

IMT School for Advanced Studies, Lucca
Lucca, Italy

**Semi-supervised and Weakly-supervised Learning
with Spatio-temporal Priors
in Medical Image Segmentation**

PhD Program in Cognitive and Cultural Systems
Track in Cognitive, Computational and Social
Neurosciences
XXXIII Cycle

By

Gabriele Valvano

2021

The dissertation of Gabriele Valvano is approved.

PhD Program Coordinator:

Prof. Maria Luisa Catoni, IMT School for Advanced Studies Lucca, Italy

Advisor:

Prof. Emiliano Ricciardi, IMT School for Advanced Studies Lucca, Italy

Co-Advisor:

Prof. Andrea Leo, University of Pisa, Italy

Co-Advisor:

Prof. Sotirios A. Tsaftaris, School of Engineering, University of
Edinburgh, UK

The dissertation of Gabriele Valvano has been reviewed by:

Prof. Ender Konukoglu, ETH Zürich, Switzerland

Prof. Caroline Petitjean, Université de Rouen, France

IMT School for Advanced Studies Lucca
2021

Contents

List of Figures	xi
List of Tables	xxiv
Ringraziamenti	xxvii
Acknowledgements	xxviii
Vita and Publications	xxx
Abstract	xxxii
1 Introduction	1
1.1 Medical Motivation	1
1.2 Common Limitations of Medical Datasets	3
1.3 Prior-driven Regularisation	4
1.4 Overview and Technical Contributions	4
2 Clinical and Medical Imaging Background	9
2.1 Medical Imaging	9
2.2 Magnetic Resonance Imaging	10
2.2.1 Physical Principles	10
2.2.2 Hardware and Signal Acquisition	13
2.3 Cardiac MRI	15
2.3.1 The Heart	17
2.4 Abdominal MRI	18
2.5 Summary	19

3	Technical Background	20
3.1	Mathematical Notation	20
3.2	Learning Algorithms	21
3.2.1	Learning Paradigms with Limited Supervision	23
3.3	Direct and Indirect Regularisation of the Data Representation	25
3.4	Priors for Direct Regularisation	26
3.4.1	Classical Approaches	27
3.4.2	Disentangled Representations	27
3.5	Priors for Indirect Regularisation	29
3.5.1	Priors as Learning Objectives	30
3.5.2	Priors as Design Bias	32
3.5.3	Priors as Data Bias	32
3.6	Representation Learning and Deep Generative Models	33
3.6.1	Autoencoders (AE)	34
3.6.2	Variational Autoencoders (VAE)	34
3.6.3	Generative Adversarial Networks (GAN)	37
3.7	Segmentation Metrics	41
3.8	Datasets	42
3.8.1	ACDC: Automatic Cardiac Diagnosis Challenge	43
3.8.2	LVSC: Left Ventricular Segmentation Challenge	45
3.8.3	M&Ms: Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Image Segmentation Challenge	45
3.8.4	CHAOS: Combined Healthy Abdominal Organ Segmentation	46
3.8.5	PPSS: Pedestrian Parsing in Surveillance Scenes	47
3.8.6	Data Augmentation	47
3.9	Summary	48
4	Regularising Disentangled Representations using Spatio-Temporal Priors	50
4.1	Introduction	51
4.1.1	Contributions	53
4.2	Related Work	54
4.2.1	Semi-Supervised Learning	55

4.2.2	Disentangled Representations	56
4.2.3	Improving Disentanglement with Temporal Transitions	57
4.3	Methods	59
4.3.1	Spatial Decomposition Network (SDNet)	59
4.3.2	Spatial Decomposition and Transformation Network (SDTNet)	62
4.4	Experiments	67
4.4.1	Data	67
4.4.2	Temporal Axis	67
4.4.3	Baselines and Evaluation	67
4.5	Results and Discussion	68
4.5.1	Semi-supervised segmentation	68
4.5.2	What does the model learn?	71
4.6	Conclusion	74
4.7	Summary	75
5	Learning Adversarial Shape Priors with GANs	77
5.1	GANs and Adversarial Shape Priors	78
5.1.1	Techniques to Improve GAN Training	80
5.1.2	Popular GAN Variants	81
5.2	GANs for Unsupervised and Semi-supervised Image Segmentation	82
5.3	GAN Regularisation: Exploiting Texture Bias in Mask Discriminators	87
5.3.1	Introduction	87
5.3.2	Related Work	88
5.3.3	Method	90
5.3.4	Experimental Setup	93
5.3.5	Results	95
5.4	Summary	98
6	Multi-scale Adversarial Shape Priors for Weak Supervision	100
6.1	Introduction	101
6.1.1	Overview of the proposed approach	102

6.1.2	Contributions	103
6.2	Related Work	105
6.2.1	Learning from Scribbles	105
6.2.2	Shape Priors in Deep Learning for Medical Imaging	106
6.2.3	Multi-scale GANs	107
6.2.4	Attention Gates	108
6.3	Proposed Approach	109
6.3.1	Method Overview	109
6.3.2	Architectures	110
6.3.3	Loss Functions and Training Details	113
6.4	Experimental Setup	115
6.4.1	Data	115
6.4.2	Baseline, Benchmark Methods and Upper Bounds .	118
6.5	Experiments and Discussion	120
6.5.1	Learning from Scribbles	121
6.5.2	Segmentation Masks vs Scribbles	123
6.5.3	Model Robustness to Limited Annotations	125
6.5.4	How Much Does the Model Rely on the Unpaired Data?	127
6.5.5	Combining Multiple Scribbles: Simulating Crowd- sourcing	127
6.5.6	Multitask Learning: Combining Masks and Scribbles	129
6.5.7	Why does Adversarial Attention Gating work? . . .	129
6.6	Conclusion	133
6.7	Summary	133
7	Self-supervised Multi-scale Consistency for Weak Supervision	135
7.1	Introduction	136
7.1.1	Contributions	137
7.2	Related Work	138
7.2.1	Weakly-supervised Learning for Image Segmenta- tion	138
7.2.2	Multi-scale Consistency and Attention	139

7.2.3	Self-supervised Learning for Medical Image Segmentation	139
7.3	Proposed Approach	140
7.3.1	Method Overview	140
7.3.2	Model Architecture and Training	140
7.4	Experiments	143
7.4.1	Data	143
7.4.2	Evaluation Protocol	144
7.4.3	Results	146
7.5	Conclusion	149
7.6	Summary	149
8	Re-using Adversarial Shape Priors for Test-Time Training	151
8.1	Introduction	152
8.1.1	Contributions	153
8.2	Related Work	153
8.2.1	Learning from Test Samples	153
8.2.2	Tackling Distribution Shifts	155
8.2.3	Shape Priors in Deep Learning for Medical Image Segmentation	155
8.2.4	Re-using Adversarial Discriminators	156
8.3	Method	156
8.3.1	Method Overview	156
8.3.2	Re-usable Discriminators: Challenges and Proposed Solutions	157
8.3.3	Architectures and Training Objectives for $\Sigma(\cdot)$ and $\Delta(\cdot)$	159
8.3.4	Adversarial Test-Time Training: Adapting $\Omega(\cdot)$	161
8.4	Experimental Setup	162
8.4.1	Data	162
8.4.2	Evaluation Protocol	163
8.5	Experiments and Discussion	163
8.5.1	Adversarial Test-time Training Under Distribution Shifts	165

8.5.2	Limitations and a Possible Solution	167
8.5.3	Toward Causal Test-time Training	168
8.5.4	Combining Adversarial TTT with Post-processing Operations	173
8.5.5	Online Continual Learning	175
8.6	Conclusion	176
9	Summary and Future Directions	177
9.1	Summary	177
9.2	Future Directions	179
A	Experimental Details of Chapter 5	183
A.1	Experimental Details of Section 5.2 and 5.3	183
A.1.1	Model Architectures	183
A.1.2	Optimisation	184
B	Additional Experiments and Results of Chapter 6	185
B.1	Dice score and Hausdorff Distance for Single Anatomical Regions	185
B.2	Results on ACDC Evaluation Platform	188
B.3	The Effect of the Dynamic Loss Weighting	188
B.4	Fully Supervised Learning	189
B.5	Additional Figures	189
C	Additional Experiments and Results of Chapter 8	191
C.1	Discriminator: Convergence and Memorisation	191

List of Figures

- 1 The effect of a radiofrequency pulse (RF) at the Larmor frequency on a set of nuclei aligned with the magnetic field B_0 . The orthogonal RF wave tilts the nuclei spins, producing components of the net magnetisation on the xy plane. The angle α between the z axis and the tilted net magnetisation vector M has the name of flip angle. 11
- 2 Summary of the physical principle behind the generation of MR signals. Inside the static magnetic field B_0 , the spins align parallel or anti-parallel to the direction of B_0 , with a net magnetisation $M > 0$. Excited by a radiofrequency (RF) pulse, the net magnetisation tilts on the xy plane. Once terminated the RF pulse, spins gradually re-align themselves with the z axis, releasing energy in the form of an RF wave that a receiver coil can capture. At the same time, spins also accumulate a different phase in their precession movement, further decaying the net transverse magnetisation. 14
- 3 Different tissues exhibit different magnetic properties. Acquiring a signal at time t_s will therefore result in a different signal amplitude measured by the receiver RF coil. 14

4	Left: substructures of the human heart. Right: the Wiggers diagram shows the temporal association between the electrical and the mechanical signals that it is possible to measure during the cardiac cycle. Figures are adapted from Wikipedia, 2020a and Wikipedia, 2020b , respectively.	16
5	Example of MR images of a 23-year old subject. ¹ On the left, MRI on the subject thorax in two orthogonal planes. On the right, a short-axis cardiac MRI.	17
6	Example of abdominal MRIs of 4 different subjects, contained in the CHAOS dataset (Kavur, Gezer, Barış, Aslan, et al., 2021). In the top row, we show examples of T1-DUAL in-phase images, while in the bottom row we report T2-SPIR images.	18
7	Comparison of different learning scenarios. Classical supervised learning requires the highest amount of supervision, while the need for fine-grained annotations diminishes moving toward unsupervised learning.	22
8	Comparison of different types of annotations. Segmentation masks (a) are the most time expensive annotations to collect. Weaker forms of annotations (b–e) are cheaper to obtain but provide less supervisory signal. The MRI image is taken from the ACDC dataset (Bernard et al., 2018). .	24
9	Variational Autoencoder (VAE) schematic. A stochastic encoder learns to map input samples to a prior probability distribution, predicting mean and variance of the distribution. The decoder attempts to reconstruct the input by drawing a sample from the predicted distribution, thanks to the reparametrization trick.	36

10	Generative Adversarial Network (GAN) schematic. A generator learns to map random inputs from a known distribution into samples of the target data distribution. Meanwhile, the discriminator is trained to say apart images generated from the generator and images sampled from the data distribution.	38
11	The generator of a GAN learns to map random inputs to samples of the data distribution. The generator is trained to minimise a distance, or a divergence, between the distributions of the real and the generated data.	39
12	Graphical visualisation of metrics used to evaluate segmentation performance.	41
13	Method overview. Given the input image x_t at time t , the model extracts a multi-channel binary representation S_t (anatomical factors) and a residual vector z (modality factors). In this work, we aim at regularising S_t constraining it to be predictable: conditioned on the temporal gap dt , a neural network must be able to predict the representation at time $t + dt$	54
14	Block diagram of SDNet (described in Section 4.3.1) and SDTNet (Section 4.3.2). The components of SDNet are represented using yellow boxes, while SDTNet also includes the transformer network Θ , represented in light blue. In SDTNet, we train Θ to predict the future modality-independent anatomical factors conditioned on the temporal information dt . Notice that improving the quality of the anatomical representation S_t can make the segmentor job easier, facilitating the extraction of high-quality segmentation masks \tilde{y}_t	59

- 15 Effect of dataset shift on a model residing in a sharp or a flat local minimum. We plot the loss landscapes of train data in grey colour and test data in red colour. Given an optimised model with parameters W^* , the same dataset shift from train to test data (blue horizontal arrows) has an increased performance impact if the model resides in a sharp minimum (dashed vertical lines). 65

- 16 Segmentation masks predicted by the considered models at various levels of training annotations on ACDC (top) and LVSC (bottom) datasets. As can be seen, using temporal consistency to regularise disentanglement (SDTNet) leads to the best performance, especially when annotations are scarce. 69

- 17 Example of anatomical factors extracted from an input image for ACDC (top row) and LVSC (bottom row) datasets. Anatomies are represented as multi-channel binary maps and can contain well defined anatomical components, such as left/right ventricle and myocardium, or other geometrical content needed for the image reconstruct through the image decoder (rightmost image). 72

- 18 Example of temporal interpolation from ED to ES cardiac phases. Images were obtained by keeping fixed the anatomical factors S_t at time $t = 0$ and ranging dt in $[0, 1]$ 73

- 19 Features maps in the transformer bottleneck. On the left, we show 16 out of the 64 features maps extracted by the anatomical representations S_t . On the right, we show the features maps predicted by the MLP (top row) when ranging the value of dt from 0 (ED cardiac phase) to 1 (ES cardiac phase). Colour maps linearly range from 0 (dark blue) to 1 (yellow). 74

- 20 Example of segmentation masks generated by an unsupervised conditional GAN trained using images from the ACDC dataset (Section 3.8.1). The generator receives an input image and produces realistic segmentation masks. To prevent posterior collapse, we used a self-supervised consistency loss between the transformed images and their associated predicted segmentation. However, without using more stringent pixel-level constraints, there are no guarantees that the generated masks overlap with the regions of interest. 83
- 21 A reconstruction cost can help unsupervised GANs to generate realistic masks that also overlap with the anatomical regions. Given an input image \mathbf{x} , the generator produces realistic segmentation masks and a residual representation using a *softmax* activation function. A mask discriminator encourages the predicted segmentations to look realistic, while a decoder combines them with the residual representations to reconstruct the input image and obtain $\tilde{\mathbf{x}} \approx \mathbf{x}$ 84
- 22 Training a segmentor $\Sigma(\cdot)$ with paired and unpaired data. When annotations are available, we optimise $\Sigma(\cdot)$ with a supervised cost function. For unlabelled images, we train $\Sigma(\cdot)$ using a data-driven adversarial loss, evaluating if the predicted mask belongs to the manifold of real segmentation masks. 85

- 23 Performance of popular GAN variants in the task of semantic segmentation. Performance is measured in terms of Dice (\uparrow) and IoU (\uparrow) scores, where arrows show the metric improvement direction. The box plots report median and inter-quartile range (IQR) of each metric on a given test set, considering as outliers those values falling outside two times the IQR. We compare the performance of a standard segmentor (UNet) and the following GAN variants: Non-saturating GAN (NSGAN), Relativistic-average GAN (RaGAN), Least-square GAN (LSGAN), Wasserstein GAN with gradient penalty (WGAN-GP). We consider the following medical datasets: ACDC (Section 3.8.1), LVSC (Section 3.8.2), and CHAOS (T1 and T2 images, Section 3.8.4). The top row in the figure shows Dice and IoU scores on the test set after training the segmentor using only 5% of labelled training samples. The bottom row reports the performance when using 25% of annotated training samples. We observe that GANs improve the segmentor training especially when training annotations are scarce. 86
- 24 Comparison of different regularisation techniques: **a)** Instance Noise adds a small random perturbation to the discriminator input; **b)** CoordConv introduces continuous spatial information by adding a spatial coordinate grid to the data; **c)** Texture layer learns to introduce continuous information in the form of a sinusoidal grid. 94
- 25 Examples of textures added by the discriminator on top of the segmentation masks. We show examples on ACDC, CHAOS-T1 and CHAOS-T2 test sets. The textures appear as a small amplitude sinusoidal grid pattern, having different phases and oscillation frequencies for each class (including the background). To easy visualization, all images are cropped around the object of interest. On the right, we report the values of the textural parameters learned by the mask discriminator, in radians. 96

26	Segmentation performance of a vanilla UNet segmentor and the analysed GAN variants when regularised with different techniques. Performance is measured in terms of Dice (\uparrow) and IoU (\uparrow) scores, where arrows show the metric improvement direction. The box plots report median and inter-quartile range (IQR) of each metric on a given test set, considering outliers those values falling outside $1.5 \times \text{IQR}$. Overall, we observe that regularisation leads to larger performance gains on NSGAN, while LSGAN and WGAN-GP have smaller gains.	97
27	In an adversarial game, our model learns to generate segmentation masks that look realistic at multiple scales and overlap with the available scribble annotations. Loopy arrows in the figure, on the segmentor, represent the proposed attention gates, which under adversarial conditioning suppress irrelevant information in the extracted features maps.	104
28	Model architectures. Top: segmentor and discriminator interact at multiple scales. Bottom: convolutional blocks detail. In yellow background, the Adversarial Attention Gate (AAG).	110
29	Adversarial Attention Gates consist of an attention block (yellow background in the figure) pairing Adversarial Deep Supervision (ADS, obtained via the connection in pink background) and a multiplicative gating operation (green background).	112

- 30 Example of generated scribbles for each dataset. In ACDC, scribbles were manually annotated inside the available segmentation masks. In CHAOS and PPSS, we obtained scribbles for each class via binary erosion of the associated segmentation mask, as in (Rajchl et al., 2017). In LVSC, the binary erosion would result in a very good approximation of the myocardium: thus, we generated scribbles with a random walk inside of each class. Please, refer to Section 6.4.1 for additional details. 116
- 31 Example of predicted segmentation masks for the considered methods on each task. Observe that our approach (bottom row) learns spatial relationships in the image, thus preventing the prediction of isolated pixels in the mask, as well as unrealistic spatial relationship among the object parts. 124
- 32 Dice score obtained on the test data by our and methods that don't use shape priors when changing the percentage of available labels in the training set (shaded bands show standard errors instead of deviation for clarity). As upper bound (U.B.) we consider UNet_D^{UB}, trained using all the densely annotated masks. Asterisks denote if differences between first and second best has statistical significance (* $p \leq 0.05$, ** $p \leq 0.01$). 126
- 33 (a) Effect of training with labels from multiple annotators; and (b) performance in presence of mixed supervision (i.e. using masks and scribbles) on ACDC. The upper bound (U.B.) is the UNet_D^{UB}, trained with all the dense segmentation masks. 128

