** and **ACDC:** we resample images to the average resolution of $1.51mm^2$, and crop or pad them to 224×224 pixel size. Lastly, we normalise data by subtracting the patient-specific median and dividing by its interquartile range (IQR).

- **M&Ms:** after resampling the images to the average resolution of $1.25mm^2$, we crop/pad them to 224×224 pixels. We normalise images by subtracting the patient-specific median and dividing by the IQR.

- **LVSC:** we resample images to the average resolution of $1.45mm^2$, and then crop or pad them to 224×224 pixel size. Finally, we normalise images by subtracting the patient-specific median and dividing by the IQR.

- **CHAOS:** we test our method on the T1 in-phase images, after resampling them to $1.89mm^2$, normalising and cropping them to 192×192 pixel size.

4.2 Evaluation Protocol

For all the experiments, we report results of 3-fold cross-validation. We measure performance in terms of segmentation quality, using Dice score, IoU score, and Hausdorff distance to compare the predicted segmentation masks with the ground truth labels available in the test sets. We assess statistical significance with the bootstrapped t-test. We use significance at $p \leq 0.05$ or $p \leq 0.01$ denoted by one (*) or two (**) asterisks, respectively.

5. Experiments and Discussion

We present and discuss the performance of our method in various experimental settings. First, Section 5.1 presents the advantage of the proposed approach during inference: either under identifiable or non-identifiable distribution shifts. In Section 5.2, we highlight the

Dataset	Adv. TTT	Dice (↑)	IoU (↑)	Hausdorff Distance (↓)
$ACDC_{1.5\rightarrow 3T}$	before	77.0_{09}	68.9_{09}	5.2_{02}
	after	$\mathbf{78.4_{08}}^{**}$	$\mathbf{70.4_{08}}^{**}$	$\mathbf{5.0_{02}}^{**}$
M&Ms	before	82.0_{08}	75.6_{08}	4.4_{03}
	after	$\mathbf{82.1_{08}}^{*}$	$\mathbf{75.7_{08}}^{*}$	$\mathbf{4.3_{03}}$
ACDC	before	74.2_{10}	66.1_{10}	7.1_{06}
	after	$\mathbf{75.0_{09}}^{**}$	$\mathbf{67.1_{10}}^{**}$	$\mathbf{6.9_{05}}$
CHAOS	before	74.0_{12}	70.3_{12}	9.1_{04}
	after	$\mathbf{74.3_{12}}^{**}$	$\mathbf{70.5_{12}}^{**}$	9.1_{04}
LVSC	before	62.6_{15}	53.1_{14}	5.8_{04}
	after	$\mathbf{65.9_{12}}^{**}$	$\mathbf{56.2_{12}}^{**}$	$\mathbf{5.7_{03}}$

Table 1: Dice (↑), IoU (↑) and Hausdorff distance (↓) obtained before and after tuning the segmentor on the individual test instances. Arrows show metric improvement direction; numbers are the average performance, with standard deviation as subscript; best results are in **bold**. Observe how adversarial Test-time Training always improves performance (bootstrapped t-test, $^{*}p < 0.05$, $^{**}p < 0.01$).

differences between adversarial TTT and post-processing operations, reporting also complementary performance gains. After that, Section 5.3 shows model potential for Online Continual Learning. Lastly, Section 5.4 discusses a limitation of the approach and defines a possible solution to overcome it building on a causal perspective: we show that including losses on image self-reconstruction during Test-time Training makes adaptation more reliable and faster.

5.1 Adversarial Test-time Training Under Distribution Shifts

We start with a qualitative example of test-time adaptation in Fig. 3, showing that it helps fix prediction mistakes. As can be seen on all datasets, our method corrects unrealistic masks by removing scattered false positives and segmentation holes.

In Table 1, we report segmentation performance before and after Test-time Training. We find performance improvements across metrics and datasets both in terms of metric average and spread. The only case where differences are not statistically significant is on CHAOS data, where the test set has a small number of samples (8 patients), and distributions are broad. Nevertheless, we observe empirical improvements in terms of Dice and IoU scores on CHAOS, too. From these results, we argue that adversarial TTT could lead to substantial benefits for medical applications, where systems must be robust and avoid trivial mistakes.

In Fig. 4, we compare our method with one using a shape prior separately learned by a DAE (i.e. TTANN, Karani et al. (2021)). We compare the performance increases obtained through adversarial TTT vs using TTANN, and discuss the pros and cons of driving the adaptation using a mask discriminator vs a DAE. Our experiments show advantages in using