- 34 UNet-like segmentor with (top) vs without (bottom) adversarial conditioning of the attention gates in its decoder. Conditioned by an adversarial shape prior (w/ ADS), the model learns semantic attention maps able to localize the object to segment at multiple scales. Also, the shape prior encourages the segmentor to learn multi-scale relationships in the objects. 130
- 35 Weight distribution for the convolutional layers at depth $d=4$ of the segmentor. We compare how the weight distribution changes during training, with and without the use of ADS on the segmentor. Notice that ADS helps the layer training, and the initially narrow distribution becomes broader in time. 131
- 36 We train a segmentor using only scribble annotations as supervision. To regularise the model to produce realistic predictions, we introduce a self-supervised multi-scale consistency objective. Coupled with a customised attention gate, this objective biases the segmentor toward predicting masks satisfying short-range and long-range dependencies in the image, ultimately improving segmentation performance. 137
- 37 Detail of a decoding block at depth level d . The convolutional block processes the input features and predicts a low-resolution version of the segmentation mask $\mathbf{y}^{(d)}$ as part of a PyAG attention module (represented in light yellow background). To ensure that the mask $\tilde{\mathbf{y}}^{(d)}$ is consistent with the final prediction $\tilde{\mathbf{y}}$, we use the self-supervised multi-scale loss described in Equation 7.1 and graphically represented in Figure 38. Using the predicted mask, we compute the probability of pixels belonging to the background and then suppress their activations in the features map $\mathbf{M}^{(d)}$ according to Equation 7.2 141

38 Self-supervised training of the segmentor. Thanks to PyAG modules, the model produces segmentation masks at multiple scales. We compare (\ominus symbol) the lower resolution masks (green squares) to those obtained under-sampling the full resolution prediction \tilde{y} (blue squares). At each level, we compute a self-supervised loss contribution $\mathcal{L}_{Self}^{(\cdot)}$ that we use as a regulariser. To prevent trivial solutions, we stop (X symbol) gradients (red arrows) from propagating through the highest resolution stream. 142

39 Segmentation performance in terms of Dice (\uparrow) and IoU scores (\uparrow) and Hausdorff distance (\downarrow), with arrows reporting metric improvement direction. The box plots report median and inter-quartile range (IQR) of each metric on a given test set, considering outliers those values falling outside $2 \times \text{IQR}$. **Left column:** our method vs baseline (UNet) and other methods regularising the prediction with a Compactness loss (UNet_{Comp.}), or CRF as post-processing (UNet_{CRF}). Observe how our method shows the best performance across datasets. **Right column:** our method vs methods regularising the predictions using a shape prior learned from unpaired masks (DCGAN, ACCL, and UNet_{AAG}). Our method has competitive performance with the best of the benchmark models, and it has the advantage of not requiring a set of unpaired masks for training. 147

40 Example of segmentation masks predicted by our model on different datasets. In most cases, the model can effectively approximate the true segmentations. 148

- 41 Whenever a test image falls outside the training data distribution, a segmentor may underperform and produce unrealistic predictions. Herein, we suggest 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. 153
- 42 We re-use GAN discriminators to correct segmentation predictions at inference. The key to our success is training stable and re-usable discriminators, as we detail in Section 8.3.2. At inference, we tune a small convolutional block $\Omega(\cdot)$ on each test sample \mathbf{x} , independently, until the predicted mask $\tilde{\mathbf{y}}$ satisfies the adversarial shape prior. We only need a single sample to do the fine-tuning. 157
- 43 We show examples of mistaken predictions and their corrections obtained 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. 164
- 44 Dice (\uparrow), IoU (\uparrow) and Hausdorff distance (\downarrow) obtained **before** and **after** tuning the segmentor on the individual test instances. Arrows show metric improvement directions. Under each violin plot, we also report the median performance, with 95% confidence interval as subscript. Observe how adversarial Test-time Training improves performance under different metrics and datasets (bootstrapped t-test, $*p < 0.05$, $**p < 0.01$). 166
- 45 Adversarial TTT has competitive performance with TTANN, 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 significant (bootstrapped t-test, $*p < 0.05$, $**p < 0.01$). 167

- 46 Failure cases. Since the information contained inside the predicted mask is limited, the discriminator will not penalise realistic but wrong segmentation masks (**top row**). In some cases, it might even encourage the segmentor to make bigger mistakes (**bottom row**). 169
- 47 The proposed approach in a causal setting. We add the adaptor Ω in front of SDNet 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 \mathbf{R} . A decoder uses both of them to reconstruct the adapted image, predicting $\tilde{\mathbf{x}}' \approx \mathbf{x}'$. A mask discriminator learns to say apart real segmentation masks from the predicted ones. **At inference**, we perform Test-time Training by minimising 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 Equation 8.2). 171
- 48 Toward Causal Test-time Training. We compare the performance of: a GAN before and after adversarial Test-time Training; the SDNet model (discussed in Section 8.5.3); 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. 172
- 49 Compatibility with post-processing techniques (PostDAE and CRF). Bar plots report average performance and standard errors. 174

- 50 Example of model failures. In both ACDC and LVSC, the apical and the basal slices of the heart are the hardest to segment, due to intrinsic uncertainty of the cardiac boundaries, resulting in over/under-segmentations in all the models. For CHAOS, we show that all models make mistakes when the organ boundaries have low contrast, though our model preserves realistic outputs. In PPSS, we show that occlusions make the segmentation task harder; for example, if two people overlap, all model will try to segment both people, rather than only one. 190
- 51 At convergence, the discriminator reaches an equilibrium stage where it always predicts the value 0, equidistant from the *true* and the *fake* labels. As a result, losses converge to the equilibrium value 1.0 both for train and validation. 192
- 52 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. . . . 192

List of Tables

- 1 Overview of the datasets used in this thesis. For each dataset, we report if it contains medical data, the object of interest (O.I.), the modality used to acquire the images, how many different subjects the dataset contains, how many classes are annotated (including the background), and the total number of images. Please, refer to Section 3.8 for additional details. 43
- 2 Dice Score average and standard deviation (subscript) for the segmentation of myocardium (MYO), left ventricle (LV) and right ventricle (RV), on ACDC dataset. We compare models at various proportions of training annotations. Results are the average of three-fold cross-validation. Best results in **bold**. 68
- 3 Hausdorff Distance average and standard deviation (subscript) for the segmentation of myocardium (MYO), left ventricle (LV) and right ventricle (RV), on ACDC dataset. We compare models at various proportions of training annotations. Results are the average of three-fold cross-validation. Best results in **bold**. 70

4	Average and standard deviation (subscript) performance for myocardium segmentation, on LVSC dataset. We report Dice Score on the left table, Hausdorff Distance on the right table. We compare models at various proportions of training annotations, reporting the average of three-fold cross-validation. Best results in bold	72
5	Type of prior used by each model.	120
6	Dice average and standard deviation (subscript) obtained from each method on the test set, for medical and vision datasets. Leftmost column indicates if the learning algorithm has been trained with full mask or scribble annotations. The best method is in bold characters, while the second best is underlined; asterisks denote if their difference has statistical significance ($* p \leq 0.05$, $** p \leq 0.01$).	122
7	Our ablations, as the name states, start with our model but remove: #1: Only gating; #2: Only ADS; #3: Both Gating and ADS; #4: Both ADS and the Discriminator; and finally #5: ADS, the Discriminator and Gating.	132
8	Ablation Study. We compare the performance of our method (Ours) after removing: adversarial Test-time Training (ablation #1), the proposed regularisation technique (<i>fake anchors</i> , #2), the smoothness constraints discussed in Section 8.3.2 (ablation #3), and the adaptor (standard GAN, #4). Performance is in terms of average Dice score on ACDC data, with standard deviation as subscript. 168	168
9	Online Continual Learning. We show that our model can continuously learn from a stream of test data, leading to gradually higher segmentation scores on the test samples. Numbers are average performance, with standard deviation as subscript. Best results in bold	175

10	Dice score and Hausdorff distance (HD) for single organs in ACDC. Abbreviations are as follows: RV: right ventricle, MYO: myocardium, LV: left ventricle.	186
11	Dice score and Hausdorff distance (HD) for single organs in LVSC. MYO stands for myocardium.	186
12	Dice score and Hausdorff distance (HD) for single organs in CHAOS-T1. Abbreviations are as follows: L: liver, RK: right kidney, LK: left kidney, S: spleen.	187
13	Dice score and Hausdorff distance (HD) for single organs in CHAOS-T2. Abbreviations are as follows: L: liver, RK: right kidney, LK: left kidney, S: spleen.	187
14	Dice score and Hausdorff distance (HD) of the proposed approach trained on all the available ACDC data, and tested on 50 extra patients using the challenge server. Note that the server does not provide information about the standard deviation, nor a higher precision for the Dice score. Abbreviations are as follows: RV: right ventricle, MYO: myocardium, LV: left ventricle.	188
15	Training our method with scribbles and with mask supervision. We report the Dice average (standard deviation as subscript) obtained on the test data for each dataset.	189

Ringraziamenti

Durante il mio dottorato ho ricevuto l'aiuto e il sostegno di molte persone, che vorrei ringraziare. Voglio ringraziare il Prof. Emiliano Ricciardi che ha creduto in me e mi ha permesso di iniziare la mia avventura in IMT. Ho imparato molto da te e non dimenticherò mai la tua etica e i tuoi consigli. Desidero esprimere la mia speciale gratitudine al Prof. Sotirios Tsafaris, la cui esperienza è stata impagabile nella formulazione dei quesiti scientifici e della metodologia di ricerca. Gran parte del mio progresso fin qui è dovuta a te. Il tuo feedback e la tua dedizione lavorativa sono stati fonte di ispirazione e hanno portato il mio lavoro a un livello superiore. Ringrazio il Dott. Andrea Leo per il suo grande aiuto e per tutti i suoi insegnamenti nel corso di questi anni, che non dimenticherò mai. Voglio ringraziare i miei colleghi dell'Università di Edimburgo: Agis, Alison, Andrei, Greg, Haochuan, Marija, Pedro, Spiros, Tian, Valerio, e Xiao. Abbiamo condiviso tanti momenti straordinari sia in ufficio che fuori dalla vita accademica. Grazie per il vostro supporto nei momenti più difficili e per le tante risate insieme. Voglio ringraziare i miei colleghi dell'Ospedale del Cuore di Massa: Daniele, Gianmarco e Nicola. Il mio viaggio è iniziato con voi e non dimenticherò mai il tempo passato assieme. Grazie per la vostra amicizia e tutto quello che mi avete insegnato. Non posso ringraziare abbastanza la mia famiglia, che mi ha sempre supportato nelle mie scelte, incoraggiandomi ad inseguire i miei sogni. Mi avete reso la persona che sono oggi. Infine, vorrei ringraziare la mia amata Noemi per essere stata al mio fianco in questi anni, attraverso gli alti e bassi del dottorato. Senza di te, niente di tutto questo sarebbe stato possibile.

Acknowledgements

During my doctoral studies, I received a great deal of support and assistance, and I would like to thank those who contributed to this work. I want to thank Prof. Emiliano Ricciardi, who believed in me and allowed me to start my adventure at IMT. I learned a lot from you, and I will never forget your ethics and teachings. I want to express my special gratitude to Prof. Sotirios Tsaftaris, whose expertise was priceless in formulating the research questions and methodology. I must thank you for most of the progress I made so far. Your feedback and your hard-working spirit have been inspiring and brought my work to a higher level. I want to thank Dr Andrea Leo for his significant help and all his teachings throughout these years, which I will never forget. I thank all my colleagues at the University of Edinburgh: Agis, Alison, Andrei, Greg, Haochuan, Marija, Pedro, Spiros, Tian, Valerio, and Xiao. We shared lots of joyful moments both in the office and outside the academic life. Thank you for your support in the toughest moments and for all the laughs together. I want to thank my colleagues at Massa Heart Hospital: Daniele, Gianmarco, and Nicola. My journey started with you, and I will never forget the time spent together. Thank you for your friendship and everything you taught me. I cannot thank enough my family, who always supported me in my choices and encouraged me to follow my dreams. You made me the person I am today. Finally, I would like to thank my beloved Noemi for being on my side in all these years, through the highs and lows of the PhD. Without you, none of this would have been possible.

Vita

- July 23, 1992** Born, Melfi (Potenza), Italy
- 2011-2014** B.Sc. in Biomedical Engineering
Final mark: 106/110
University of Pisa, Pisa, Italy
- 2014-2017** M.Sc. in Biomedical Engineering
Final mark: 110/110 cum laude
University of Pisa, Pisa, Italy
- 2017** Qualification as Information Engineer
University of Pisa, Pisa, Italy
- 2017-Date** Ph.D. in Cognitive, Computational and Social Neuro-
sciences
IMT Institute for Advanced Studies Lucca, Lucca, Italy
- 2019-Date** Visiting Researcher
University of Edinburgh, School of Engineering, Edin-
burgh, UK

Publications

1. **Valvano, Gabriele**, Andrea Leo, and Sotirios A. Tsaftaris (2021d). “Re-using Adversarial Mask Discriminators for Test-time Training under Distribution Shifts”. In: *arXiv preprint arXiv:2108.11926*
2. **Valvano, Gabriele**, Andrea Leo, and Sotirios A Tsaftaris (2021b). “Stop Throwing Away Discriminators! Re-using Adversaries for Test-Time Training”. In: *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Springer, pp. 68–78
3. **Valvano, Gabriele**, Andrea Leo, and Sotirios A Tsaftaris (2021a). “Self-supervised Multi-scale Consistency for Weakly Supervised Segmentation Learning”. In: *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Springer, pp. 14–24
4. **Valvano, Gabriele**, Andrea Leo, and Sotirios A. Tsaftaris (2021c). “Learning to Segment From Scribbles Using Multi-Scale Adversarial Attention Gates”. In: *IEEE Transactions on Medical Imaging*. DOI: [10.1109/TMI.2021.3069634](https://doi.org/10.1109/TMI.2021.3069634)
5. **Valvano, Gabriele**, Andrea Leo, and Sotirios A. Tsaftaris (2021e). “Regularising Disentangled Representations With Anatomical Temporal Consistency”. In: *Under Review at: Biomedical Image Synthesis and Simulations, Elsevier*
6. Liu*, Xiao, Spyridon Thermos*, **Gabriele Valvano***, Agisilaos Chartsias, Alison O’Neil, and Sotirios A. Tsaftaris (2021). “Measuring the Biases and Effectiveness of Content-Style Disentanglement”. In: *Proceedings of the British Machine Vision Conference 2021*. British Machine Vision Association
7. **Valvano, Gabriele**, Agisilaos Chartsias, Andrea Leo, and Sotirios A. Tsaftaris (2019). “Temporal Consistency Objectives Regularize the Learning Of Disentangled Representations”. In: *Domain Adaptation and Representation Transfer (DART)*. Springer, pp. 11–19. ISBN: 978-3-030-33391-1
8. Martini, Nicola, Alessio Vatti, Andrea Ripoli, Sara Salaris, Gianmarco Santini, **Gabriele Valvano**, Maria Filomena Santarelli, Dante Chiappino, and Daniele Della Latta (2019). “Robust Reconstruction of Cardiac T1 Maps Using RNNs”. In: *Medical Imaging with Deep Learning (MIDL)*
9. **Valvano, Gabriele**, Gianmarco Santini, Nicola Martini, Andrea Ripoli, Chiara Iacconi, Dante Chiappino, and Daniele Della Latta (2019). “Convolutional Neural Networks for the Segmentation of Microcalcification in

- Mammography Imaging". In: *Journal of Healthcare Engineering* 2019. DOI: [10.1155/2019/9360941](https://doi.org/10.1155/2019/9360941)
10. **Valvano, Gabriele**, Nicola Martini, Andrea Leo, Gianmarco Santini, Daniele Della Latta, Emiliano Ricciardi, Dante Chiappino, and Pietro Pietrini (2019). "Evaluation of Planar and Volumetric Convolutional Neural Networks for Brain Segmentation". In: *Organization for Human Brain Mapping (OHBM)*
 11. Della Latta, Daniele, Gianmarco Santini, **Gabriele Valvano**, Nicola Martini, Andrea Ripoli, Francesco Avogliero, Alberto Clemente, Carla Luisa Susini, Dante Chiappino, et al. (2018). "Contrast-free Estimation of Cardiac Volumes from CT Scans Using Deep Learning". In: *European Congress of Radiology (ECR)*. DOI: [10.1594/ecr2018/C-1413](https://doi.org/10.1594/ecr2018/C-1413)
 12. **Valvano, Gabriele**, Daniele Della Latta, Nicola Martini, Gianmarco Santini, Andrea Gori, Chiara Iacconi, Andrea Ripoli, Luigi Landini, and Dante Chiappino (2017). "Evaluation of a Deep Convolutional Neural Network Method for the Segmentation of Breast Microcalcifications in Mammography Imaging". In: *European Medical and Biological Engineering Conference & Nordic-Baltic Conference on Biomedical Engineering and Medical Physics (EM-BEC & NBC)*. Springer, pp. 438–441. DOI: [10.1007/978-981-10-5122-7_110](https://doi.org/10.1007/978-981-10-5122-7_110)
 13. Santini, Gianmarco, Daniele Della Latta, Nicola Martini, **Gabriele Valvano**, Andrea Gori, Andrea Ripoli, Carla L. Susini, Luigi Landini, and Dante Chiappino (2017). "An Automatic Deep Learning Approach for Coronary Artery Calcium Segmentation". In: *European Medical and Biological Engineering Conference & Nordic-Baltic Conference on Biomedical Engineering and Medical Physics (EM-BEC & NBC)*. Springer, pp. 374–377. DOI: [10.1007/978-981-10-5122-7_94](https://doi.org/10.1007/978-981-10-5122-7_94)

Abstract

Over the last decades, medical imaging techniques have played a crucial role in healthcare, supporting radiologists and facilitating patient diagnosis. With the advent of faster and higher-quality imaging technologies, the amount of data that is possible to collect for each patient is paving the way toward personalised medicine. As a result, automating simple image analysis operations, such as lesion localisation and quantification, would greatly help clinicians focus energy and attention on tasks best done by human intelligence.