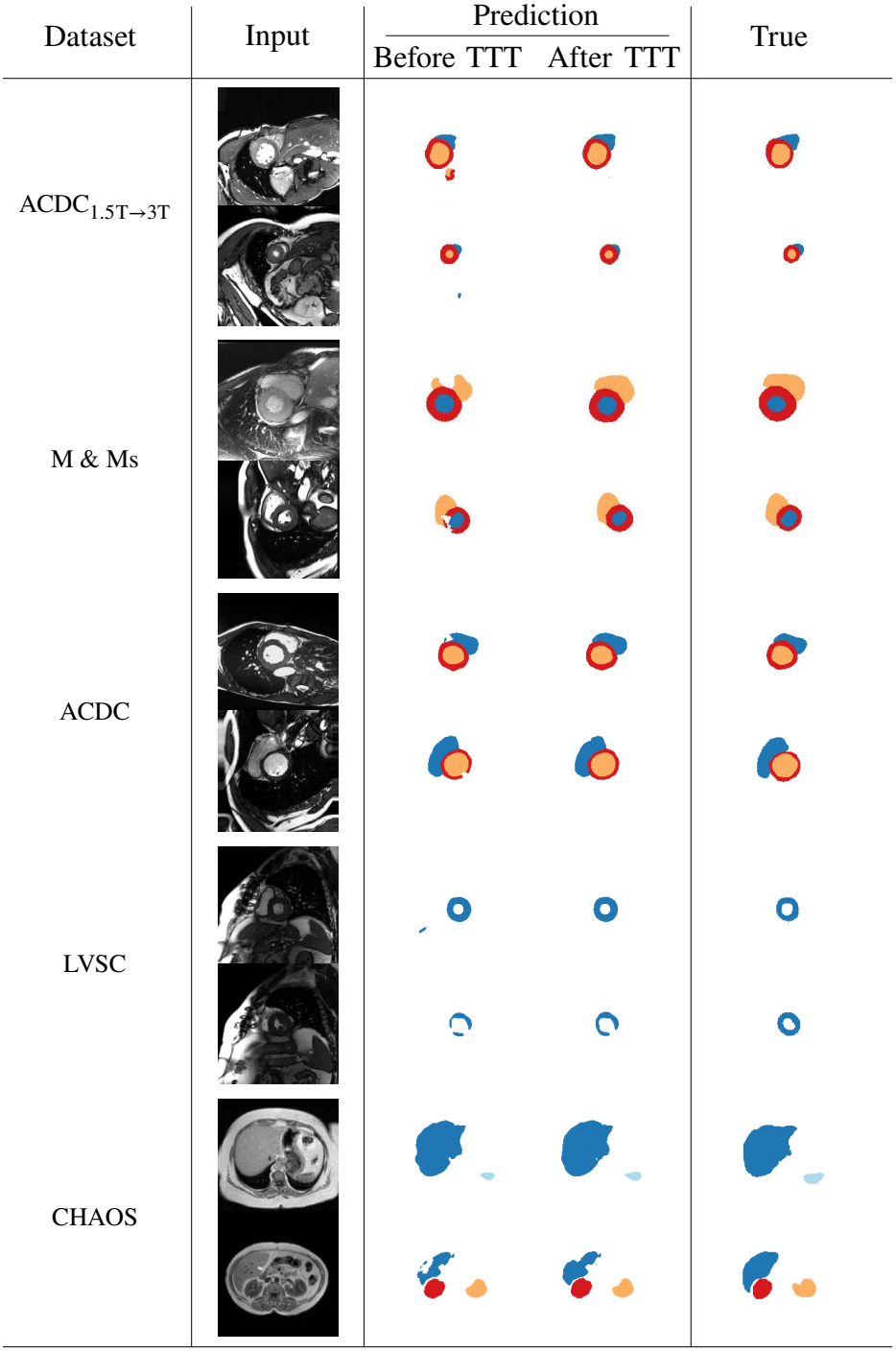

Figure 3: We show examples of prediction mistakes and their corrections after the adversarial Test-time Training. We group pairs of examples by dataset. As can be observed, the segmentor corrects the initially erroneous segmentation masks to make them realistic, according to the learned adversarial shape prior.

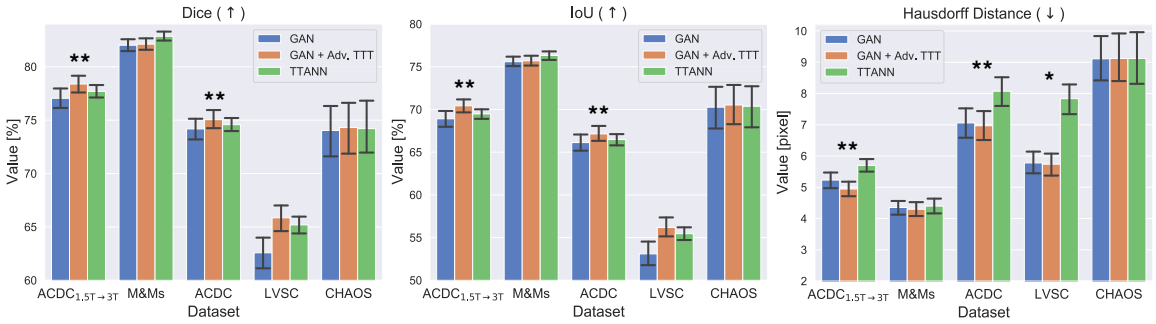

Figure 4: Adversarial TTT has competitive performance with TTANN (Karani et al., 2021), and it has the advantage of re-using an already available GAN component. Bar plots report average performance and standard errors. Stars on top of the bar plots show if differences between adversarial TTT and TTANN are statistically significant (bootstrapped t-test, $^{*}p < 0.05$, $^{**}p < 0.01$).

our method. Although performance gains appear small and TTANN performs better on M&Ms data, using adversarial TTT leads to statistically significant improvements in most of the cases. Probably, the performance increase derives from the optimisation procedure, as we train the discriminator to detect the segmentor mistakes. On the contrary, DAEs are optimised independently of the segmentor by only artificially simulating prediction mistakes. Thus, DAEs may have never seen specific mask corruptions during training, as also observed by Larrazabal et al. (2020).

Ablation Study: In Table 2 we show results ablating the adaptor, the smoothness constraints and the proposed *fake anchors* regularisation on the model. As shown, the techniques improve training and make the adversarial shape prior stronger. Consequently: i) the adversarial loss trains a better segmentor, and ii) the re-usable discriminator can increase test-time performance. For comparison, training a simple UNet on the same data leads to an average Dice score of 70.1 (standard deviation of 13). We also report an ablation study comparing the contributions of the each type of fake anchors regulariser in Table 3. In this experiment, we trained the model using no fake anchors regulariser, only patch swap, only binary noise, or both. After training, we evaluated the segmentor on the test set before TTT. As shown in the table, both randomly swapping segmentation patches and adding binary noise contribute to train better segmentors, and TTT improves upon it further.

5.2 Combining Adversarial TTT with Post-processing

Adversarial TTT should not be confused with post-processing operations because it does not modify the prediction independent of the input. Indeed, our approach lets the model adapt to the input image. Moreover, contrary to standard post-processing, our method has the advantage that it can also learn from a continuous stream of data, as we will show in the next section. However, this does not mean that post-processing cannot follow after test-time training, which we now explore.

	Adaptor Ω	Smoothness Constraints	Fake Anchors	Adversarial TTT	Performance
Ours	✓	✓	✓	✓	75.0_{09}
#1	✓	✓	✓		74.2_{10}
#2	✓	✓			72.8_{12}
#3	✓				72.7_{12}
#4					70.0_{12}

Table 2: Ablation Study. We compare the performance of our method (Ours) after removing: adversarial Test-time Training (ablation #1), the proposed regularisation technique (*fake anchors*, #2), the smoothness constraints discussed in Section 3.2 (ablation #3), and the adaptor (standard GAN, #4). Performance is in terms of average Dice score on ACDC data, with standard deviation as subscript.