Most recently, Artificial Intelligence (AI) research is accelerating in healthcare, providing tools that often perform on par or even better than humans in conceptually simple image processing operations. In our work, we pay special attention to the problem of automating semantic segmentation, where an image is partitioned into multiple semantically meaningful regions, separating the anatomical components of interest.

Unfortunately, developing effective AI segmentation tools usually needs large quantities of annotated data. Conversely, obtaining large-scale annotated datasets is difficult in medical imaging, as it requires experts and is time-consuming. For this reason, we develop automated methods to reduce the need for collecting high-quality annotated data, both in terms of the number and type of required annotations. We make this possible by constraining the data representation learned by our method to be semantic or by regularising the model predictions to satisfy data-driven spatio-temporal priors. In the thesis, we also open new avenues for future research using AI with limited annotations, which we believe is key to developing robust AI models for medical image analysis.

Chapter 1

Introduction

Artificial Intelligence (AI) in healthcare and radiology is a promising research direction, which has already provided remarkable success in clinical practice (Benjamins, Dhunoo, and Meskó, 2020). A challenging problem of machine learning, particularly in medical imaging, is to have access to large amounts of annotated data. With limited annotations, AI models reduce their ability to generalise on unseen data and face consistent performance drops, which we cannot ignore in clinical applications. For this reason, it is essential to constrain models development using regularisation techniques.

In this thesis, we improve the ability to generalise on unseen data by introducing unsupervised and self-supervised regularisation methods. The proposed approaches improve model robustness and make AI models more reliable for clinicians. More specifically, we present methods for semantic segmentation and demonstrate their utility in several medical applications.

1.1 Medical Motivation

Medical imaging is a significant part of many medical diagnosis and treatment processes. Physicians use medical images for clinical analysis and to plan medical interventions.

Currently, radiologists do most of the medical image analysis in person. However, image interpretation performed by humans exposes the results to subjectivity, fatigue, personal experience, and it is time expensive, drastically limiting the ability of healthcare to advance toward more evidence-based and personalised medicine. Furthermore, with the advent of new technologies, the amount of data that is possible to collect from patients is enormously increasing, resulting in an unprecedented need for automated procedures that lessen the time required for simple but time expensive operations.

In recent years, AI automated many image analysis operations, directly learning clinical tasks from data: from the detection of pathologies to their quantification and characterisation (Zhou, Greenspan, et al., 2021; Petersen, Abdulkareem, and Leiner, 2019a). Unfortunately, AI tools, such as neural networks, are data greedy, which limits their applicability in the medical field. Medical images are heterogeneous, present numerous disease patterns, have sparse and noisy annotations, and data samples are imbalanced and follow multi-modal distributions (Zhou, Greenspan, et al., 2021). Consequently, the use of AI is often challenging in practice, especially for tasks requiring large carefully-annotated datasets, such as image segmentation.

Semantic segmentation is a central task in medical imaging, consisting of partitioning an image into smaller meaningful regions, based on some homogeneity characteristics. Image segmentation is often the first step for extracting quantitative measurements from an image. For example, it allows to measure ejection fraction (Bernard et al., 2018) and the calcium score in cardiac imaging (Santini et al., 2017; Agatston et al., 1990), or it can provide information about tumour size and location (Havaei et al., 2017).

Since the advent of deep learning, semantic segmentation has witnessed significant progress in the design and performance of automated procedures. However, developing automatic tools for medical image segmentation is challenging because large-scale fully-annotated datasets are rare. The lack of labelled data motivates the research of new approaches that overcome the limitations of traditional supervised learn-

ing (Cheplygina, de Bruijne, and Pluim, 2019; Tajbakhsh et al., 2020).

In this thesis, we focus on medical image segmentation learned using limited levels of supervision. We tackle the problem of missing annotations by introducing prior knowledge in AI models. In particular, we focus on spatio-temporal constraints to satisfy when predicting a segmentation mask for unlabelled or weakly labelled images.

1.2 Common Limitations of Medical Datasets

The scarcity of annotations is a common problem for AI in medical image segmentation, as annotating data is time-expensive and requires expert knowledge. As a result, collecting large-scale labelled datasets is usually impossible. To address this limitation in computer vision, traditional solutions include data augmentation, pre-training the models on natural images, and the use of weight regularisation. However, these techniques only partially address this problem. For example, data augmentation suffers from large correlations between the available data and the augmented samples; natural images statistics are usually very different from those of medical imaging; weight regularisation tries to solve the problem only by constraining the model capacity. For this reason, there has been a considerable effort to include additional information into machine learning models. Many approaches improve model performance using unlabelled data for their optimisation (Cheplygina, de Bruijne, and Pluim, 2019). Others learn to model the possible image variations that characterise the data distribution (Zhao, Balakrishnan, et al., 2019). Others try to include data from multiple data sources (e.g., different imaging modalities) (Zhou, Ruan, and Canu, 2019), or to use weaker forms of supervision for training (e.g., scribbles, bounding boxes or image-level annotations) (Tajbakhsh et al., 2020). Finally, there is a large body of the literature (Nosrati and Hamarneh, 2016) focusing on including a priori data constraints into the model, which can encourage the model predictions to be realistic.

In this thesis, we investigate methods using unlabelled data (Chapter 4, 5), weakly annotated data (Chapter 6, 7), and other real-world con-

straints to encourage predictions to be realistic (Chapter [4](#), [7](#), and [8](#)).

1.3 Prior-driven Regularisation

Semantic segmentation is strictly related to the concept of shape. Shapes define regions of interest that satisfy specific properties inside an image. For example, the size of anatomical organs must belong to anatomically plausible ranges, organs usually have smooth outlines, and their position inside the body is approximately known a priori. Including this information into machine learning provides increased robustness on unseen data, and can limit unrealistic predictions even if the models are optimised using limited annotated data.

Recent years have seen an increasing interest in using priors in deep learning. Generally speaking, it is possible to introduce these priors in the form of *training objectives* to optimise (Kervadec, Dolz, Tang, et al., [2019](#); Zhou, Li, et al., [2019](#)), *architectural designs* (Zheng et al., [2015](#); Kohl et al., [2018](#)), or *pixel-level refinement* of the predicted segmentation masks (Krähenbühl and Koltun, [2011](#); Painchaud et al., [2019](#)). Moreover, we can use priors to recover the missing information from sparse labels, such as scribbles (Grady, [2006](#); Valvano, Leo, and Tsaftaris, [2021c](#)).

In Chapter [3](#), we detail each of the above categories of priors, which we extensively use for the rest of the thesis.

1.4 Overview and Technical Contributions

We now give a brief overview of the thesis and its contributions.

The following two chapters present the background needed for our work. To better understand the utility and the practical implications of the developed methods, **Chapter [2](#)** introduces the most commonly used techniques in Medical Imaging, with a particular focus on Magnetic Resonance Imaging (MRI). Here, we describe the physical principles and the hardware used to acquire MR images, which significantly impact the generated image appearance. Then, we discuss the importance of cardiac and abdominal MRI in clinical practice.

We provide the technical background behind the developed methods in **Chapter 3**, where we introduce the fundamental concepts and Machine Learning tools that we used. In the chapter, we define learning algorithms and describe relevant learning paradigms used in this thesis. Then, we motivate the need for regularisation techniques to stabilise model optimisation with limited data. We discuss prior-driven regularisation categories, which we subdivide as acting at features level or prediction level. After that, we offer an overview of generative models and their use to learn high-quality data representations. Finally, the chapter gives an overview of the mathematical notation, the metrics and the datasets used for our experiments.

To limit the need for a large number of annotations, **Chapter 4** presents a novel approach in the context of disentangled representation learning for semantic segmentation. In the chapter, we regularise the learning of high-level data representations based on self-supervised spatio-temporal consistency. We focus on cardiac segmentation of cine MRI images and leverage the intrinsically available temporal information to encourage coherent model predictions between subsequent temporal frames. As a result, the model learns to detect image components that share similar spatio-temporal dynamics (such as the heart), and we increase performance on several medical datasets. The content of this chapter is based on two publications:

- Valvano, Gabriele, Agisilaos Chartsias, Andrea Leo, and Sotirios A. Tsaftaris (2019). "Temporal Consistency Objectives Regularize the Learning Of Disentangled Representations". In: *Domain Adaptation and Representation Transfer (DART)*. Springer, pp. 11–19. ISBN: 978-3-030-33391-1
- Valvano, Gabriele, Andrea Leo, and Sotirios A. Tsaftaris (2021e). "Regularising Disentangled Representations With Anatomical Temporal Consistency". In: *Under Review at: Biomedical Image Synthesis and Simulations, Elsevier*

Disentanglement-based methods are effective in semi-supervised learning. However, they usually require balancing many loss functions

during training, and they also risk limiting the model flexibility significantly. On the other hand, simpler models regularising training by only imposing constraints at the output level are often effective. Among these, methods learning data-driven shape priors, such as Generative Adversarial Networks (GANs), are promising approaches. In **Chapter 5** we analyse the use of GANs in semi-supervised learning. In this chapter, we regularise learning from limited supervision levels by imposing constraints on the model predictions rather than on the learned data representation. As a result, we allow more freedom on the high-level data representation while also encouraging the prediction of realistic segmentation masks. In the chapter, we compare several GAN variants to learn adversarial shape priors for semi-supervised learning regularisation. Finally, we present a novel method to stabilise GAN training when dealing with semantic segmentation maps.

While effective in many situations, GANs have some limitations, too. For example, their standard formulation can only regularise the model at a “global” level without distinguishing between long-range and short-range dependencies in the image. To address this issue, in **Chapter 6** we develop a novel and computationally efficient multi-scale GAN. The proposed method provides a powerful shape prior, able to drive the segmentor learning in weakly supervised settings. In particular, the adversarial framework complements the partial labels with a data-driven loss which recovers the missing label information. We report the state-of-the-art performance on several medical and non-medical datasets, and we also release a new dataset of weak annotations. The content of this chapter is based on the paper:

- Valvano, Gabriele, Andrea Leo, and Sotirios A. Tsaftaris (2021c). “Learning to Segment From Scribbles Using Multi-Scale Adversarial Attention Gates”. In: *IEEE Transactions on Medical Imaging*. DOI: [10.1109/TMI.2021.3069634](https://doi.org/10.1109/TMI.2021.3069634)

Chapter 7 further extends the method developed in **Chapter 6** by introducing multi-scale spatial consistency in the predicted masks using a self-supervised objective. The advantage of this method is to re-

move the need for unpaired segmentation masks during training. We show that the method achieves similar performance to that of multi-scale GANs while being widely applicable and independent of the segmentation masks availability. This chapter is based on:

- Valvano, Gabriele, Andrea Leo, and Sotirios A Tsaftaris (2021a). “Self-supervised Multi-scale Consistency for Weakly Supervised Segmentation Learning”. In: *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Springer, pp. 14–24

Finally, we observe that when the test images fall outside the distribution of the training data, an already optimised segmentor may underperform and produce unrealistic outputs. In these cases, detecting trivial mistakes can be helpful to increase model reliability, which is crucial for medical applications. Toward this goal, **Chapter 8** introduces a novel approach to regularise model predictions at test-time while recycling components previously developed during training. We consider methods such as those of Chapter 5 and Chapter 6, and we demonstrate that it is possible to re-use adversarially learned shape priors at inference, increasing model performance and robustness under distribution shifts. The content of this chapter is based on our publications:

- Valvano, Gabriele, Andrea Leo, and Sotirios A Tsaftaris (2021b). “Stop Throwing Away Discriminators! Re-using Adversaries for Test-Time Training”. In: *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Springer, pp. 68–78
- Valvano, Gabriele, Andrea Leo, and Sotirios A. Tsaftaris (2021d). “Re-using Adversarial Mask Discriminators for Test-time Training under Distribution Shifts”. In: *arXiv preprint arXiv:2108.11926*

To conclude our manuscript, **Chapter 9** draws general considerations on the methods used for this thesis and discusses what we think are promising research directions.

To facilitate replicating our work, we open-source new data and the code developed for our experiments at the following URLs.

- **Chapter 4.** Code: <https://github.com/vios-s/sdtnet>;
- **Chapter 6.** Code and data: <https://vios-s.github.io/multiscale-adversarial-attention-gates>;
- **Chapter 7.** Code: <https://vios-s.github.io/multiscale-pyag>;
- **Chapter 8.** Code: <https://vios-s.github.io/adversarial-test-time-training>.

Chapter 2

Clinical and Medical Imaging Background

In this thesis, we use data from medical imaging modalities, and specifically from Magnetic Resonance Imaging (MRI). Magnetic Resonance is a powerful tool for medical diagnosis, and it allows radiologists to acquire non-invasive images of the patient anatomy while characterising the biological tissues based on their magnetic properties.

Computer vision techniques do not usually consider image acquisition physics. However, to allow for a better understanding of the applications developed in this thesis, we briefly introduce the most common medical imaging techniques in Section [2.1](#). Then, we describe the fundamentals of MRI in Section [2.2](#), with a specific focus on cardiac and abdominal imaging, in Sections [2.3](#) and [2.4](#).

2.1 Medical Imaging

Medical imaging refers to the techniques and processes used to create visual representations of different parts of the human body. In the clinical practice, pathologies diagnosis and treatment often involves one or more imaging techniques, which we can broadly categorise as structural or functional modalities. Structural modalities include MRI, X-rays and

computed tomography (CT), and they have the goal of reproducing the inside of the body without the necessity of invasive surgical procedures. On the contrary, functional modalities aim at representing the functioning of the body parts in terms of movement or metabolism. The latter category of modalities includes ultrasound, PET and SPECT, but also contrast-enhanced CT and MRI, fMRI and perfusion MRI.

All of these techniques belong to Radiological Imaging, which we can further categorise as using ionising or non-ionising radiations. In particular, non-ionising radiations expose patients to a reduced radiological risk when undergoing a clinical exam, such as MRI or ultrasound. Instead, ionising radiations are associated with a higher radiological risk, but they often allow to measure information that is complementary to that of non-ionising techniques.

2.2 Magnetic Resonance Imaging

In this thesis, we give a particular focus on MRI images because magnetic resonance does not expose patients to a radiological risk, and it uses a flexible image acquisition procedure. As a result, MRI has become part of many clinical diagnosis processes today, and automating its analysis and interpretation would greatly help clinicians.

To give more context to the technique and better understand the following chapters, we now provide an overview of MRI physical principles (Section [2.2.1](#)) and the effect of the acquisition process on the generated images (Section [2.2.2](#)).

2.2.1 Physical Principles

The specific property that allows biological tissues to interact with a magnetic field is the *spin angular momentum* of their particles. If the nuclei of biological tissues have net spin values, they have magnetic properties that a scanner for Nuclear Magnetic Resonance can measure. For example, it is possible to acquire MR signals using ^1H , ^{13}C and ^{23}Na nuclei. In the clinical practice, MRI is typically performed on ^1H nuclei, because

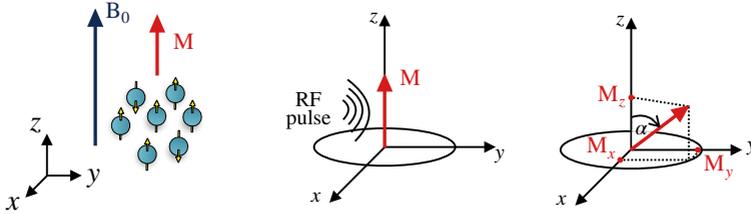


Figure 1: The effect of a radiofrequency pulse (RF) at the Larmor frequency on a set of nuclei aligned with the magnetic field B_0 . The orthogonal RF wave tilts the nuclei spins, producing components of the net magnetisation on the xy plane. The angle α between the z axis and the tilted net magnetisation vector M has the name of flip angle.

they are contained in the water molecules and are the most abundant in the human body. In the following, we briefly describe the physical principles of MRI using the standard nomenclature and formalism.

Inside a magnetic field B_0 , a set of particles with a given spin aligns with its direction and provides a net magnetisation M to the matter. It is possible to manipulate the produced magnetisation in many possible ways to generate MR images highlighting different magnetic properties. In particular, it is possible to modify the equilibrium polarisation of M with an electromagnetic wave having a proper frequency:

$$f_L = \frac{\gamma}{2\pi} B_0.$$

The frequency f_L is called Larmor frequency, and γ is the gyromagnetic ratio: a property of the element we want to excite (e.g., for ^1H , $\gamma \approx 267.51\text{MHz/T}$). Depending on the strength of B_0 and on the nucleus under examination, f_L can have different values, usually in the range of radio frequencies. For example, on a 3T MR scanner for ^1H imaging $f_L \approx 127.7\text{MHz/T}$.