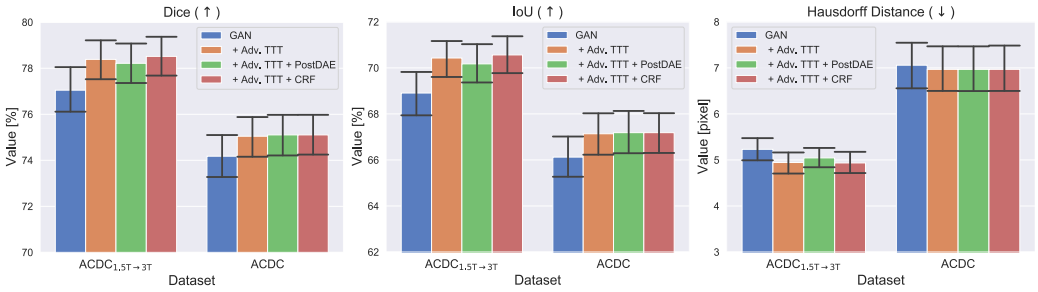

Figure 5: Compatibility with post-processing techniques (PostDAE and CRF). Bar plots report average performance and standard errors.

As examples, we consider two popular post-processing techniques. First, we examine post-processing with Conditional Random Fields (CRF), as in the DeepLab framework (Chen et al., 2017). Note that CRF adapts the *predicted mask* to the image, assigning nearby pixels with similar colours to the same semantic class. Our method, instead, adapts the *model* to the image. Second, we consider correcting the segmentation mistakes with a Denoising Autoencoder, as in PostDAE (Larrazabal et al., 2020). This method maps corrupted masks on a previously learned manifold of realistic masks without considering the associated images.

	None	Patch Swap	Binary Noise	Both	Both + Adversarial TTT
Performance	72.8_{12}	73.2_{11}	72.9_{10}	74.2_{10}	75.0_{09}

Table 3: Ablation study on the fake anchors regularisers. Performance are in terms of average Dice score, with standard deviation as subscript, on the ACDC test set. Both patch swapping and adding binary noise contribute to train better segmentors.

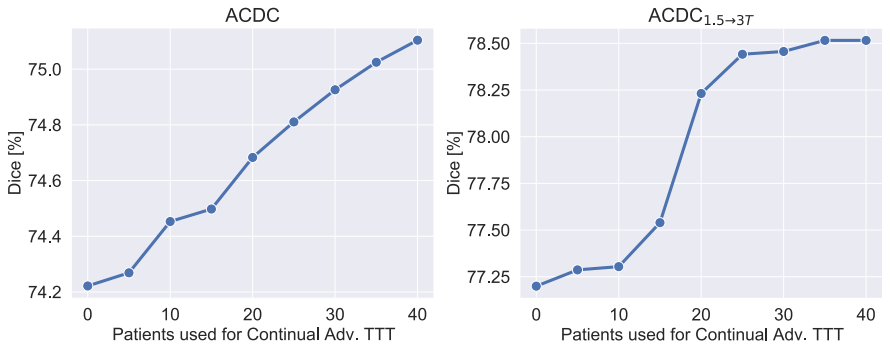

Figure 6: Effect of increasing the number of patients used to adapt the model in Online Continual Learning. For this experiment, we continually adapt the model on k test patients, and then perform standard inference on the remaining test patients. We span k between 0 (i.e. no adaptation) and 40 (i.e. continual learning on the whole test set). As new data become available, the model adapts to new distribution shifts and improves overall test performance.

Fig. 5 shows that our method can be combined with both techniques and, sometimes, the combination can improve performance. We find that PostDAE does not always help: probably because adversarial TTT already adapts the model using a data-driven shape prior (via the discriminator) and the additional DAE may be uninformative or even harmful. On the contrary, CRF increases performance because it introduces a different type of prior in the model (Zheng et al., 2015), from which the segmentor can benefit (similar to what happens in model ensembling).

5.3 Online Continual Learning of Adversarial TTT

We now experiment with the possibility of using our method for Online Continual Learning (Delange et al., 2021; Mai et al., 2021), i.e. learning from a continuous stream of non-stationary data (in our case, data affected by distribution shifts). Learning from new data, the model performance should gradually increase. Moreover, as the model gets better on the test distribution, the need for TTT should decrease, making Test-time Training faster.

We conduct experiments for both ACDC and $\text{ACDC}_{1.5\to3\text{T}}$, and report results in Figure 6 and Table 4. In this continual learning scenario, we do not restart TTT from zero when testing new data, but we continue the learning process from one patient in the test set to another. For each patient we perform the adaptation only once.

Overall, we find that the segmentor benefits from the continuous stream of test data, with performance increasing from one patient to another (Fig. 6), achieving even higher scores than using adversarial TTT on each test subject separately (Table 4).

More interestingly, we find that the average number of TTT steps needed for tuning the adaptor in Online Continual Learning decreases from 322 to 315 on ACDC data and from 120 to 114 on $\text{ACDC}_{1.5\to3\text{T}}$. This reduced number of steps suggests that gradually introducing

Dataset	Adv. TTT	Continual	Dice (↑)	IoU (↑)	Hausdorff Distance (↓)
ACDC	✗	✗	74.2_{10}	66.1_{10}	7.1_{05}
	✓	✗	75.0_{09}	67.1_{10}	6.9_{05}
	✓	✓	$\mathbf{75.1_{09}}$	$\mathbf{67.2_{09}}$	$\mathbf{6.9_{05}}$
$ACDC_{1.5\rightarrow 3T}$	✗	✗	77.0_{09}	68.9_{09}	5.2_{02}
	✓	✗	78.4_{08}	70.4_{08}	5.0_{02}
	✓	✓	$\mathbf{78.6_{08}}$	$\mathbf{70.6_{08}}$	$\mathbf{4.9_{02}}$

Table 4: Online Continual Learning. Our model can continuously learn from a stream of test data, gradually improving segmentation performance. Numbers are average performance, with standard deviation as subscript. Best results in **bold**.

new knowledge into the model lessens the need for adaptation, and the segmentor might be able to do without TTT after a while.

5.4 Towards Causal Test-time Training

We experimentally observed that, in few and rare cases, adversarial TTT can make segmentation worse (see such an example in Fig. 7). This happens because the discriminator learns to approximate the shape prior characterised by the probability distribution $p(\mathbf{y})$ rather than the joint distribution $p(\mathbf{x}, \mathbf{y})$. Thus, the discriminator will not penalise a realistic mask even when it is the wrong segmentation for the given image (see Fig. 7). Hence, it is natural to wonder whether considering both the image and the segmentation mask to drive the adaptation process may provide additional context and help Test-time Training. We emphasise that this problem also exists in TTANN (Karani et al., 2021) and in all the methods learning the marginal $p(\mathbf{y})$ instead of the joint distribution of images and masks.

As we discuss in Appendix B, we can give a causal interpretation to this problem, highlighting that image-related information would make Test-time Training causal and more effective. Below, we investigate if we can include such information and how it affects TTT.

For our experiments, we consider a method designed for semi-supervised segmentation learning: SDNet (Chartsias et al., 2019), which we describe schematically in Fig. 8. SDNet disentangles images using an encoder to find latent representations regarding the image content (anatomy) and style (modality), and a decoder to reconstruct the input images. Task specific networks learn to predict segmentations from the content latent, using either supervised losses or adversarial costs. As such, SDNet learns the joint distribution between images and masks (for additional details: see Appendix B), making SDNet a perfect test bench for evaluating both the use of discriminators and image reconstruction for TTT.

We explore if reconstructing the test sample at inference adds benefits during adaptation in terms of *performance* and *adaptation speed*. We include an adaptor Ω in front of SDNet (as shown in Fig. 8). Then, we train the full model according to the SDNet training objectives, which include an adversarial loss $\mathcal{L}_A$ and a reconstruction loss $\mathcal{L}_R$.

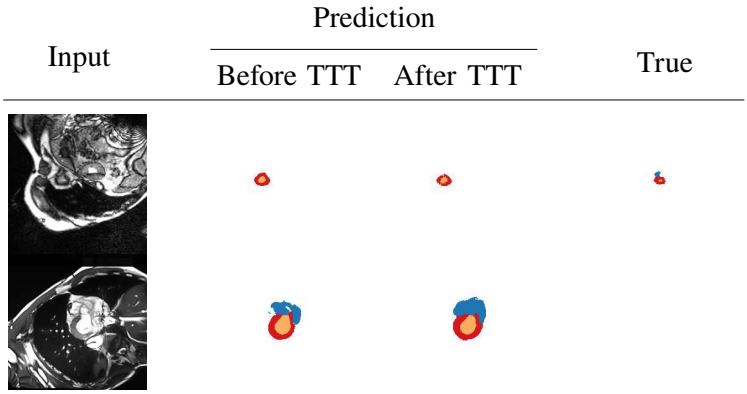

Figure 7: Since the information contained in the predicted mask is limited, the discriminator does not penalise realistic but wrong predictions (**top row**). Sometimes, it may even encourage to make bigger mistakes (**bottom row**).

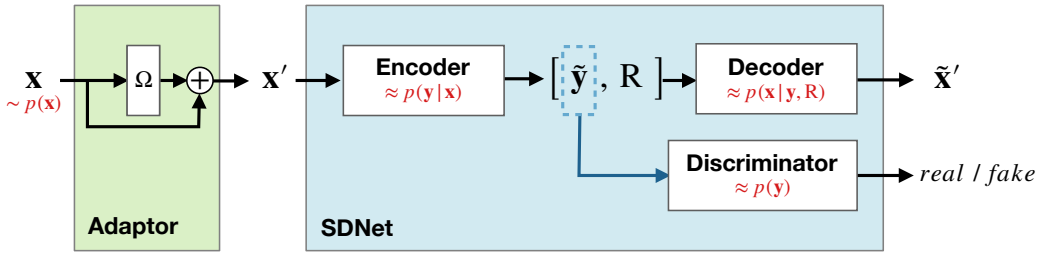

Figure 8: The proposed approach in a causal setting. We add the adaptor Ω in front of SDNet (Chartsias et al., 2019) to transform an image $\mathbf{x} \sim p(\mathbf{x})$ into its adapted version $\mathbf{x}'$. **During training**, the SDNet encoder extracts the segmentation mask $\tilde{\mathbf{y}}$ and a residual representation R. A decoder uses both of them to reconstruct the adapted image, predicting $\tilde{\mathbf{x}}' \approx \mathbf{x}'$. Meanwhile, a mask discriminator learns to tell apart real segmentation masks from the predicted ones. **At inference**, we perform Test-time Training and adapt Ω to minimise the sum of the reconstruction cost (computed comparing $\mathbf{x}'$ and $\tilde{\mathbf{x}}'$) and the adversarial loss (computed on the predicted $\tilde{\mathbf{y}}$ according to Eq. 2).

The adversarial loss $\mathcal{L}_A$ is the same as defined in Eq. 2, and we also follow the precautions discussed in Section 3.2. We leave the reconstruction term $\mathcal{L}_R$ as in the original SDNet framework, to minimise the mean absolute error between an image and its reconstruction. However, we train the model to reconstruct the adapted image $\mathbf{x}' = \Omega(\mathbf{x})$ rather than the input $\mathbf{x}$. We motivate this specific change by observing that under a distribution shift between training and inference data, the SDNet decoder may not be able to reconstruct the

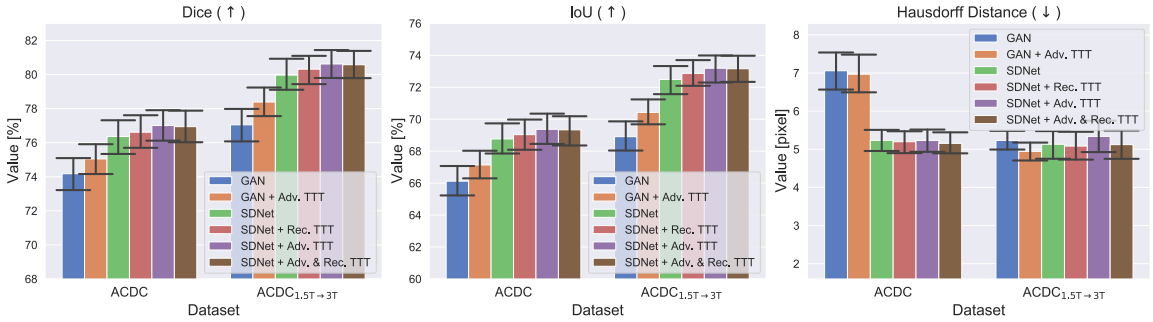

Figure 9: We compare the performance of: a GAN before and after adversarial Test-time Training; the SDNet model (discussed in Section 5.4); the SDNet after Test-time Training performed minimising only a reconstruction cost ("+ Rec. TTT"), only an adversarial cost ("+ Adv. TTT") and their sum ("+ Adv. & Rec. TTT"). Bar plots report average performance and standard errors.

test image correctly. Instead, after tuning Ω to the test image, the SDNet can effectively reconstruct the adapted image $\mathbf{x}'$.[6] We leave the rest of the SDNet model unchanged.