Let us consider a magnetic field B_0 which is parallel to the physical z axis of the MR scanner. A radiofrequency (RF) pulse oscillating to the Larmor frequency will tilt the net magnetisation M toward the orthogonal xy plane, forming a *flip angle* α between M and the z axis (see Figure [1](#)). As a consequence, M will have non-zero components along

the xyz axis, which we can describe in terms of α as:

$$\begin{cases} M_z = M_0 \cos \alpha \\ M_{xy} = M_0 \sin \alpha \end{cases},$$

where M_0 corresponds to the initial magnetisation, parallel to B_0 . The flip angle value depends on the strength and duration of the transmitted RF pulse, and it can vary from 1° to 180° . Once terminated the radiofrequency pulse, the nuclei go back to their initial equilibrium state, releasing the acquired energy in the form of an electromagnetic wave that the MR scanner can measure. In particular, the released wave depends on the interactions that multiple spins have with each other, which cause a dephasing of the magnetisation and make M_{xy} decay in time:

$$M_{xy}(t) = M_{xy}(0) e^{-\frac{t}{T_2}}.$$

The decay is exponential, and we can characterise it with the time constant T_2 , which defines the *spin-spin relaxation time*. During the decay of M_{xy} , also the longitudinal magnetisation M_z returns to its equilibrium value, following the exponential dynamics:

$$M_z(t) = M_z(0) \left(1 - e^{-\frac{t}{T_1}}\right),$$

characterised by the time constant T_1 . The T_1 constant depends on the interactions that the spins have with each other, and it is called *spin-lattice relaxation time*. Together with T_2 , it allows us to define the temporal evolution of the net magnetisation using the Bloch equations:

$$\frac{\partial \mathbf{M}}{\partial t} = \mathbf{M} \times \gamma \mathbf{B} - \left[\frac{M_x \hat{x} + M_y \hat{y}}{T_2} + \frac{(M_z - M_0) \hat{z}}{T_1} \right],$$

where $\mathbf{M} = [M_x, M_y, M_z]^T$ and $\mathbf{B} = [B_x, B_y, B_z]^T$ are the magnetisation and the magnetic field vector, \times denotes a vector product, and \hat{x} , \hat{y} and \hat{z} are the versors along the x , y and z axis, respectively.

The magnetisation vector \mathbf{M} has a precession movement around the z axis at the Larmor frequency f_L and it creates a time-varying magnetic flux which, according to Faraday's law, we can measure with a receiver RF coil to produce an image.

A graphical summary of the generation process of the MR signal can be seen in Figure 2 and Figure 3.

2.2.2 Hardware and Signal Acquisition

In the rest of the thesis, we use several medical datasets whose images were generated using different magnetic field strengths or acquisition sequences. Understanding what this means in terms of image appearance is important for the rest of the thesis, and especially in Chapter 8 where we emphasise that different acquisition protocols may affect model performance at inference. To gain familiarity with these concepts, we now briefly discuss how the MRI signal is acquired and list a few acquisition protocols used in clinical practice.

Acquiring MR images requires dedicated hardware to manipulate the magnetisation vector and measure the induced signals. In the first place, MR scanners need a source of the static and homogeneous magnetic field B_0 to correctly polarise the spins. MR scanners generate B_0 using a magnet which, depending on the field strength ranging from 0.3 T to 7 T, can be a permanent, an electromagnetic or a superconducting magnet.

Once aligned the spins with B_0 , an RF coil, tuned to the Larmor frequency of the nucleus of interest, excites them with a transverse electromagnetic pulse B_1 .

However, to acquire spatially-localised signals and produce a meaningful image, it is necessary to excite only specific locations inside B_0 . For example, assuming a two-dimensional axial acquisition, we must be able to select specific slices in the xy plane. It is possible to acquire the desired 2D slice by using dedicated coils to introduce a magnetic field gradient G_z during the acquisition. Parallel to B_0 , a linear gradient G_z leads to lightly different Larmor frequencies across the 2D slices and makes it possible to excite only the subset of spins in the z coordinate of interest:

$$f_L(z) = \frac{\gamma}{2\pi} (B_0 + zG_z).$$

Within the desired 2D slice, it is possible to introduce an additional encoding on the xy plane, obtained by including the gradients G_x and

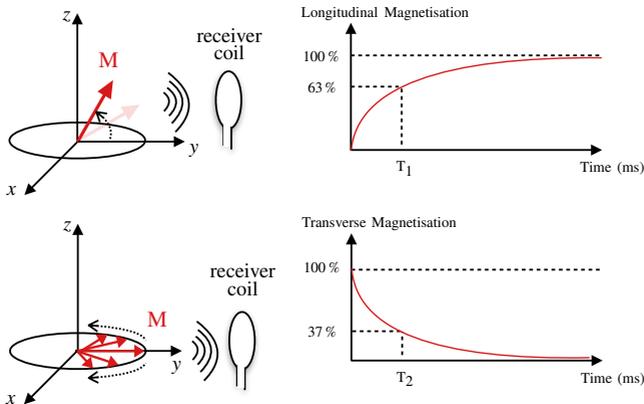
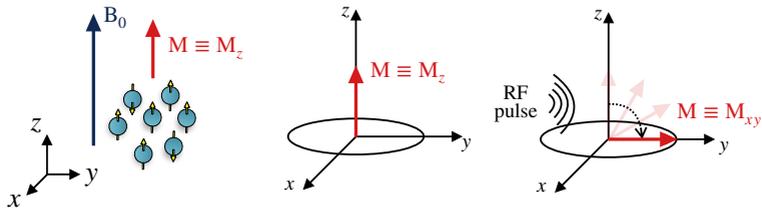


Figure 2: Summary of the physical principle behind the generation of MR signals. Inside the static magnetic field B_0 , the spins align parallel or anti-parallel to the direction of B_0 , with a net magnetisation $M > 0$. Excited by a radiofrequency (RF) pulse, the net magnetisation tilts on the xy plane. Once terminated the RF pulse, spins gradually re-align themselves with the z axis, releasing energy in the form of an RF wave that a receiver coil can capture. At the same time, spins also accumulate a different phase in their precession movement, further decaying the net transverse magnetisation.

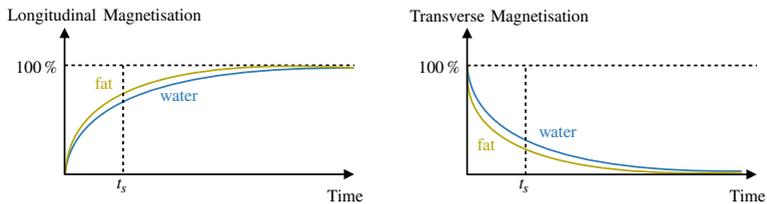


Figure 3: Different tissues exhibit different magnetic properties. Acquiring a signal at time t_s will therefore result in a different signal amplitude measured by the receiver RF coil.

G_y at the end of the excitation. With the activated gradients the spins accumulate a phase that depends by their position $r = [x^*, y^*]$ on the xy plane, and by the excitation time:

$$\phi(t, r) = \gamma \int_0^t G(\tau) \cdot r \, d\tau,$$

with $G = [G_x, G_y]^T$ and \cdot denoting a scalar product. Denoting:

$$k(t) = \frac{\gamma}{2\pi} \int_0^t G(\tau) \, d\tau,$$

we can write the equation of the acquired signal as:

$$s(t) = \int M_{xy}(r) e^{-i2\pi k(t) \cdot r} \, dr.$$

The latter equation states that the MRI signal is the Fourier transform of the transverse magnetisation at a given spatial location $k(t)$. In other terms, the acquired signal corresponds to a profile in the so-called *k-space*. By repeating multiple pulse sequences with different G_y (which allows collecting a signal from different locations), it is possible to fill the k -space matrix. At the end of the procedure, we can produce the final MR image using the inverse Fourier transform to go back from the frequency domain (i.e. the k -space) to the spatial domain.

It is possible to use several different gradient waveforms and radio-frequency pulses, allowing to generate diverse image contrasts. Examples of widely adopted pulse sequences are the T1, T2 and the proton-density (PD) weighted sequences, as well as the gradient echo and spin echo, the inversion recovery, and the diffusion-weighted sequences. These techniques can be used to acquire several anatomical parts, including the cardiac structures and the abdomen. In the following, we briefly discuss cardiac and abdominal MRI and their clinical importance.

2.3 Cardiac MRI

Cardiac MRI allows investigating both the functional and anatomical properties of the heart. In the clinical practice, physicians use cardiac

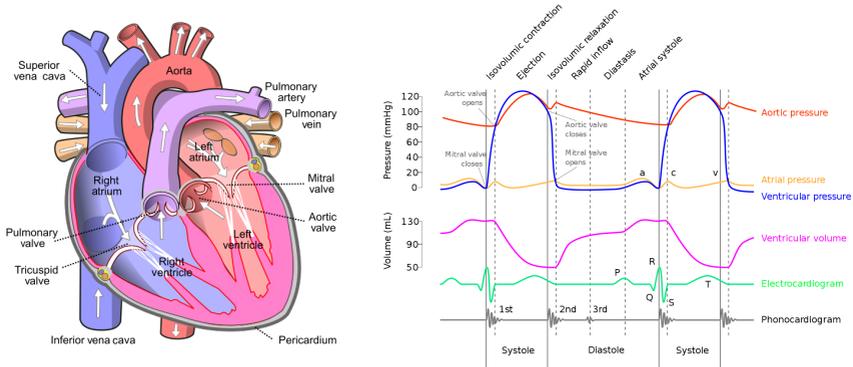


Figure 4: Left: substructures of the human heart. Right: the Wiggers diagram shows the temporal association between the electrical and the mechanical signals that it is possible to measure during the cardiac cycle. Figures are adapted from Wikipedia, [2020a](#) and Wikipedia, [2020b](#) respectively.

MRI for several quantitative measures, including left and right ventricular volumes quantification, ventricular wall thickness, diameters of the great vessels, myocardial infarction size, blood flow measurements.

Cardiac MRI is considered the “gold” standard technique for the non-invasive characterisation of the cardiac function, and it proved to be an effective tool for the diagnosis of complex cardiomyopathies (Petersen, Abdulkareem, and Leiner, [2019b](#)). However, it typically needs clinicians to manually inspect and take measurements on the images, which is time demanding and suffers from human inter-subject and intra-subject variability. Developing fast and reliable techniques for the automated identification of the cardiac structures in an MRI would allow physicians to focus their energies on tasks that cannot be automated and that are best done by human intelligence.

In the remaining chapters of the thesis, we will present automated techniques for cardiac segmentation in MRI. Below, we give a brief description of cardiac anatomy.

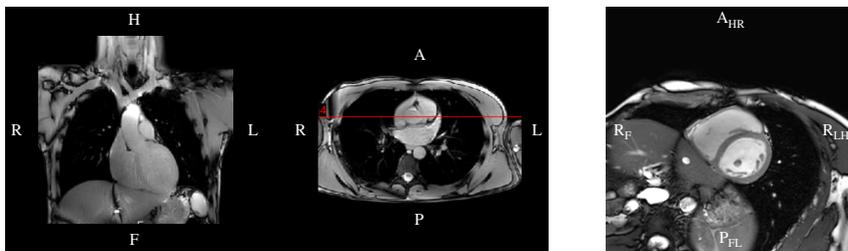


Figure 5: Example of MR images of a 23-year old subject¹. On the left, MRI on the subject thorax in two orthogonal planes. On the right, a short-axis cardiac MRI.

2.3.1 The Heart

The heart is a muscular organ responsible for circulating the blood inside the body, and it is at the centre of the cardiovascular system. At the beginning of a cardiac cycle, the heart receives non-oxygenated blood to the right atrium, which is gathered in the right ventricle and then pumped to the lungs through the pulmonary artery. In the lungs, the blood is oxygenated, and through the blood pressure, it is sent back to the heart again, in the left atrium. From here, the cardiac contraction pumps the oxygenated blood to the left ventricle first, then to the rest of the body through the aorta.

The cardiac contraction process is named *systole*, and it is responsible for “pushing” the blood forward in the cardiovascular system. The dilation process is instead the *diastole*, and it allows to collect blood from the periphery of the body. Cycling between systolic and diastolic phases is made possible by electrical signals that travel from the right atrium to the ventricles and generate the cardiac contraction. We summarise the cardiac structures and the associations between the electrical and mechanical properties of the cardiac cycle in Figure 4. We report an example of short-axis cardiac MRI in Figure 5.

¹This is a younger version of myself while volunteering for an MRI experiment.

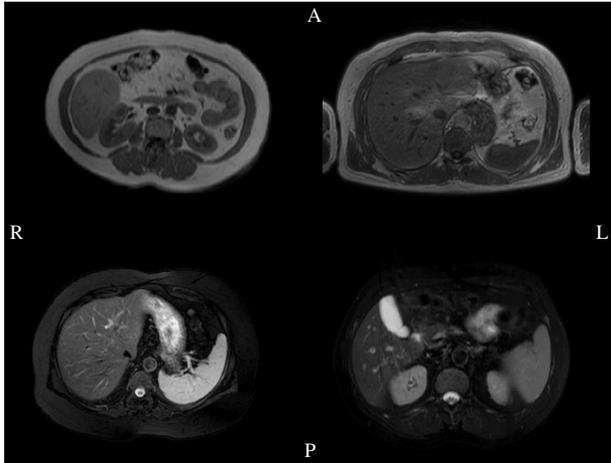


Figure 6: Example of abdominal MRIs of 4 different subjects, contained in the CHAOS dataset (Kavur, Gezer, Barış, Aslan, et al., 2021). In the top row, we show examples of T1-DUAL in-phase images, while in the bottom row we report T2-SPIR images.

2.4 Abdominal MRI

Abdominal imaging has an important role in diagnostic radiology. It can detect emergencies that require immediate treatment or intervention, such as aortic aneurysms and acute liver failures. In the clinical practice, clinicians use CT, MRI and ultrasonography to acquire images of abdominal organs and structures, including bladder, kidneys, liver, prostate and pancreas. Semantic segmentation of the abdominal organs plays a significant role in many clinical applications, such as pre-surgical planning and shape monitoring for several diseases (Kavur, Selver, et al., 2019). Since contrast-enhanced MRI with gadolinium agent is considerably safer than contrast-enhanced CT (Neto et al., 2008), MRI is often the preferred modality to examine patients with severe allergies or chronic renal failure. Moreover, as discussed by Neto et al., 2008, MRI has a higher soft-tissue contrast resolution, and a greater sensitivity to intravenous contrast means (such as gadolinium) than does CT (using iodine-based means).

For these reasons, the semantic segmentation of abdominal organs in MRI is an important challenge for medical image analysis, and we will investigate it in the course of this thesis.

We report an example of abdominal MRI in Figure [6](#).

2.5 Summary

This chapter discussed the medical imaging background of the thesis. We gave an overview of the most common medical imaging techniques, with a particular focus on MRI. Furthermore, we briefly discussed why automated procedures for cardiac and abdominal MRI segmentations are becoming essential for clinical practice.

In the next chapter, we present the technical background related to the Machine Learning tools used for this thesis, and we give an overview of the notation, metrics and datasets used for our experiments.

Chapter 3

Technical Background

In this chapter, we present the technical background of the thesis. We first introduce the mathematical notation used in the remainder of the thesis. Then, we give an overview of learning algorithms and briefly describe learning with different supervision levels. After that, we argue that obtaining well-performing AI models with limited annotations needs the extraction of good data representations. In Section 3.3 we discuss how we can directly or indirectly constrain the model to learn good representations, using prior knowledge about the data, or about the representation itself. Afterwards, we discuss recent literature stating that we can learn good representations using generative models, such as Variational Autoencoders and Generative Adversarial Networks, described in Section 3.6. Lastly, we present the most important metrics to measure segmentation performance, and the datasets used for our experiments in Section 3.7 and 3.8.

3.1 Mathematical Notation

We now describe the mathematical notation used in the following sections and chapters. To ease the reader, we will periodically recall the used notation when defining the mathematical objects. However, this section provides a helpful overview and summary.

We denote sets of data points as Ω_k , where the subscript k defines the type of set (e.g., Ω_k are sets of points of type k). We assume that these data points are sampled from a probability distribution $p(\cdot)$, and we write $\sim p(\cdot)$ to highlight the sampling process.

We use italic lowercase letters to denote scalars s and underlined italic lowercase letters for vectors \underline{v} . Two-dimensional images (matrices) are denoted with bold lowercase letters, as $\mathbf{x} \in \mathbb{R}^{n \times m}$, where $n, m \in \mathbb{N}$ are scalars denoting the matrix dimensions. We refer to tensors $\mathbb{T} \in \mathbb{R}^{r \times s \times t}$ using uppercase letters, where $r, s, t \in \mathbb{N}$.

Lastly, we generally denote functions $\Phi(\cdot)$ using Greek capital letters; however, we sometimes emphasise *loss* functions using calligraphic fonts, as $\mathcal{L}(\cdot)$.

3.2 Learning Algorithms

Machine Learning algorithms are characterized by the use of data to learn a mapping function between input and output data distributions. More formally, given two sets of data points $\Omega_x \equiv \{\underline{x}_i\}_{i=1}^N$ and $\Omega_y \equiv \{\underline{y}_i\}_{i=1}^M$ sampled from an input and an output data distribution, $\underline{x}_i \sim p(\underline{x})$, and $\underline{y}_i \sim p(\underline{y})$, respectively, a Machine Learning algorithm finds a function $\Phi : p(\underline{x}) \rightarrow p(\underline{y})$.