During inference, we fix the SDNet weights, and do Test-time Training to tune the adaptor on each sample. We set the number of TTT steps n_{iter} as described in Section 3.4, using the adaptation loss described in three different settings:

- "SDNet + Rec. TTT", where we do TTT using only the reconstruction loss $\mathcal{L}_R$;
- "SDNet + Adv. TTT", where we drive adaptation using only the adversarial loss $\mathcal{L}_A$;
- "SDNet + Adv. & Rec. TTT", where we use the sum of the adversarial and the reconstruction cost $\mathcal{L}_{tot} = \mathcal{L}_A + \mathcal{L}_R$, leading to a consistent causal-driven adaptation.

For the experiments, we considered both the case of clearly identifiable distribution shifts ($\text{ACDC}_{1.5\to 3\text{T}}$ data) and non-identifiable shifts (ACDC data). We report per-dataset results in terms of segmentation quality in Fig. 9.

From this figure, we observe that all three types of TTT improve SDNet performance, confirming that the framework is general and widely applicable. In fact, both the adversarial discriminator and the decoder used to reconstruct the image provide useful priors to drive the adaptation process. There is only one experimental exception: the Hausdorff distance of "SDNet + Adv. TTT" on $\text{ACDC}_{1.5\to 3\text{T}}$ data. In this case, while Dice and IoU scores increase, the Hausdorff distance worsens. We believe this happens because "SDNet + Adv. TTT" makes more errors in the most apical and basal slices of the heart[7]: a behaviour that we also observe for "GAN" and "GAN + Adv. TTT", where Hausdorff distances are high.

6. An alternative to reconstructing $\mathbf{x}'$ would be to introduce an "inverted" adaptor Ω^{-1} at the decoder output. However, this would require extra computational cost and reconstructing $\mathbf{x}'$ is simpler.
7. By definition, the Hausdorff distance between two binary masks has the maximum possible value (i.e. the image size) when one of the two masks is empty. In this case, even one missed or one extra pixel in the apical and basal slices leads to high values of the metric, decreasing performance.

Analysing the contribution of the reconstruction loss in detail, we observe that it mainly helps when optimising the model on the training data (i.e. before Test-time Training). In fact, if we compare the performance of SDNet with that of a GAN *before* TTT ("SDNet" vs "GAN", in Fig. 9), we find a big improvement in all the metrics. On the contrary, when we analyse the effect of $\mathcal{L}_R$ during Test-time Training, we find that it only slightly affects the metrics (compare "SDNet + Adv. TTT" vs "SDNet + Adv. & Rec. TTT").

Instead, including $\mathcal{L}_R$ during TTT has a bigger impact on the test-time training speed. In fact, we find that the number of TTT iterations needed for convergence halves. Specifically, using only the adversarial cost during TTT, the average optimal n_{iter} is 111 on ACDC, and 206 on $\text{ACDC}_{1.5\to 3\text{T}}$ data. By adding also the reconstruction term, the average number of TTT steps becomes 66 on ACDC and 119 on $\text{ACDC}_{1.5\to 3\text{T}}$. Our results are also in line with recent findings arguing that correct causal structures adapt faster (Bengio et al., 2020).

These experiments highlight that causal TTT, using the causal structure herein, leads to marginal improvements in segmentation quality, but it makes adaptation to the test samples considerably faster.

6. Conclusion

In this work, we suggest re-using adversarial discriminators at inference. First, we identify simple design assumptions that must be satisfied to make discriminators useful once training is complete. Then, we show how to use discriminators to detect and correct segmentation mistakes on the test data. The proposed approach is simple. compatible with common post-processing techniques, and it increases test-time performance on the most challenging images. Our approach also benefits from continual learning, making test-time inference gradually more accurate and faster. Lastly, we showed that reconstruction losses could complement mask discriminators and improve the inference speed of our model.

We believe that learning without supervision on new test data is a promising research avenue. However, while this works only touches upon the potential use of discriminators and their combination with other components at inference time, there are still several challenges to solve. First, Test-time Training methods rely on gradient descent to fine-tune the model on the test data. While effective, this solution slows down model prediction time. Hence, future work should focus on strategies to make TTT faster, increasing parallelisation and possibly adapting the model using just a single training step. For example, Bartler et al. (2021) have recently explored the combination of Meta-Learning techniques with TTT, to make adaptation faster. Furthermore, when moving towards continual learning, alleviating the risk of forgetting previous experience is crucial. To this end, it would be interesting to combine our method with approaches aiming to solve these challenges, such as Elastic Weight Consolidation (Kirkpatrick et al., 2017).

More generally, re-using adversarial discriminators to fix generator mistakes may open several research opportunities, even outside image segmentation. Thanks to their flexibility and the ability to learn data-driven losses, GANs have been widely adopted in medical imaging, from domain adaptation to image synthesis tasks (Yi et al., 2019). In this context, we believe that improved architectures and regularisation techniques (Kurach et al., 2019; Chu et al., 2020) will make adversarial networks even more popular. Thus, training stable

and re-usable discriminators opens opportunities for re-using flexible data-driven losses at test time and make inference better.

Acknowledgments

This work was partially supported by the Alan Turing Institute (EPSRC grant EP/N510129/1). S.A. Tsaftaris acknowledges the support of Canon Medical and the Royal Academy of Engineering and the Research Chairs and Senior Research Fellowships scheme (grant RCSRF1819\8\25). We thank NVIDIA for donating the GPU used for this research.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we don't have conflicts of interest.

References

Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. A Critical Analysis of Self-Supervision, or What We Can Learn From a Single Image. *International Conference on Learning Representations (ICLR)*, 2020.

Alexander Bartler, Andre Bühler, Felix Wiewel, Mario Döbler, and Bin Yang. MT3: Meta Test-Time Training for Self-Supervised Test-Time Adaption. *arXiv preprint arXiv:2103.16201*, 2021.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms. *International Conference on Learning Representations (ICLR)*, 2020.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Išgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.