An example of an algorithm used to learn Φ is the Artificial Neural Network (ANN), which is a universal function approximator (Leshno et al., 1993). Inspired by the human visual cortex, feedforward ANNs consist of a series of layers containing hidden units (neurons) which, connected, process the input data to extract gradually higher-level representation, and finally predict an output signal. When the number of stacked layers is more than two, the ANN is named Deep Neural Network (DNN) and the learning process is termed Deep Learning.

Learning a function $\Phi : p(\underline{x}) \rightarrow p(\underline{y})$ defines discriminative models, which map input samples \underline{x} to output data \underline{y} . These models can be learned in a supervised, semi/weakly-supervised, and unsupervised manner, depending on the availability of input-output pairs in the dataset, and the type of pairing. Specifically, when the learning

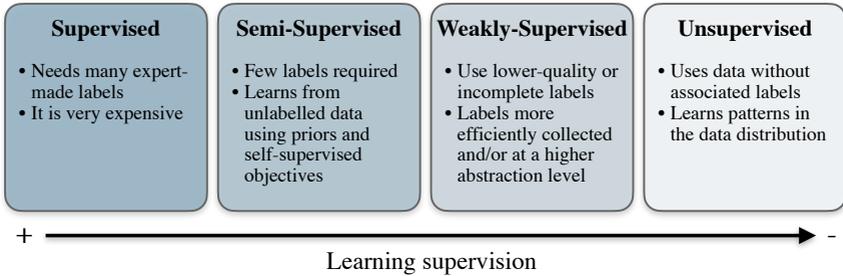


Figure 7: Comparison of different learning scenarios. Classical supervised learning requires the highest amount of supervision, while the need for fine-grained annotations diminishes moving toward unsupervised learning.

happens pairing each \underline{x}_i to a unique output sample \underline{y}_i (that is, given a dataset $\Omega, \forall \underline{x}_i \in \Omega : \exists! \underline{y}_i \in \Omega \text{ s.t. } (\underline{x}_i, \underline{y}_i)$ is an input-output pair), the learning is named supervised. On the contrary, unsupervised learning is the process of learning Φ having access to \underline{x}_i or \underline{y}_i only, but no pairs. Semi-supervised approaches are instead mixed settings, where there is the availability of some paired and some unpaired data. Finally, weakly-supervised learning is in between semi-supervised and unsupervised learning; weak annotations are partial labels or image-level labels that can only give loose supervision to the model; in this case, instead of directly pairing $(\underline{x}_i, \underline{y}_i)$, we pair \underline{x}_i to \underline{y}_i through a weaker semantic connection $\tilde{\underline{y}}_i$.

Recent years have seen a groundbreaking success of *supervised* Deep Learning, which reported state-of-the-art performance in many computer vision tasks. Unfortunately, access to paired data is not always possible, as they do require experts to obtain annotations. Instead, large amounts of unlabeled or weakly annotated data can be considerably easier to collect, motivating the research of better semi-supervised and weakly-supervised techniques to substitute the supervised approaches. Below, we briefly define these learning paradigms, and we graphically represent them in Figure 7.

3.2.1 Learning Paradigms with Limited Supervision

In this thesis, we consider several different learning problems. In all cases, our goal is to limit model dependence on the number of fine-grained annotations. We consider two possible strategies to achieve this goal: reduce the number of annotated data (Semi-supervised Learning) or reduce the time and quality of annotations (Weakly-supervised Learning). We use the first type of approaches in Chapter 4, 5 and 8. Instead, we use Weakly-supervised Learning methods in Chapter 6 and 7. Finally, to stabilise training and obtain better results in a lack of high-quality labels, we present Self-supervised Learning strategies in Chapter 4 and 7.

Below, we give a brief overview of each of these learning paradigms.

Semi-supervised Learning

In this scenario, we assume there are two sets of data: labelled samples Ω_L and unlabeled samples Ω_U . The goal of semi-supervised learning (SSL) is to use Ω_U to improve the model performance compared to training using only Ω_L . Most commonly, the optimisation of DNNs happens as a multi-task learning process. Usually, the unlabeled data are part of the optimisation of self-supervised objectives, such as a reconstruction cost, or linear regression. An alternative to self-supervised objectives is the introduction of prior knowledge into the model, in the form of a loss function. For example, it is possible to penalise the model when it predicts labels of rare classes, or when it predicts unrealistic outputs.

The basic assumption behind the formulation of SSL as a multi-task learning problem is that the unsupervised costs will also improve the performance on the main (supervised) task. As suggested by Chapelle, Scholkopf, and Zien, [2009], such an assumption is reasonable only if we assume that the model predicts outputs based on data features representations satisfying: i) *manifold assumption*: the high dimensional data lie on a low-dimensional manifold; ii) *smoothness assumption*: if two data points are close on this manifold, then the model should output similar predictions; and iii) *clustering assumption*: data points falling inside the same cluster have a high probability of belonging to the same class. Under

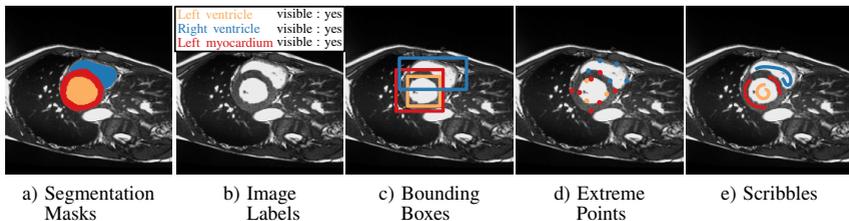


Figure 8: Comparison of different types of annotations. Segmentation masks (a) are the most time expensive annotations to collect. Weaker forms of annotations (b–e) are cheaper to obtain but provide less supervisory signal. The MRI image is taken from the ACDC dataset (Bernard et al., 2018).

these assumptions, quite a lot can be done with limited labels, or even using just one annotated data (Chaitanya, Karani, et al., 2019; Zhao, Balakrishnan, et al., 2019). On the contrary, if the additional assumptions do not hold, the model could perform even worse than in simple supervised learning (Fabio Cozman, 2006).

Weakly-supervised Learning

An alternative approach to reduce the effort associated with label collection is to rely less on fine-grained annotations (Zhou, 2018; Tajbakhsh et al., 2020). Weak annotations are lower-quality labels more easily collected or at a higher abstraction level. They include image-level labels (e.g., the statement “there is an object of the class *dog* in the picture”), bounding boxes (squares containing the object of interest inside), extreme points (sparse points in correspondence of the object boundaries), and scribbles (curvy lines inside the object of interest, which can be interpreted as partial annotations). We report examples of these types of annotations in Figure 8.

Obtaining weak annotation is generally faster than collecting fine-grained ones. For example, bounding boxes can be collected up to $15\times$ faster than segmentation masks (Lin, Maire, et al., 2014), resulting in an increased number of labelled data per annotation time. However, training with reduced supervision is more challenging because the training

signals can be noisy, and the model optimisation can fail. For this reason, weakly supervised approaches require other forms of prior knowledge to limit the flexibility of the model and learn proper data representations.

Self-supervised Learning

Self-supervised learning is an unsupervised paradigm where we train a model without human annotations. In self-supervised learning, the data provide the supervision needed to optimise the model through a proxy loss function. Ideally, to solve the proxy objective effectively, the model will learn the task we are interested in.

The literature reports several pretext tasks for self-supervised learning, including in-painting and out-painting of the images (Zhou, Sodha, et al., 2019), context restoration (Chen, Bentley, et al., 2019), superpixel segmentation (Ouyang et al., 2020), learning image correspondences (Wang, Jabri, and Efros, 2019; Vondrick et al., 2018), coordinate prediction (Bai, Chen, et al., 2019), and contrastive learning (Chaitanya, Erdil, et al., 2020).

Designing a good pretext task can be difficult. Thus, pre-trained self-supervised models are often fine-tuned on the supervised objective of interest (also known as *transfer learning*).

3.3 Direct and Indirect Regularisation of the Data Representation

As discussed in the previous sections, a challenging problem of Machine Learning is reducing the dependence from annotated data. Since annotations are often scarce, supervised learning is not always possible and preventing the model from overfitting can be hard.

Bengio, Courville, and Vincent, 2013 argued that the success of learning algorithms heavily depends on their ability to learn good data representations. In other terms, we should build algorithms that extract all the useful information from the data while removing redundancy and

nuisance factors. To make this possible, one can restrict the learned representations to satisfy specific learning constraints.

In general, we can improve the representations adopting regularisation techniques, which we can categorise as *direct* or *indirect*. Direct approaches directly constrain the learned features. For example, using sparsity or amplitude constraints in the cost function (e.g., ℓ_1 and ℓ_2 regularisations), penalise inter-variable dependencies (Kingma and Welling, 2014), or restricting the representation to depend only by specific data variations (i.e. *features disentanglement*¹). On the other hand, indirect approaches encourage model predictions to satisfy some prior knowledge about the data. As a result, the model autonomously adapts the representation to accomplish the task at hand, and the regularisation assumes an indirect form.

In this thesis, we improve model generalisation using both direct and indirect approaches. In particular, after briefly introducing disentangled representations in Section 3.4 we demonstrate that it is possible to improve such representations using spatio-temporal priors (Chapter 4). Moreover, after introducing popular forms of indirect regularisation in Section 3.5 we present specific applications in Chapter 5 6 7 and 8 demonstrating their benefits in semi-supervised and weakly supervised learning.

3.4 Priors for Direct Regularisation

We now give an overview of priors used for the direct regularisation of the data representation. We first discuss classical approaches, then we review the concept of features disentanglement, which we will extensively use in Chapter 4

¹For an overview of disentanglement in computer vision tasks, when it helps and possible metrics to measure it, interested readers may refer to Liu*, Thermos*, et al., 2021

3.4.1 Classical Approaches

Two classical approaches used to regularise representations are the regularisations ℓ_1 (also known as Lasso) and ℓ_2 (also known as ridge or Tikhonov's regularisation). The ℓ_1 regularisation consists of penalising the ℓ_1 -norm of the features maps extracted by the model from the input data. This technique aims to obtain sparse features representations, introducing invariance and an information bottleneck penalising redundant or stochastic signals.

Similar to the ℓ_1 , the ℓ_2 regularisation penalises the ℓ_2 -norm of the features maps. Intuitively, penalising the ℓ_2 -norm leads to a reduced variance in the extracted features maps, and it is less aggressive than Tikhonov's regularisation because it does not encourage sparsity (thus, more activations can coexist).

There are also other advanced regularisation techniques, such as constraining part of the representation to activate only in correspondence of specific properties of the image. Formally, given a features vector $\underline{v} = \Phi(\mathbf{x})$ extracted by Φ from an image \mathbf{x} , we can decompose it in two subvectors $\underline{v} = \{\underline{v}_1, \underline{v}_2\}$ that respond only to specific transformations of \mathbf{x} . In other terms, given the transformations T_1 and T_2 :

$$\begin{aligned}\underline{v} &= \{\underline{v}_1, \underline{v}_2\} \\ \underline{v}' &= \{\underline{v}'_1, \underline{v}_2\} = \Phi(T_1 \circ \mathbf{x}) \\ \underline{v}'' &= \{\underline{v}_1, \underline{v}'_2\} = \Phi(T_2 \circ \mathbf{x}),\end{aligned}$$

where $\underline{v} \neq \underline{v}' \neq \underline{v}''$ and $T_1 \neq T_2$. If T_1 and T_2 change independent aspects of the data, such as the colour and position of an object, this form of regularisation generates the so-called *disentangled representations*, which we introduce below.

3.4.2 Disentangled Representations

Recent literature (Bengio, Courville, and Vincent, 2013; Bengio, 2009) discusses that we should consider data samples as generated from independent generating factors, or factors of variation. As a result, we should

train models to learn representations that separate out data generating factors into separate subsets of features: a process having the name of *features disentanglement*.

Disentangled representations divide data explanatory factors into disjoint subsets. Formally, we can define disentangled representations starting from the concept of “symmetry transformations”, i.e. those transformations that can change specific aspects of the real world state, while keeping other aspects unchanged, or invariant. According to Higgins, Amos, et al., [2018], a vector representation is disentangled if it can be decomposed into a number of subspaces, each one of which is compatible with, and can be transformed independently by a unique symmetry transformation. Vice versa, changes happening in the encoded features are sparse over real-world transformations.

Disentangling the generating factors of the data would be of great importance for increasing interpretability of the extracted features and improving generalization on unseen data, thanks to the concept of equivariance (Hinton et al., [2012]). Moreover, the isolation of factors of variations allows interpretable latent code manipulation, which is desirable in several applications, from Image-to-Image translation to video editing.

A shared definition of disentanglement is still open to debate. However, many researchers believe that a factorial representation, i.e. a representation with statistically independent variables, could be a good starting point for disentanglement (Kim and Mnih, [2018]; Watanabe, [1960]). This representation is a compact and meaningful information encoding, can improve model generalisation (Bengio, Courville, and Vincent, [2013]), and is more robust against adversarial attacks (Alemi et al., [2016]).

Decoupling Image Shape and Appearance

In computer vision, researchers have extensively used disentanglement to decouple information about image shape and position (usually named *content*) from that regarding image appearance (often called *style*). Outside Image-to-Image translation (Liu, Breuel, and Kautz, [2017]; Lee, Tseng, et al., [2018]; Huang et al., [2018]), content-style disentanglement has been used in other applications, such as pose estimation (Charles et

al., [2013] and semantic segmentation (Chartsias, Joyce, et al., [2019]; Chen, Ouyang, et al., [2019]; Yang et al., [2019]). In the context of medical image segmentation, a shared consensus is that being able to decouple the stylistic information (i.e. the imaging modality) from the image content (i.e. the patient anatomy) would allow training modality-independent segmentation methods. In fact, object segmentation is a content-specific task which, thanks to disentanglement, can be improved using data from multiple sources, such as scanners or imaging modalities, sharing similar content-related information. In this scenario, learning good content representations is crucial, as the extraction of object shape and position directly affects the downstream segmentation task.

In Chapter 4, we show that it is possible to use a spatio-temporal prior to constraint the content representation to satisfy the physical properties of the real world, such as the smooth temporal variations of the patients' anatomy. By regularising the content features, we show to also improve the downstream segmentation task.

3.5 Priors for Indirect Regularisation

It is possible to regularise data representations indirectly as well, by introducing additional constraints to the model's predictions (Nosrati and Hamarneh, [2016]). For example, in image segmentation, the optimised models must be robust to the presence of noise, perform well with low contrasted images, and take into account high object variability. However, when optimised with limited annotations, models can be hard to train and are likely to overfit the training distribution. In these cases, a common strategy for obtaining good data representations is to regularise the optimisation using prior knowledge about the expected results. For example, traditional methods require the presence of homogeneity inside the segmented object (Nosrati and Hamarneh, [2016]). Unfortunately, homogeneity is not necessarily present in an image, and magnetic field biases or other acquisition artifacts can alter the appearance of anatomy within the same image.

A better description of real-world objects should consider a combina-

tion of the notions we have about the entities to segment. For this reason, there has been a considerable effort in designing better prior-driven training objectives or to effectively post-process the model predictions. We can introduce prior information in several forms: via user interaction, object appearance, topology, location, size, shape, and relative position to other regions of the image. User interaction requires an expert to correct the model predictions, or to include a seed area to make the model work. For example, the random walker algorithm proposed by Grady, 2006 requires a user to draw scribbles inside the image. On the contrary, we can directly include the other forms of priors in the automated procedure: as *learning objectives*, *design biases*, or *data biases*. As we will largely use these priors, we describe each of them below.

3.5.1 Priors as Learning Objectives

Learning a model for image segmentation is the problem of finding a function Φ to map an image \mathbf{x} to a label map \mathbf{y} where each label corresponds to an independent and semantically meaningful region of the image. Ideally, we can solve this problem by minimising a training objective containing two terms: \mathcal{L} that evaluates the correctness of the mapping on a single annotated data sample; and a second term \mathcal{R} , which regularises the model to learn a general concept of the object. In other terms, given a model Φ_θ parametrised by θ , and given the model prediction on the input image $\tilde{\mathbf{y}} = \Phi_\theta(\mathbf{x})$, we define the optimisation procedure as: $\min_\theta \mathcal{L}(\mathbf{y}, \tilde{\mathbf{y}}) + \mathcal{R}(\tilde{\mathbf{y}})$.

In particular, we use \mathcal{R} to include prior knowledge about the objects. For example, if we know that the object appearance falls within a range of intensities (e.g., the calcium has a specific intensity range, in CT images), we can use a dissimilarity metric to penalise those predictions $\tilde{\mathbf{y}}$ that include voxels out of this range. We can measure the dissimilarity using different formulations. For instance, we can assume a probabilistic prior about the average value of the object intensity, and then compute

the probability of the image pixels to belong to the k -th object as:

$$p = \frac{1}{s_k \sqrt{2\pi}} e^{-\frac{(z-m_k)^2}{2s_k^2}},$$

where m_k and s_k are the mean and standard deviation of the intensity value of the k -th object, and z is the pixel of interest²

As thoroughly discussed by Jurdi et al., [2020] it is also possible to include other types of information. For example, Kervadec, Dolz, Tang, et al., [2019] and Zhou, Li, et al., [2019] argue that anatomical organs cannot have arbitrary dimension, and suggest to include prior knowledge about realistic organ sizes in the model optimisation. However, including these priors requires the knowledge of well-defined statistics that sometimes do not fully represent the actual population. For instance, in medical imaging, we usually deal with data obtained from out-of-distribution subjects, which suffer from diseases or can have injuries and abnormal anatomies. In these cases, off-the-shelf population statistics such as the “normal” size of the heart, “normal” cerebral structure, etc., poorly represent the data on which we test the developed models.

An alternative approach is learning a more general form of prior regarding organ size, geometry, location, and other defining aspects directly from the data. In particular, it is possible to use a neural network to learn the data distribution explicitly, via a Variational Autoencoder (VAE), or implicitly, through a Generative Adversarial Network (GAN). We describe VAEs and GANs in detail in Section 3.6.2 and Section 3.6.3 respectively. Within Chapter 5, we analyse in detail several methods falling within the GAN family. Furthermore, we will demonstrate that it is possible to learn priors that can regularise a segmentor model both at train time (Chapter 6 and 7) and during inference (Chapter 8).

²There are also alternative formulations replacing the probability p with the distance between the pixel intensity and a target value. Examples are the use of *Log-Euclidean tensor distance*, *Kullback-Leiber divergence*, and *Rao distance*. Interested readers may refer to Nosrati and Hamarneh, [2016] for a thorough discussion and comparison.

3.5.2 Priors as Design Bias

Another mechanism of introducing prior knowledge into a model is by using design biases. A classic example for multi-class segmentation is the *one-hot* encoding of the model output. With such an encoding, an *argmax* operation associates each pixel of the image to one single object, having the highest score. Thus, it is not possible to assign a pixel to multiple classes at the same time.

There exist different possible biases in the model design. For example, Zheng et al., [2015] suggest to stack a series of convolutional layers to process the features maps extracted by a neural network before classifying their pixels. The suggested design is used to model a Conditional Random Field directly inside the network. Other biases can derive from specific designs of the latent representations, such as the use of hyperbolic and hyper-spherical embeddings (Nickel and Kiela, [2017]; Khrulkov et al., [2020]; Kong and Fowlkes, [2018]), or can be obtained under specific constraints on the features maps variations (Liu*, Thermos*, et al., [2021]).

In Chapter 4 we show how to constrain the latent representations of an autoencoder to be disentangled by binarising the quantitative content-related information of the image, and then imposing an information bottleneck on the residuals. In Chapter 6 we report an example of adversarial and conditioning of attention maps which, linked to an adversarial discriminator, force a segmentor to learn multi-scale relationships in the object to segment.

3.5.3 Priors as Data Bias

Obviously, data can be a useful bias either via biased sampling or measurement errors. For example, a common classifier problem is *class imbalance*, where the dataset contains many instances of class A but limited samples of class B. In medical imaging, the class imbalance can happen because of the population study, which can cover many different biological aspects or be biased toward a sex-gender or a specific age range.

A large body of research focuses on solving the imbalance problem to avoid the model from focusing on the majority class rather than learn-

ing from all the classes. For example, it is possible to sample balanced batches via undersampling of the majority class (Kubat, Matwin, et al., 1997) or the oversampling of the minority class. The latter category of techniques was introduced with the name of Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) and proved to be better than undersampling the majority class. There are also more recent improvements of the SMOTE methodology, including those proposed by Shrivastava, Gupta, and Girshick, 2016 and Dong, Gong, and Zhu, 2017, where the oversampling is adapted to the training process.

An alternative to sampling techniques, it is often convenient to weigh more the loss of the under-represented classes, such as in the case of weighted cross-entropy and focal loss (Lin, Goyal, et al., 2017).

There are other forms of unintentional data biases that regard the intrinsic properties of the image. For instance, Kayhan and Gemert, 2020 proved that CNNs can also exploit the spatial position of the objects for their predictions. A related issue is the data distribution shift problem. For example, models trained on scanners obtained from a vendor A do not necessarily perform well on scanners of the vendor B, as well as optimising a model on MRI images will not necessarily make it work on CT images. In these cases, one can obtain robust models by including data from multiple data sources during training, or adapting the model to work on a different data distribution (Xu, 2019; Toldo et al., 2020).

3.6 Representation Learning and Deep Generative Models

Training with limited supervision is difficult. Thus, it is necessary to constraint models to learn good high-level representations of the data, which must be robust to confounding factors and be informative enough to generalise on new data. There is a shared belief that generative models are a great option to learn good data representations as the ability to synthesise the observed data distribution entails some form of understanding it (Karpathy et al., 2016). In other words, by learning the data generating factors, generative models can potentially use them to improve down-

stream tasks, too. Variational Autoencoders and Generative Adversarial Networks are two popular classes of generative models that we often use in this thesis. Below, we first describe standard autoencoding architectures, then we briefly review the main concepts of these generative frameworks.

3.6.1 Autoencoders (AE)

Autoencoders (AE) consist of an encoding and a decoding module. An encoder maps input data samples to lower-dimensional feature representations z , named *code*, which is used by a decoder to reconstruct the input again. The core idea behind AEs is that by constraining the information flow through a bottleneck, the encoder is forced to extract a compact and meaningful representation z of the input, ignoring nuisance factors and propagating enough information to allow a good reconstruction through the decoder. For example, Denoising Autoencoders (Vincent et al., 2008) minimise the reconstruction error after applying a stochastic corruption to the input, and thus learn a corruption-free representation.

It has been proved (Bourlard and Kamp, 1988) that AEs using only one linear hidden layer and the mean squared error criterion to train, extract a latent representation corresponding to the k principal components of the data, where k is the number of hidden units. On the contrary, if the hidden layer is non-linear, the AE can capture multi-modal aspects of the input distribution (Japkowicz, Hanson, and Gluck, 2000).

Autoencoders are widely used in Machine Learning, and they are often the building blocks that inspired many popular architectures, such as the Variational AE and the UNet (Kingma and Welling, 2014; Ronneberger, Fischer, and Brox, 2015).

3.6.2 Variational Autoencoders (VAE)

Variational Autoencoders (VAE) (Kingma and Welling, 2014; Rezende, Mohamed, and Wierstra, 2014; Kingma, Salimans, et al., 2016) are a probabilistic approach to the AE framework. These models are considered as “latent variable models” because they pair a set of observable variables

to a set of latent variables using an encoder and a decoder neural network (sometimes also termed *inference model* and *generator*, respectively). The encoder maps latent variables \underline{z} sampled from a prior distribution $p(\underline{z})$ to samples of the data distribution, $\mathbf{x} \sim p(\mathbf{x} | \underline{z})$, while the encoder learns to predict the latent variables associated to the data samples, $\underline{z} \sim p(\underline{z} | \mathbf{x})$. From the Bayes's theorem, we know that given the marginal probabilities $p(\mathbf{x})$ and $p(\underline{z})$ (i.e., the probabilities of observing \mathbf{x} and \underline{z} , respectively), and given the likelihood of observing \mathbf{x} given \underline{z} , the posterior is $p(\underline{z} | \mathbf{x}) = \frac{p(\mathbf{x}|\underline{z})p(\underline{z})}{p(\mathbf{x})}$. Hence, computing the posterior requires evaluating the integral $p(\mathbf{x}) = \int p(\mathbf{x} | \underline{z})p(\underline{z})d\underline{z}$, which is intractable. For this reason, instead of computing it analytically, one can try to approximate the posterior $p(\underline{z} | \mathbf{x})$ with a parametric distribution $q(\underline{z} | \mathbf{x})$. In VAEs, $q(\underline{z} | \mathbf{x})$ is a multivariate Gaussian distribution, whose parameters (mean and variance) are predicted by a stochastic encoder. As a result, an input \mathbf{x} can be associated with different levels of probability with multiple possible \underline{z} .

Training is done by maximizing the marginal log-likelihood $\log p(\mathbf{x})$:

$$\begin{aligned} \log p(\mathbf{x}) &= \mathbb{E}_{q(\underline{z}|\mathbf{x})} \left[\log \frac{p(\mathbf{x}, \underline{z})}{p(\underline{z} | \mathbf{x})} \right] = \mathbb{E}_{q(\underline{z}|\mathbf{x})} \left[\log \frac{p(\mathbf{x}, \underline{z})}{p(\underline{z} | \mathbf{x})} \frac{q(\mathbf{x}, \underline{z})}{q(\underline{z} | \mathbf{x})} \right] \\ &= \mathbb{E}_{q(\underline{z}|\mathbf{x})} [\log p(\mathbf{x}, \underline{z}) - \log q(\underline{z} | \mathbf{x})] + \mathbb{E}_{q(\underline{z}|\mathbf{x})} \left[\log \frac{q(\underline{z} | \mathbf{x})}{p(\underline{z} | \mathbf{x})} \right] \\ &= \mathbb{E}_{q(\underline{z}|\mathbf{x})} [\log p(\mathbf{x}, \underline{z}) - \log q(\underline{z} | \mathbf{x})] + D_{KL}(q(\underline{z} | \mathbf{x}) || p(\underline{z} | \mathbf{x})), \end{aligned} \tag{3.1}$$

and since the KL divergence term D_{KL} is non-negative³, it follows that:

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q(\underline{z}|\mathbf{x})} [\log p(\mathbf{x}, \underline{z}) - \log q(\underline{z} | \mathbf{x})] := \text{ELBO}. \tag{3.2}$$

The term on the right of the equation is called Evidence Lower Bound (ELBO) and maximising it is equivalent to maximising the log-likelihood

³For the Jensen inequality, given the probability distributions p_1 and p_2 , we have $D_{KL}(p_1 || p_2) := \mathbb{E}_{p_1} \left[-\log \left(\frac{p_2}{p_1} \right) \right] \geq -\log \mathbb{E}_{p_1} \left[\frac{p_2}{p_1} \right] = -\log \left(\int p_1 \cdot \frac{p_2}{p_1} \right) = -\log(1) = 0$, being $-\log$ a convex function and $\int p_2 = 1$.

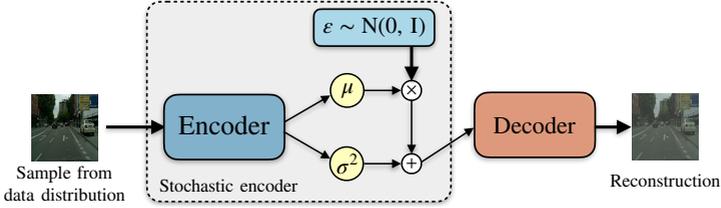


Figure 9: Variational Autoencoder (VAE) schematic. A stochastic encoder learns to map input samples to a prior probability distribution, predicting mean and variance of the distribution. The decoder attempts to reconstruct the input by drawing a sample from the predicted distribution, thanks to the reparametrization trick.

term. We can further expand this term as:

$$\begin{aligned}
 \log p(\mathbf{x}) &\geq \text{ELBO} \\
 &= \mathbb{E}_{q(\underline{z}|\mathbf{x})} [\log p(\mathbf{x}, \underline{z}) - \log q(\underline{z} | \mathbf{x})] \\
 &= \mathbb{E}_{q(\underline{z}|\mathbf{x})} [\log p(\mathbf{x} | \underline{z}) + \log p(\underline{z}) - \log q(\underline{z} | \mathbf{x})] \quad (3.3) \\
 &= \mathbb{E}_{q(\underline{z}|\mathbf{x})} [\log p(\mathbf{x} | \underline{z})] + \mathbb{E}_{q(\underline{z}|\mathbf{x})} [\log p(\underline{z}) - \log q(\underline{z} | \mathbf{x})] \\
 &= \mathbb{E}_{q(\underline{z}|\mathbf{x})} [\log p(\mathbf{x} | \underline{z})] - D_{KL}(q(\underline{z} | \mathbf{x}) \| p(\underline{z})).
 \end{aligned}$$

As can be seen, the right hand side of the equation contains a term measuring the likelihood of the reconstructed data output of the decoder ($\mathbb{E}_{q(\underline{z}|\mathbf{x})}[\log p(\mathbf{x} | \underline{z})]$), and a term measuring the KL divergence between the posterior distribution $q(\underline{z} | \mathbf{x})$ and the prior $p(\underline{z})$.

In particular, we can model $p(\underline{z})$ as a standard Gaussian distribution $N(\underline{\mu} = 0, \underline{\sigma} = 1)$, which allows us to write a closed form solution for the KL term (Kingma and Welling, 2014):

$$-D_{KL}(q(\underline{z} | \mathbf{x}) \| p(\underline{z})) = \frac{1}{2} \left[1 + \log(\underline{\sigma}_q^2) - \underline{\sigma}_q^2 - \underline{\mu}_q^2 \right], \quad (3.4)$$

where $\underline{\mu}_q, \underline{\sigma}_q^2$ are the mean and the variance of the approximate distribution $q(\underline{z} | \mathbf{x})$. As a result, the VAE can be trained to maximize:

$$\mathcal{G} = \mathbb{E}_{q(\underline{z}|\mathbf{x})} [\log p(\mathbf{x} | \underline{z})] + \frac{1}{2} \left[1 + \log(\underline{\sigma}_q^2) - \underline{\sigma}_q^2 - \underline{\mu}_q^2 \right], \quad (3.5)$$

or, equivalently, to minimize the loss $\mathcal{L} = -\mathcal{G}$.

We can use a decoder neural network to estimate $p(\mathbf{x} | \mathbf{z})$, which corresponds to the reconstruction error of \mathbf{x} , and we can use an encoder neural network to estimate the mean $\underline{\mu}_q$ and variance $\underline{\sigma}_q^2$. Finally, to be able to train the VAE bypassing the sampling procedure, we can back-propagate gradients from the decoder to the encoder thanks to the reparametrization trick, which converts stochastic sampling to the deterministic operation: $z_i = \underline{\mu}_i + \underline{\sigma}_i \cdot \varepsilon$, where $\varepsilon \sim N(0, 1)$. We summarize the VAE framework in a schematic in Figure 9.

It is important to note that, if we use a Gaussian prior $p(\mathbf{z})$ (i.e., its probability has a diagonal covariance matrix), we can think of the latent representation as a set of independent additive Gaussian noise channels z_i transmitting independent information about the input \mathbf{x} . In this context, the KL divergence term of the VAE objective acts as an upper bound on the information that can be transmitted via the latent channels (Burgess et al., 2018; Higgins, Matthey, et al., 2017), which can be made even tighter by multiplying it by a scalar $\beta > 1$ (Higgins, Matthey, et al., 2017). Reconstructing under this restrictions encourages the VAE to embed samples that look similar in the data space into nearby positions of the latent space, constructing a smooth latent manifold.

3.6.3 Generative Adversarial Networks (GAN)

Generative Adversarial Networks (Goodfellow et al., 2014) are deep generative models trained in a *mini-max* game. During this game, generator neural network Γ plays against against a discriminator neural network Δ (Fig. 10). The role of the discriminator is to compute the probability that an input \mathbf{x} belongs to the distribution of data (i.e., $\mathbf{x} \sim p(\mathbf{x})$) rather than being synthesized by the generator (i.e., $\mathbf{x} \sim q(\mathbf{x})$), or in other terms, it learns to distinguish *real* from *fake* input samples, respectively. Concurrently, the generator learns a mapping from a prior distribution $p(\mathbf{z})$ to the data distribution $p(\mathbf{x})$, such that the discriminator cannot distinguish its predictions from the real data.

During the adversarial game, the discriminator is trained in a supervised manner to distinguish *real* vs. *fake* samples, while the generator is

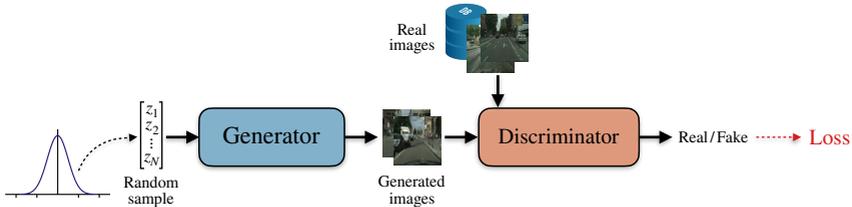


Figure 10: Generative Adversarial Network (GAN) schematic. A generator learns to map random inputs from a known distribution into samples of the target data distribution. Meanwhile, the discriminator is trained to say apart images generated from the generator and images sampled from the data distribution.

trained leveraging the gradients of the discriminator. In its vanilla formulation (Goodfellow et al., [2014]), this game can be formalized as:

$$\min_{\Gamma} \max_{\Delta} V(\Delta, \Gamma) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} [\log \Delta(\mathbf{x})] + \mathbb{E}_{z \sim p(z)} [\log(1 - \Delta(\Gamma(z)))]. \quad (3.6)$$

In practice, we train Δ to maximize the probability of assigning the correct label to both training and generated samples; we simultaneously train Γ to minimize $\log(1 - \Delta(\Gamma(z)))$. In this setup, the optimal discriminator is:

$$\Delta^*(\mathbf{x}) = \frac{p(\mathbf{x})}{p(\mathbf{x}) + q(\mathbf{x})}. \quad (3.7)$$

If we consider an optimal discriminator in Equation [3.6], we obtain:

$$V(\Delta^*, \Gamma) = 2 \cdot D_{JS}(p(\mathbf{x}) \| q(\mathbf{x})) - \log 4, \quad (3.8)$$

that is, the generator is trained to minimize the Jensen–Shannon divergence D_{JS} (symmetric form of the KL divergence) between generated and real data distributions (Goodfellow et al., [2014]), leading generated samples to become gradually more realistic (Fig. [11]).

There are alternative ways of training GANs, which minimize the Pearson divergence (Mao, Li, et al., [2017]), or the Wasserstein distance (Arjovsky, Chintala, and Bottou, [2017]; Gulrajani et al., [2017]) between real and fake data distributions, making the training process easier. We will discuss these and other GAN variants more in depth in Chapter [5], where we study their application for semi-supervised image segmentation.