Victor M. Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M. Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, et al. Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge. *IEEE Transactions on Medical Imaging*, 2021.

Daniel C. Castro, Ian Walker, and Ben Glocker. Causality Matters in Medical Imaging. *Nature Communications*, 11(1):1–10, 2020.

Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David E. Newby, Rohan Dharmakumar, and Sotirios A. Tsaftaris. Disentangled Representation Learning In Cardiac Image Analysis. *Medical Image Analysis*, 58: 101535, 2019.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic Image Segmentation With Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

Veronika Cheplygina, Marleen de Bruijne, and Josien P. W. Pluim. Not-so-Supervised: A Survey of Semi-Supervised, Multi-Instance, and Transfer Learning In Medical Image Analysis. *Medical Image Analysis*, 54:280–296, 2019.

Casey Chu, Kentaro Minami, and Kenji Fukumizu. Smoothness and Stability in GANs. *International Conference on Learning Representations (ICLR)*, 2020.

James R. Clough, Ilkay Oksuz, Nicholas Byrne, Veronika A. Zimmer, Julia A. Schnabel, and Andrew P. King. A Topological Loss Function for Deep-Learning Based Image Segmentation Using Persistent Homology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Adrian V. Dalca, John Guttag, and Mert R. Sabuncu. Anatomical Priors in Convolutional Networks for Unsupervised Biomedical Segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9290–9299, 2018.

Adrian V. Dalca, Evan Yu, Polina Golland, Bruce Fischl, Mert R. Sabuncu, and Juan Eugenio Iglesias. Unsupervised Deep Learning For Bayesian Brain MRI Segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 356–365. Springer, 2019.

Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A Continual Learning Survey: Defying Forgetting in Classification Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial Feature Learning. *International Conference on Learning Representations (ICLR)*, 2017.

Qi Dou, Daniel C. Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain Generalization via Model-Agnostic Learning Of Semantic Features. *arXiv preprint arXiv: 1910.13580*, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014.

Hao Guan and Mingxia Liu. Domain Adaptation for Medical Image Analysis: A Survey. *arXiv preprint arXiv:2102.09508*, 2021.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved Training of Wasserstein GANs. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NeurIPSs 30*, pages 5767–5777. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/7159-improved-training-of-wasserstein-GANs.pdf`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning (ICML)*, pages 1321–1330. PMLR, 2017.

Yufan He, Aaron Carass, Lianrui Zuo, Blake E Dewey, and Jerry L Prince. Autoencoder Based Self-Supervised Test-Time Adaptation for Medical Image Analysis. *Medical Image Analysis*, page 102136, 2021.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training By Reducing Internal Covariate Shift. In *International Conference on Machine Learning (ICML)*, pages 448–456. PMLR, 2015.

Thomas Joyce, Agisilaos Chartsias, and Sotirios A. Tsaftaris. Robust Multi-Modal MR Image Synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 347–355. Springer, 2017.

Rosana El Jurdi, Caroline Petitjean, Paul Honeine, Veronika Cheplygina, and Fahed Abdallah. High-level Prior-based Loss Functions for Medical Image Segmentation: A Survey. *arXiv preprint arXiv:2011.08018*, 2020.

Rosana El Jurdi, Caroline Petitjean, Paul Honeine, Veronika Cheplygina, and Fahed Abdallah. A Surprisingly Effective Perimeter-based Loss for Medical Image Segmentation. *Medical Imaging with Deep Learning (MIDL)*, 2021.

Neerav Karani, Ertunc Erdil, Krishna Chaitanya, and Ender Konukoglu. Test-Time Adaptable Neural Networks for Robust Medical Image Segmentation. *Medical Image Analysis*, 68:101907, 2021.

A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. CHAOS Challenge-combined (CT-MR) Healthy Abdominal Organ Segmentation. *Medical Image Analysis*, 69:101950, 2021.

Hoel Kervadec, Jose Dolz, Meng Tang, Eric Granger, Yuri Boykov, and Ismail Ben Ayed. Constrained-CNN Losses for Weakly Supervised Segmentation. *Medical Image Analysis*, 54:88–99, 2019.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations (ICLR)*, 2017.

Youngjin Kim, Minjung Kim, and Gunhee Kim. Memorization Precedes Generation: Learning Unsupervised GANs With Memory Networks. In *International Conference on Machine Learning (ICML)*, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*, 2015.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming Catastrophic Forgetting in Neural Networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Karol Kurach, Mario Lučić, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. A Large-Scale Study on Regularization and Normalization in GANs. In *International Conference on Machine Learning (ICML)*, pages 3581–3590. PMLR, 2019.

Agostina J. Larrazabal, César Martínez, Ben Glocker, and Enzo Ferrante. Post-DAE: Anatomically plausible segmentation via post-processing with Denoising Autoencoders. *IEEE Transactions on Medical Imaging*, 2020.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain Generalization With Adversarial Feature Learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5400–5409, 2018.

Yuejiang Liu, Parth Kothari, and Alexandre Alahi. Collaborative Sampling in Generative Adversarial Networks. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 4948–4956, 2020.

Xudong Mao, Qing Li, Haoran Xie, Raymond Yiu Keung Lau, Zhen Wang, and Stephen Paul Smolley. On the Effectiveness of Least Squares Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. Online Continual Learning in Image Classification: An Empirical Survey. *arXiv preprint arXiv:2101.10423*, 2021.

Xin Mao, Zhaoyu Su, Pin Siang Tan, Jun Kang Chow, and Yu-Hsing Wang. Is Discriminator a Good Feature Extractor? *arXiv preprint arXiv:1912.00789*, 2019.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral Normalization for Generative Adversarial Networks. *International Conference on Learning Representations (ICLR)*, 2018.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When Does Label Smoothing Help? In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4694–4703, 2019.

Vaishnavh Nagarajan, Colin Raffel, and I. Goodfellow. Theoretical Insights Into Memorization in GANs. In *NeurIPS Workshop*, 2018.

Phuc Cuong Ngo, Amadeus Aristo Winarto, Connie Khor Li Kou, Sojeong Park, Farhan Akram, and Hwee Kuan Lee. Fence GAN: Towards Better Anomaly Detection. In *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 141–148. IEEE, 2019.

Masoud S. Nosrati and Ghassan Hamarneh. Incorporating Prior Knowledge in Medical Image Segmentation: A Survey. *arXiv:1607.01092*, 2016.

Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Stuart A. Cook, Antonio de Marvao, Timothy Dawes, Declan P. O'Regan, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation. *IEEE Transactions on Medical Imaging*, 37(2):384–395, 2017.

Nathan Painchaud, Youssef Skandarani, Thierry Judge, Olivier Bernard, Alain Lalande, and Pierre-Marc Jodoin. Cardiac MRI Segmentation With Strong Anatomical Guarantees. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 632–640. Springer, 2019.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning With Deep Convolutional Generative Adversarial Networks. *arXiv preprint arXiv:1511.06434*, 2015.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do Cifar-10 Classifiers Generalize to Cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. Semi-supervised Learning in Causal and Anticausal Settings. In *Empirical Inference*, pages 129–141. Springer, 2013.

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning From Simulated and Unsupervised Images Through Adversarial Training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2107–2116, 2017.

Abhishek Sinha, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, and Stefano Ermon. Negative Data Augmentation. *International Conference on Learning Representations (ICLR)*, 2021.

Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. *International Conference on Learning Representations (ICLR)*, 2017.

Avan Suinesiaputra, Brett R. Cowan, Ahmed O. Al-Agamy, Mustafa A. Elattar, Nicholas Ayache, Ahmed S. Fahmy, Ayman M. Khalifa, Pau Medrano-Gracia, Marie-Pierre Jolly, Alan H. Kadish, Daniel C. Lee, Ján Margeta, Simon K. Warfield, and Alistair A. Young. A Collaborative Resource to Build Consensus for Automated Left Ventricular Segmentation of Cardiac MR Images. *Medical Image Analysis*, 18(1):50–62, 2014.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-Time Training With Self-Supervision for Generalization Under Distribution Shifts. In *International Conference on Machine Learning (ICML)*, pages 9229–9248. PMLR, 2020.

Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N. Chiang, Zhihao Wu, and Xiaowei Ding. Embracing Imperfect Datasets: A Review of Deep Learning Solutions for Medical Image Segmentation. *Medical Image Analysis*, page 101693, 2020.

Gabriele Valvano, Andrea Leo, and Sotirios A. Tsaftaris. Stop Throwing Away Discriminators! Re-using Adversaries for Test-Time Training. In *Domain Adaptation and Representation Transfer (DART)*. 2021a.

Gabriele Valvano, Andrea Leo, and Sotirios A. Tsaftaris. Learning to Segment From Scribbles Using Multi-Scale Adversarial Attention Gates. *IEEE Transactions on Medical Imaging*, 2021b. doi: 10.1109/TMI.2021.3069634.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, Trevor Darrell, UC Berkeley, and Adobe Research. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations (ICLR)*, volume 4, page 6, 2021.

Xin Yi, Ekta Walia, and Paul Babyn. Generative Adversarial Network in Medical Imaging: A Review. *Medical Image Analysis*, 58:101552, 2019.

Houssam Zenati, Manon Romain, Chuan-Sheng Foo, Bruno Lecouat, and Vijay Chandrasekhar. Adversarially Learned Anomaly Detection. In *IEEE International Conference on Data Mining (ICDM)*, pages 727–736. IEEE, 2018.

Marvin Zhang, Sergey Levine, and Chelsea Finn. MEMO: Test Time Robustness via Adaptation and Augmentation. *arXiv preprint arXiv:2110.09506*, 2021.

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional Random Fields as Recurrent Neural Networks. In *International Conference on Computer Vision (ICCV)*, pages 1529–1537, 2015.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain Generalization: A Survey. *arXiv preprint arXiv:2103.02503*, 2021.

Appendix A. Discriminator: Convergence and Memorisation

During adversarial training, we optimise the segmentor to produce realistic masks, which the discriminator cannot distinguish from the real ones. At convergence, the discriminator can either reach an equilibrium stage or collapse.

Below, we show empirical examples of the training and validation losses for the GAN discriminator Δ. For these experiments, we use a Least-square GAN (Mao et al., 2018), whose discriminator loss to minimise is:

$$\mathcal{V}_{LS}(\Delta) = \frac{1}{2}\underbrace{E_{\mathbf{y}\sim\mathcal{Y}}[(\Delta(\mathbf{y})-1)^2]}_{\text{loss on } \textit{real} \text{ samples}} + \frac{1}{2}\underbrace{E_{\mathbf{x}\sim\mathcal{X}}[(\Delta(\Sigma(\mathbf{x}))+1)^2]}_{\text{loss on } \textit{fake} \text{ samples}} \tag{3}$$

where $+1$ and -1 are the labels for *real* and *fake* (generated) images, respectively.

We report examples of convergence modes in Fig. 10 and Fig. 11. On the left side: losses on the training set; on the right: losses on the validation set. Observe that – despite the single loss components have different values – the total loss $\mathcal{V}_{LS}(\Delta)$ on the validation set is the same in both cases.

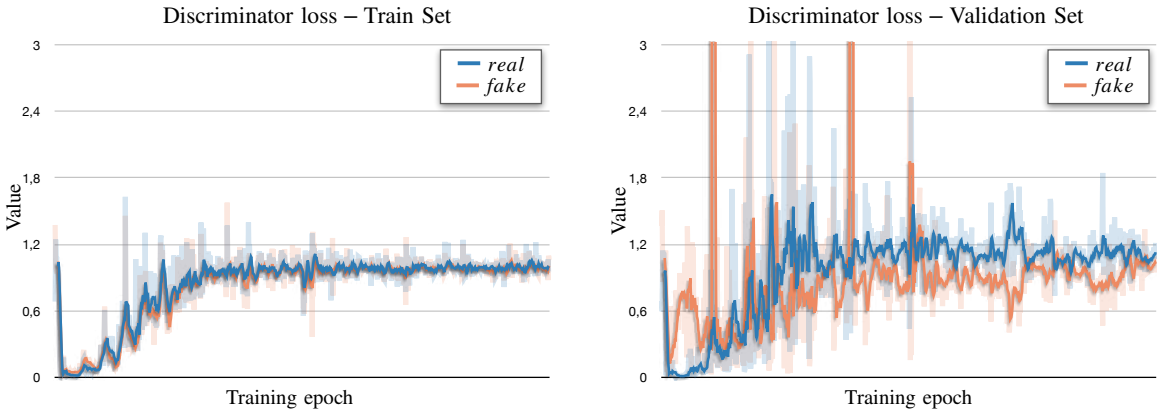

Figure 10: At convergence, the discriminator reaches an equilibrium where it always predicts the value 0, equidistant from the *real* and the *fake* labels. Thus, losses for *real* and *fake* masks converge to the equilibrium value 1.0 in both train and validation.