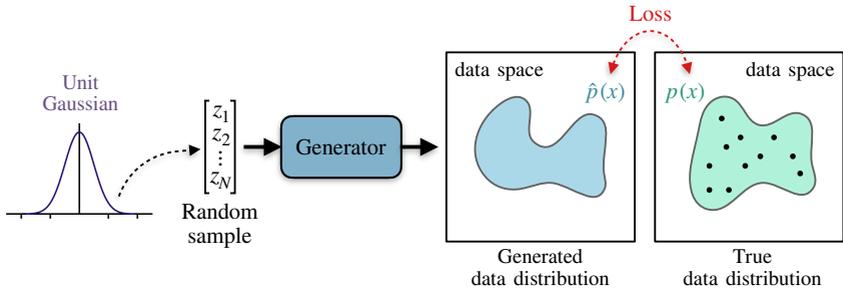


Figure 11: The generator of a GAN learns to map random inputs to samples of the data distribution. The generator is trained to minimise a distance, or a divergence, between the distributions of the real and the generated data.

Conditional GANs

In the context of image-to-image translation and semantic segmentation, a popular variant of the GAN framework is the *conditional* GAN proposed by Mirza and Osindero, [2014]. In a conditional GAN, a supplementary input conditions the model to predict an image with definite properties, such as reporting content of a semantic class, or specific appearance. For example, the Pix2Pix framework (Isola et al., [2017]) uses a generator receiving as input a random sample from a Normal distribution and an input image in one domain (which acts as conditioning factor). Given the two inputs, Pix2Pix optimises the generator to predict an output image in a target domain. The network is trained with a supervised cost when input-output image pairs are available, with the adversarial loss for unpaired data. Another popular conditional GAN is the CycleGAN (Zhu et al., [2017]), which overcomes the lack of paired training data using the cycle consistency loss, based on the principle that images translated from domain A to a domain B , and back to A again, should remain identical to themselves. These and other conditional GANs are widely used in semantic segmentation, as they can introduce a data-driven shape prior in the model, and thus provide unsupervised training signals on the unlabeled images. In our thesis, we use them in the context of semi-supervised learning in Chapter [4] and [5], and for weakly annotated data

in Chapter 6. We also present their usage to improve performance on challenging test samples, in Chapter 8.

The Stability Problem

A common problem with GANs is that they often exhibit unstable behaviour during training (Chu, Minami, and Fukumizu, 2020; Arjovsky and Bottou, 2017; Gulrajani et al., 2017; Miyato et al., 2018). This problem has motivated the research of alternative training objectives (Mao, Li, et al., 2017; Arjovsky, Chintala, and Bottou, 2017; Gulrajani et al., 2017) to make model convergence easier, and to prevent the generator from learning only a small subset of the target data distribution (a phenomenon known as *mode collapse*).

As recently pointed out by Sønderby et al., 2017, we can find the reasons for GANs' instability in three main assumptions that we usually make about GANs and which may not always be satisfied. In the first place, we commonly assume the log-likelihood ratio $\log \frac{q(\mathbf{x})}{p(\mathbf{x})}$ is finite. Secondly, we expect the Jensen-Shannon divergence to be a well-behaved function in the weights search space. Finally, we assume there is a single optimal discriminator. When at least one of these hypothesis does not hold, GANs fail to converge.

In the literature, many strategies attempt to address the convergence problem of GANs. As highlighted by Chu, Minami, and Fukumizu, 2020, most approaches focus on the discriminator network. The primary motivation for this choice is that the discriminator produces the signal used to drive the generator training. Thus, having good discriminators is fundamental to train powerful generators.

In this context, aside from using alternative loss functions for training GANs, several additional strategies aim to boost discriminator performance. In particular, most methods force the discriminator to learn a smooth function or to limit its overfitting. We describe these techniques more in detail in Chapter 5.

The stability problem of GANs is yet to be solved. In Chapter 5, we will introduce a complementary technique to stabilise training of adversarial mask discriminators. Such a technique is not related to the smooth-

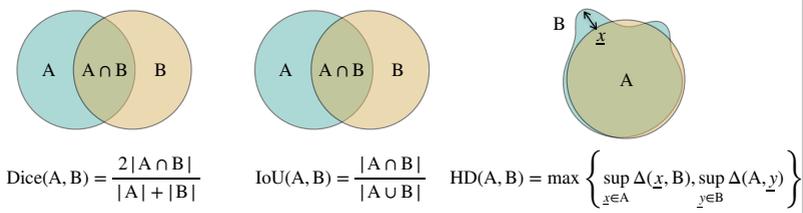


Figure 12: Graphical visualisation of metrics used to evaluate segmentation performance.

ness of the discriminator, but it improves its training when dealing with flat segmentation masks.

3.7 Segmentation Metrics

We evaluate the methods presented in the subsequent chapters by comparing with state-of-the-art benchmark models for semantic segmentation. In the following, we briefly describe the metrics used to evaluate performance, which we also summarise in Figure [12](#).

Given a two sample sets A and B , we define the following metrics:

- **Sørensen–Dice coefficient**, also known as Dice score (Dice, [1945](#); Sørensen, [1948](#)). It is defined as twice the size of the intersection divided by the sum of the two sets:

$$\text{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|} = \frac{2|A \cap B|}{|A \cup B| + |A \cap B|}.$$

The score can have values in the $0 \div 1$ range, where 0 reflects a complete mismatch between two segmentation masks and 1 their perfect overlap. In the classic formulation, the Dice score is computed for each class considered in a segmentation task, and then averaged over the classes. There exist also a multi-class formulation (Crum, Camara, and Hill, [2006](#)), where all the classes are considered as one set, and the intersection is over the entire set of classes.

- **Intersection over Union (IoU)**, also known as the Jaccard similarity coefficient (Jaccard, [1912]; Tanimoto, [1958]). The coefficient measures the similarity between two finite sample sets as the size of the intersection divided by the size of the union:

$$\text{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Given two segmentation masks, the IoU score can have values in the $0 \div 1$ range, with 0 reflecting a complete segmentation mismatch and 1 their perfect overlap.

- **Hausdorff distance** (Blumberg, [1920]): measures the distance $\Delta(\cdot)$ between two subsets of a metric space, and it is defined as the greatest of all the distances from a point \underline{x} in one set to the closest point \underline{y} in the other set. Mathematically, we can write:

$$\text{HD}(A, B) = \max \left\{ \sup_{\underline{x} \in A} \Delta(\underline{x}, B), \sup_{\underline{y} \in B} \Delta(A, \underline{y}) \right\}.$$

The Hausdorff distance has the minimum possible value of 0 when there is a perfect overlap of two segmentation masks. There is no fixed upper bound, and the metric is not defined when one of the two masks is empty. In the latter case, it is possible to ignore the mask comparison or to assign to the metric the value of the maximum possible distance we can measure (i.e. the mask dimension) (Reinke et al., [2021]). In this thesis, we will adopt the second approach.

3.8 Datasets

Throughout this work we use public clinical and non-clinical datasets. We describe these datasets below and give an overview in Table [1]. Medical datasets contain data obtained by a number between 20 and 320 different patients, which result in approximately 1,600 to 98,000 images. This number of images can be considered small compared to popular

Dataset	Medical	O. I.	Modality	Subjects	Classes	Images
ACDC	Yes	Cardiac	Cine-MR	150	4	38,346
LVSC	Yes	Cardiac	Cine-MR	100	2	23,218
M&Ms	Yes	Cardiac	Cine-MR	320	4	98,810
CHAOS	Yes	Abdomen	MR (T1, T2)	20	5	1,594
PPSS	No	Pedestrians	Surveillance Cameras	3,961	7	3,961

Table 1: Overview of the datasets used in this thesis. For each dataset, we report if it contains medical data, the object of interest (O.I.), the modality used to acquire the images, how many different subjects the dataset contains, how many classes are annotated (including the background), and the total number of images. Please, refer to Section 3.8 for additional details.

computer vision datasets, such as ImageNet (Deng et al., 2009) that contains about 14 million images. However, this is a typical size for medical datasets. In fact, collecting medical data is more challenging, and it entails ethical and privacy processes. Despite the size of the employed dataset is sufficient for research purposes, it would be necessary to conduct a large-scale study to evaluate the use of commercial applications for the clinical setting (Park and Han, 2018).

We now present a description of these datasets for cardiac and abdominal images. We also describe a computer vision dataset which we use to explore the model generalisation capability on non-medical data. After describing each dataset, we provide details on the pre-processing operations we performed before using it. Lastly, Section 3.8.6 details the data augmentation strategies used in the thesis to artificially increase the dataset size.

3.8.1 ACDC: Automatic Cardiac Diagnosis Challenge

This dataset contains data from the Automatic Cardiac Diagnosis Challenge (Bernard et al., 2018), presented at MICCAI 2017.

The ACDC dataset was created from real clinical exams and consists of 2-dimensional cine-MR images acquired by 100 patients using various 1.5T and 3T MR scanners. It is possible to test a segmentation method on additional 50 patients, for which annotations are not

provided, using the challenge server⁴. Overall, data covers five evenly distributed subgroups of patients: healthy subjects, subjects with previous myocardial infarction, subjects with dilated cardiomyopathy, subjects with dilated cardiomyopathy, and subjects with abnormal right ventricle. The dataset also provides additional information regarding each subject: weight, height, and diastolic and systolic phase instants.

Images have a spatial resolution between 1.22 and 1.68 $mm^2/pixel$, and for each patient, there is a number of cardiac phases ranging between 28 and 40 images. In the dataset, there are manual segmentations provided in correspondence of the end-diastolic (ED) and end-systolic (ES) cardiac phases for three anatomical structures: right ventricle (RV), left ventricle (LV) and left myocardium (MYO). In total, there are 1,902 images with manual segmentations (corresponding to ED and ES instants) and 23,449 images with no segmentation (from the remaining cardiac phases). Moreover, the challenge server allows testing the model for a limited number of times on 12,995 images.

ACDC Scribble Annotations

In our work, we collected manual scribble annotations for the ACDC annotated patients that is possible to download. It is possible to use these annotations to experiment with weakly supervised methods. After training, these methods can be evaluated through the ACDC challenge server or using the available fully annotated masks. We published these data in Valvano, Leo, and Tsaftaris, 2021c, and we will discuss them further in Chapter 6. These weak annotations were manually drawn within the available segmentation masks for RV, LV and MYO, in the ES and ED cardiac instants.

Pre-processing

Given a patient volume scan, we consider outliers and clip image pixels with values outside the 5th to 95th percentiles interval. We resample

⁴<https://www.creatis.insa-lyon.fr/Challenge/acdc/databasesTesting.html>

images to the average resolution of $1.51mm^2$. For each patient, we normalise data by removing the patient-specific median and dividing by the inter-quartile range. In Chapter 4, we deal with memory constraints by cropping images to 176×176 pixel size. In the other chapters of the thesis, we instead crop or pad them to 224×224 image sizes.

3.8.2 LVSC: Left Ventricular Segmentation Challenge

The Left Ventricular Segmentation Challenge dataset (Suinesiaputra et al., 2014) is part of the Cardiac Atlas Project (Fonseca et al., 2011) and has been introduced at the STACOM 2011 MICCAI workshop. It contains gated Steady-State Free Precession (SSFP) MRI pulse sequence, in short- and long-axis views, acquired by patients with myocardial infarction.

Images were acquired using a mix of 1.5T scanner types and imaging parameters and have a spatial resolution between 138×192 and 512×512 pixels. The temporal resolution is between 19 and 30 frames per patient. In total, there are manual segmentations for 100 subjects, for a total of 23,218 images, which cover the left ventricular myocardium (MYO) in all the cardiac phases.

Pre-processing

Given a volume scan, we clip outliers outside the 5^{th} to 95^{th} percentiles interval. Then, we resample images to the average resolution of $1.45mm^2$. Finally, we normalise data by removing the median and dividing by the inter-quartile range computed for each patient. Similar to ACDC, in Chapter 4, we deal with memory constraints cropping images to 176×176 pixel size. In the other chapters of the thesis, we instead crop/pad images to 224×224 pixel size.

3.8.3 M&Ms: Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Image Segmentation Challenge

This dataset (Campello et al., 2021) contains cardiac images obtained from 320 different subjects scanned on six clinical centres in 3 different

countries. In total, there are 98,810 images, acquired using four different scanner vendors, detailed below:

- Siemens (vendor A): 95 subject, acquired with a spatial resolution of $1.32\text{mm}^2/\text{pixel}$ and with a number of 25 temporal frames.
- Philips (vendor B): 125 subjects, acquired with a spatial resolution of $1.30\text{mm}^2/\text{pixel}$ and with a number of temporal frames ranging between 18 and 30.
- General Electric (vendor C): 50 subjects, acquired with a spatial resolution of $1.37\text{mm}^2/\text{pixel}$ and with a number of temporal frames ranging between 20 and 30.
- Canon (vendor D): 50 subjects, acquired with a spatial resolution of $0.85\text{mm}^2/\text{pixel}$ and with a number of temporal frames ranging between 20 and 36.

Similar to ACDC, the dataset reports manual segmentation masks for the right ventricle (RV), left ventricle (LV) and left myocardium (MYO) in correspondence of the end-systolic and end-diastolic temporal instants.

Pre-processing

We clip pixel intensities outside the 5^{th} to 95^{th} percentiles interval, considering them outliers. We resample images to the average resolution of 1.25mm^2 . Then, we crop or pad them to 224×224 pixel size. Lastly, we normalise data by subtracting the patient-specific median and dividing by the interquartile range.

3.8.4 CHAOS: Combined Healthy Abdominal Organ Segmentation

The Combined Healthy Abdominal Organ Segmentation dataset (Kavur, Gezer, Barış, Aslan, et al., [2021](#)) contains data released for the abdominal segmentation challenge (Kavur, Gezer, Barış, Aslan, et al., [2021](#); Kavur, Gezer, Barış, Şahin, et al., [2020](#)), that was part of the ISBI 2019. Images

were acquired with 1.5 MR scanners, using T1-dual inphase and T2-SPIR sequences. In total, there are 1,594 DICOM images, with 256×256 spatial resolution. The dataset contains abdominal MR images of 20 subjects, alongside with segmentation masks of liver, kidneys, and spleen.

Pre-processing

For each patient, we clip intensity values outside the 5^{th} to 95^{th} percentiles interval, which we consider outliers. Similar to Chartsias, Joyce, et al., [2019], we resample data to $1.89mm^2$ resolution and then normalise in between -1 and 1. Finally, we crop images to 192×192 pixel size.

3.8.5 PPSS: Pedestrian Parsing in Surveillance Scenes

The Pedestrian Parsing in Surveillance Scenes dataset (Luo, Wang, and Tang, [2013]) contains 3,961 RGB images of pedestrian, derived from 171 surveillance videos. Images were obtained using different cameras and resolutions, and present occlusion. The authors recommend using the first 100 surveillance scenes for training, and images from the remaining 71 cameras for testing. Besides images, ground truth segmentations are given for seven parts of the pedestrians: hair, face, upper clothes, arms, legs, shoes, and background. Images have a resolution varying between 224×212 and 808×404 pixels, while segmentation masks have resolution of 80×160 pixels.

Pre-processing

We resample all the RGB images to the same spatial resolution of the segmentation masks: 80×160 pixel size. Finally, we normalise them by rescaling values in the $[0, 1]$ range.

3.8.6 Data Augmentation

Data augmentation is an effective strategy to prevent overfitting, especially when lacking large-scale labelled datasets (Tajbakhsh et al., [2020]). For this reason, data augmentation has become a standard practice in the

image processing pipeline of learning algorithms, synthesising new data through random transformations of the images available in the training set. Despite having high mutual information with the data used to generate it, the augmented data helps to introduce transformation *equivariance* and *invariance* in the optimised model (Tajbakhsh et al., 2020).

In our experiments, we always use data augmentation, which we apply on the 2D images during training, at run-time. The operations used to augment the data are:

- **Image translation:** between $\pm 10\%$ of the pixels on both vertical and horizontal axis;
- **Image rotation:** between $\pm \pi/2$ on the medical datasets (ACDC, LVSC, M&Ms and CHAOS) and between $\pm \pi/6$ on the vision dataset (PPSS);
- **Random image intensity perturbation:** addition of a small *random noise*, sampled from a Normal distribution $N(\mu = 0; \sigma = 0.02)$; *brightness* transformations, with a maximum delta of 0.025; and *contrast* changes of $\pm 5\%$ of the image intensity range.

We perform all these operations using standard libraries available in TensorFlow (Abadi et al., 2016).

3.9 Summary

This chapter discussed the technical background of the thesis. We introduced the used mathematical notation, and we defined the learning algorithms and possible learning paradigms. When dealing with partial or scarce annotations, we explained that it is necessary to regularise models, either directly or indirectly constraining their learned data representations. Then, we gave an overview of recent representation learning families of algorithms that we use in this thesis: Autoencoders (AEs), Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). Finally, we introduced segmentation metrics, datasets and data processing operations used for our experiments.

The following chapters present our approach to semantic segmentation when missing abundant and high-quality annotations. In each chapter, we also include a related work section highlighting the state-of-the-art literature when we first released the method.

Chapter 4

Regularising Disentangled Representations using Spatio-Temporal Priors

▣

Deep neural networks have shown to be promising approaches for medical image analysis. However, their training is most effective when they learn robust data representations using large-scale annotated datasets, which are tedious to acquire in clinical practice. As medical annotations are often limited, there has been an increasing interest in making data representations robust in a lack of data. As we briefly discussed in the previous chapter (Section [3.4](#)), a spate of research aims to do so by constraining the learned representations to be interpretable and able to separate out, or *disentangle*, the data explanatory factors.