Appendix B. Causal Test-time Training

Causal machine learning is recently gaining considerable attention in medical imaging (Castro et al., 2020) because it could identify the best suited approaches to solve a specific task (Schölkopf et al., 2013; Castro et al., 2020), or make learning faster (Bengio et al., 2020).

In our model, we optimise the segmentation modules of a GAN (Ω and Σ, described in Section 3.1) to approximate the conditional distribution $p(\mathbf{y}|\mathbf{x})$, and the discriminator to learn the marginal $p(\mathbf{y})$. Training this type of GAN has practical advantages because the discriminator regularises the segmentor and allows to use unlabelled data for training.

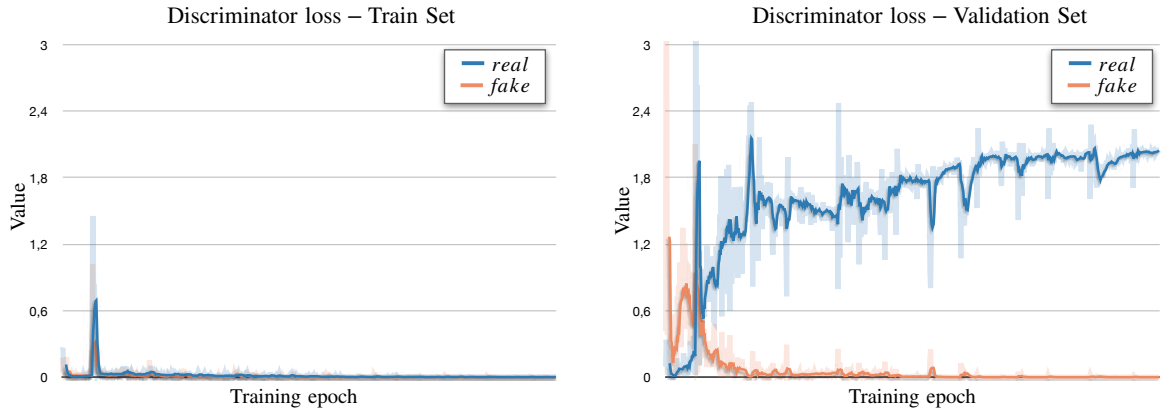

Figure 11: At convergence, the discriminator shows signals of memorisation. The discriminator memorises the *real* training images, and it predicts the label *fake* (i.e. the value -1) for any other case. During validation, the *fake* images are still classified correctly, while the *real* ones are classified as *fake* and the associated loss converges to the value of 2.0.

However, tuning the adaptor based only on the adversarial loss may be considered as driving the adaptation using an anti-causal model, while the process is instead causal. In other terms, it is an image that causes the predicted mask because experts draw masks on top of the images, and not vice versa (Castro et al., 2020). Instead, GANs whose discriminator only receives segmentation masks as input would penalise the segmentor without considering the image causing that mask.

Thus, from a causal perspective, our approach is non-optimal. To capture the causal structure better and update the model parameters, we should also consider the inverse conditional probability $p(\mathbf{x}|\mathbf{y})$, which would improve the approximation $p(\mathbf{y}|\mathbf{x})$ (i.e. the segmentation modules Ω and Σ) according to Bayes theorem:

$$p(\mathbf{y}|\mathbf{x}) = p(\mathbf{x}|\mathbf{y})\frac{p(\mathbf{y})}{p(\mathbf{x})}. \tag{4}$$

Hence, to obtain a more coherent description, we should learn an inverse function mapping the masks to their respective images: $p(\mathbf{x}|\mathbf{y})$. Unfortunately, segmentation masks do not contain all the information needed to go from $\mathbf{y}$ to $\mathbf{x}$, as one mask can be associated to many different images, also known as the *one-to-many problem*. Since this inversion is not possible, rather than learning the two components $p(\mathbf{x}|\mathbf{y})$ and $p(\mathbf{y})$ separately (Eq. 4), one may attempt to directly learn the joint distribution $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$. To have this type of model, we can optimise the discriminator Δ providing input pairs $(\mathbf{x}, \mathbf{y})$ rather than just unpaired masks. As a result, Δ would implicitly learn to approximate the joint probability distribution of image-mask pairs, rather than the distribution of masks, and we would obtain a coherent causal description. However, this approach also has several problems. In the first place, the discriminator would be subject to pixel-intensity distribution shifts of $\mathbf{x}$: thus, we would move our problem from adapting Ω to the test images,

to the problem of adapting Δ. Moreover, since the discriminator would need paired data for training, we would not be able to use the framework in semi-supervised settings where we have unpaired images and unpaired segmentation masks (such as in the scenarios of non-identifiable distribution shift, described in Section 4.1).[8]

Another alternative to learning $p(\mathbf{x}|\mathbf{y})$ is to substitute it with a proxy distribution. For example, we could learn $p(\mathbf{x}|\mathbf{y}, \mathrm{R})$, where R is a residual representation containing complementary information that is not present in $\mathbf{y}$ and is necessary to go from a mask $\mathbf{y}$ to the respective image $\mathbf{x}$, breaking the one-to-many/many-to-one problem described above. In this case, we would establish the relationship:

$$p(\mathbf{y}|\mathbf{x}) \leftrightarrow p(\mathbf{x}|\mathbf{y}, \mathrm{R})\frac{p(\mathbf{y})}{p(\mathbf{x})}. \tag{5}$$

As discussed in the paper, an example of such a model is SDNet (Chartsias et al., 2019), which uses the extracted mask and its residuals to reconstruct the image[9], while also having an adversarial discriminator learning $p(\mathbf{y})$.

8. For completeness, we also conducted an experiment in fully-supervised learning, where all the training images are associated with a segmentation mask and where the discriminator can learn the joint distribution. In this case, we observed that the discriminator was more prone to overfitting the training data, and its generalisation under distribution shifts got worse.
9. To be more precise, this holds assuming that the "anatomy encoder" of SDNet performs a segmentation task and extracts $\mathbf{y}$ within the anatomical representation of a patient. For the purpose of our experiments, we assume it is a reasonable approximation.