This chapter is based on:

- Valvano, Gabriele, Agisilaos Chatsias, Andrea Leo, and Sotirios A. Tsaftaris (2019). “Temporal Consistency Objectives Regularize the Learning Of Disentangled Representations”. In: *Domain Adaptation and Representation Transfer (DART)*. Springer, pp. 11–19. ISBN: 978-3-030-33391-1
- Valvano, Gabriele, Andrea Leo, and Sotirios A. Tsaftaris (2021e). “Regularising Disentangled Representations With Anatomical Temporal Consistency”. In: *Under Review at: Biomedical Image Synthesis and Simulations, Elsevier*

This chapter discusses recent disentanglement frameworks, with a particular focus on image segmentation. We build on a recent approach to disentangling cardiac medical images into disjoint *patient anatomy* and *imaging modality* dependent representations. In the model, we incorporate a purposely designed architecture (which we term “temporal transformer”) which, from a given cine MR image and a time-gap, can estimate anatomical representations of the image at a future time-point of the cardiac cycle. The transformer’s role is to introduce a self-supervised objective to encourages the emergence of temporally coherent data representations. We show that such a regularisation improves the quality of disentangled representations, ultimately increasing semi-supervised segmentation performance when annotations are scarce.

4.1 Introduction

The performance of machine learning algorithms largely depends on their ability to extract good high-level representations from the data (Bengio, Courville, and Vincent, 2013), which is a challenging problem and usually requires large quantities of labelled data. Unfortunately, collecting large-scale fully-annotated medical datasets is expensive and requires experts.

On the other hand, semi-supervised learning (SSL) suggests that it is possible to include unlabelled data to train better models, exploiting data correlations. For example, in medical image segmentation, physicians may annotate only the end-diastolic and the end-systolic temporal instances of a cardiac cine MRI (Bernard et al., 2018). Yet all the images in the cardiac cycle may be used to add knowledge into the model. It is common to formulate SSL as a multi-task learning problem (Cheplygina, de Bruijne, and Pluim, 2019; Ouali, Hudelot, and Tami, 2020; Salimans et al., 2016; Chatsias, Joyce, et al., 2019), where one minimises a supervised cost on the annotated images, but also other unsupervised or self-supervised objectives, which do not require labels. For example, it is possible to train a model to perform object segmentation while also minimising a self-reconstruction cost. Sharing model parameters across

tasks leads to more rich and meaningful data representations.

However, improving the supervised task in multi-task learning is only possible when the tasks do not compete, which is not always the case (Gong et al., 2018). For example, optimising one learning objective may require data representations which instead hamper the convergence of another training objective. A possible workaround to the problem is constraining the representation to separate out, or *disentangle*, features useful for both tasks from those that are task-specific (Bengio, 2009; Achille and Soatto, 2018; Van Steenkiste et al., 2019).

Recently, there has been an increasing interest in learning disentangled representations for many computer vision applications, such as image-to-image translation (Liu, Breuel, and Kautz, 2017; Lee, Tseng, et al., 2018; Huang et al., 2018), semantic segmentation (Chartsias, Joyce, et al., 2019), and landmark detection (Lorenz et al., 2019). These methods usually decompose an image into two subsets of representations: the *content* and the *style*. The image content aims to capture spatial information required for spatially-equivariant tasks, such as object detection and segmentation. On the other hand, the style representation captures image appearance in terms of colour intensity and textures. The hope of such decomposition and desired equivariances and invariances is to push semantic meaning into the different information contents. In medical imaging, we can associate the image content with the anatomical information varying across patients. Instead, the image style contains the imaging modality's information, which changes with scanner and acquisition physics. It is also possible to further factorise the representation. For example, decoupling the spatial information related to specific anatomical structures assists semantic segmentation tasks (Chartsias, Joyce, et al., 2019). Disentangling pathology helps for pathology segmentation and pseudo-healthy image synthesis (Jiang et al., 2020; Xia, Chartsias, and Tsaftaris, 2020). Disentangling artefacts helps to improve the image quality and the subsequent analysis (Liao et al., 2019).

In this chapter, we focus on the task of cardiac image segmentation. We discuss whether it is possible to regularise the learning of disentangled representation in cardiac MRI by exploiting the anatomical region-

specific spatio-temporal dynamics. In particular, we show that inductive biases, such as temporal coherence, are of fundamental importance to encourage the model to deal with real-world dynamics and improve generalisation. Herein, we leverage the temporal evolution of the heart’s contraction as captured by unlabelled cine MRIs. We use a self-supervised objective to constrain the latent representation to be predictable in time. As a result, we improve segmentation performance on unseen data.

4.1.1 Contributions

In the remaining of this chapter, we adopt SDNet, a framework that Chartsias, Joyce, et al., [2019] introduced in the context of medical imaging for object segmentation via disentangled representations. SDNet decouples factors specific to the imaging modality (style) from those related to the patient anatomy (content). Compared to other frameworks for content-style disentanglement (Yang et al., [2019]; Qin et al., [2019]), SDNet discretises content representation to preserve pixel-to-pixel correspondences with the image whilst encouraging the removal of continuous modality-related information from the spatial content representation. This additional discretisation bottleneck encourages disentanglement but also provides a more interpretable content representation (Liu*, Thermos*, et al., [2021]), which is of importance for healthcare applications.

We endow SDNet with the ability to predict anatomical temporal dynamics, inherently building a better representation that increases model robustness in a scarcity of annotations. We graphically present the method in Figure [13] and summarise the key aspects as follows:

- We regularise the learning of disentangled representations in SDNet through a modality invariant transformer that, conditioned on the temporal information, transforms the anatomical factors to predict future instants in a cine MRI.
- We show that the transformer provides a self-supervised signal which has a regularising effect on the extracted representation, and it helps to perform the cardiac segmentation task.

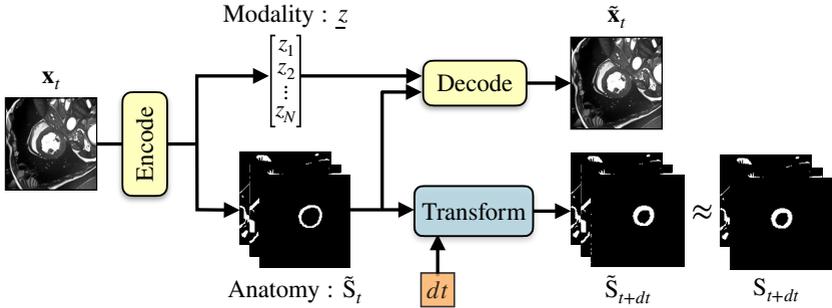


Figure 13: Method overview. Given the input image x_t at time t , the model extracts a multi-channel binary representation S_t (anatomical factors) and a residual vector z (modality factors). In this work, we aim at regularising S_t constraining it to be predictable: conditioned on the temporal gap dt , a neural network must be able to predict the representation at time $t + dt$.

- We report increased performance compared to SDNet for semi-supervised learning when the amount of available annotations decreases, and we achieve comparable results to a fully supervised training using fewer labels.
- We show an example of how it is possible to employ our model for cardiac temporal synthesis.

We made code for reproducing the experiments available at: <https://github.com/vios-s/sdtnet>.

4.2 Related Work

Deep neural networks are excellent tools for medical image analysis (Zhou, Greenspan, et al., 2021), and the UNet (Ronneberger, Fischer, and Brox, 2015) is a popular and effective approach for image segmentation. The UNet has an auto-encoding architecture characterised by skip-connections, i.e. interconnections between the encoder and the decoder at multiple depth levels. Fundamental to the UNet success, the skip-connections limit the gradient vanishing problem and improve

the segmentation of high-resolution details. Unfortunately, training a UNet in standard fully-supervised learning is expensive in terms of label collection, and the model does not perform well when annotations are scarce. For this reason, semi-supervised methods have emerged as appealing alternatives to increase model accuracy while keeping low the labelling cost.

In the following, we first discuss semi-supervised approaches, their assumptions and limitations. Then we discuss why disentangled representations help to learn in semi-supervised settings. Finally, we discuss the importance of learning temporal transitions and how they can increase data representation quality.

4.2.1 Semi-Supervised Learning

Semi-supervised learning is the process of training using both annotated and unannotated data. Under specific assumptions (Chapelle, Scholkopf, and Zien, 2009), optimising training objectives on the unlabelled data also improves the supervised task, for which annotations may be scarce. Among SSL methods, several approaches regularise the training process requiring that the model predictions remain consistent after applying realistic perturbations on the unlabelled data (Taigman et al., 2014; Zhang, Zhang, Odena, et al., 2020; Chaitanya, Erdil, et al., 2020). Other approaches leverage pre-trained models to predict additional labels for unannotated samples (Ouali, Hudelot, and Tami, 2020), which are used for co-training (Blum and Mitchell, 1998; Qiao et al., 2018), self-training (Bai, Oktay, et al., 2017; Ouyang et al., 2020) and multi-view learning (Zhao, Xie, et al., 2017; Noroozi et al., 2018). To SSL also belong generative models that learn the data distribution while also performing a supervised objective. Once learned high-quality features for the generative task, they use them to perform the supervised objective, too (Salimans et al., 2016; Kohl et al., 2018; Yi, Walia, and Babyn, 2019). Finally, graph-based methods consider labelled and unlabelled data as nodes inside a graph, and they learn to propagate labels from the labelled nodes to the unlabelled ones (Grady, 2006; Zheng et al., 2015).

Each of the above SSL categories relies on at least one of the following hypothesis: i) the *manifold assumption*: high dimensional data lie on a low-dimensional manifold. ii) The *smoothness assumption*: if two data points are close, the corresponding model predictions should be close. And iii) the *clustering assumption*: two points that are in the same cluster most likely belong to the same class. In this work, we assume that the smoothness assumption holds even in disentangled features space, and we force the model to map similar images to similar representations. In particular, we regularise SDNet (Chartsias, Joyce, et al., 2019) anatomical representations using their temporal consistency as a self-supervised objective, and improve the segmentation task.

4.2.2 Disentangled Representations

Recently, disentangled representations have been used in many computer vision tasks, endowing machine learning models of the possibility to extract explanatory factors from the data. Higgins, Amos, et al., 2018 recently proposed a formal definition of disentangled representations which exploits the concept of “symmetry transformations”. Symmetry transformations are transformations changing only specific aspects of the real world state while keeping other aspects unchanged (or invariant). According to this definition, a vector representation is disentangled if it can be decomposed it into several sub-spaces, each one of which is compatible with and can be transformed independently by a unique symmetry transformation (Higgins, Amos, et al., 2018). As a result, we must assume that changes in the world state only sparsely affect the representation. Vice versa, localised changes in the encoded data are sparse over real-world transformations.

For these properties, disentangled representations increase the model interpretability, and improve its generalisation on unseen data, thanks to the concept of equivariance. Moreover, confining single factors of variations in specific subsets of features allows interpretable latent code manipulation, which is desirable in many applications, such as modality transfer (Huang et al., 2018; Lee, Tseng, et al., 2018), image generation

(Li, Singh, et al., 2020; Nie et al., 2020), and domain adaptation (Chartsias, Papanastasiou, et al., 2020; Yang et al., 2019; Meng et al., 2020).

A shared definition of disentanglement is still open to debate. However, many researchers think that disentangled representations should be *factorised*: i.e. they should contain statistically independent latent variables (Kim and Mnih, 2018). Obtaining this type of representations would allow compact and meaningful information encoding, which is useful to increase model generalisability (Van Steenkiste et al., 2019) and to increase the robustness against nuisance factors and adversarial attacks (Alemi et al., 2016).

In the context of medical imaging, Chartsias, Joyce, et al., 2019; Chartsias, Papanastasiou, et al., 2020 recently explored the use of factorised representations for semi-supervised learning and multi-modal image segmentation. Qin et al., 2019 exploited disentangled representations for unsupervised domain adaptation. Jiang et al., 2020 used disentanglement in the context of pathology segmentation. However, to the best of our knowledge, this is the first work exploiting temporal information to regularise the learning of disentangled representations and improve the segmentation task.

4.2.3 Improving Disentanglement with Temporal Transitions

Real-world transformations preserve a considerable amount of invariant structure, which profoundly influences the biological vision. For example, objects have smooth temporal dynamics, and thus temporal smoothness facilitates the development of object recognition in humans (Wood, 2016). More broadly, the Sensorimotor Contingencies Theory states that human perception emerges from a sensorimotor flux of data, e.g. experiencing how sensory experience changes in time, or as a consequence of our actions (O'Regan and Noë, 2001). This flux of data has properties and constraints that are learned by our brain to better understand the world around us.

Driven by these observations, Caselles-Dupré, Garcia-Ortiz, and Fil-

liat, [2019] argued that the transitions from one state of the system into another are necessary for learning good disentangled representations. In particular, rather than simply training a neural network with unrelated samples $\{a, b, c\}$, we can introduce temporal transitions by teaching the model to go from the state at time t : $\{a_t, b_t, c_t\}$ to the state at time $t + 1$: $\{a_{t+1}, b_{t+1}, c_{t+1}\}$.

Even in medical imaging, the use of temporal information has been explored. For example, Krebs et al., [2019] used the temporal information contained in cine MRIs to detect cardiac abnormalities via a probabilistic registration model. In the context of image segmentation, Bai, Suzuki, et al., [2018] and Qin et al., [2019] used the temporal information for label propagation on unannotated images. In this chapter, we show that we can use cardiac temporal dynamics to improve the quality of disentangled representations.

Outside medical imaging, Hsieh et al., [2018] proposed to decompose the input images in a set of time-dependent representations (pose) and a set of fixed representations (content). Specifically, they suggest using such a decomposition for video prediction, where they keep the content fixed and make inference on the pose vector to predict future frames in a temporal sequence. With a similar idea, we decompose the image in time-dependent *anatomical* factors and fixed *imaging modality* factors. After such a decomposition, we predict future temporal frames only based on the time-dependent representation. However, we should highlight that our objective is not to obtain good temporal predictions. Rather we demonstrate that learning temporal dynamics regularises the (learning of) disentangled representations, encouraging them to change smoothly and consistently in time. As a direct consequence, we show that this also improves the segmentation capabilities of the model. In other words, this chapter demonstrates once more what is quite known for a while that several correlated tasks in a multi-task learning setting encourage the learning of better representations (Caruana, [1997]). We describe below two self-supervised tasks that aid the performance of a supervised (segmentation) task.

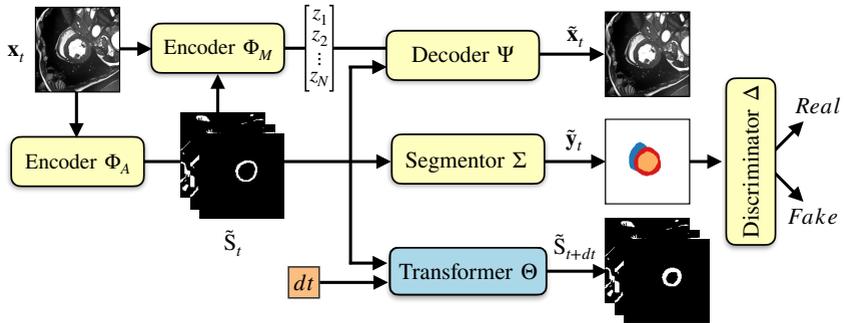


Figure 14: Block diagram of SDNet (described in Section 4.3.1) and SDTNet (Section 4.3.2). The components of SDNet are represented using yellow boxes, while SDTNet also includes the transformer network Θ , represented in light blue. In SDTNet, we train Θ to predict the future modality-independent anatomical factors conditioned on the temporal information dt . Notice that improving the quality of the anatomical representation S_t can make the segmentor job easier, facilitating the extraction of high-quality segmentation masks \tilde{y}_t .

4.3 Methods

4.3.1 Spatial Decomposition Network (SDNet)

Many medical imaging modalities contain spatial information about the patient’s anatomy modulated by modality-specific characteristics. The SDNet (Chartsias, Joyce, et al., 2019) decouples anatomical factors from their appearance, obtaining: i) improved performance with limited annotations compared to other supervised approaches, and ii) more interpretable representations. Below, we briefly review SDNet, upon which we build our model.

Model

Overall, we can interpret SDNet as an autoencoder that receives a 2-dimensional image x as input and decomposes it into disjoint anatomical components S and modality-dependent factors \underline{z} (we present the block diagram of SDNet as the yellow boxes in Figure 14). The general idea

is that jointly within these two representations, all of the available information is captured, and thus it should be possible to (perfectly) reconstruct the input image. This is a self-supervised task. Moreover, using only the anatomical information should be enough to perform (supervised) tasks that only need information about image anatomy, such as semantic segmentation. With this decomposition of an image \mathbf{x} in a tuple of modality and anatomical factors (\underline{z}, S) , we can train SDNet to minimise supervised, self-supervised, and adversarial objectives, resulting in a multi-task learning problem.

Supervised Objective With the goal of performing semantic segmentation, we train a segmentor $\Sigma(\cdot)$ to extract label maps $\tilde{\mathbf{y}}$ from the anatomical representation of the image, such that the prediction $\tilde{\mathbf{y}} = \Sigma(S)$ approximates the available ground-truth mask \mathbf{y} .

Unsupervised Objective Conditioned on the anatomical representation S , a modality encoder $\Phi_M(\mathbf{x})$ learns to map the image \mathbf{x} to factors \underline{z} which are image modality-dependent. In particular, we encourage \underline{z} to follow a multivariate Gaussian distribution, as in the VAE framework (Kingma and Welling, 2014). A decoder $\Psi(\cdot)$ combines \underline{z} and S to reconstruct the input image $\tilde{\mathbf{x}} = \Psi(\underline{z}, S) \approx \mathbf{x}$, providing a loss signal used to improve both modality and anatomy factors based on the image reconstruction error.

Adversarial Objective Disentanglement is not trivial and requires inductive biases (Locatello et al., 2019; Locatello et al., 2020; Liu*, Thumos*, et al., 2021). One strong bias is data, but expending annotations to provide such bias is perhaps conflicting to semi-supervised learning. One possibility is to provide shape priors that can help constrain the obtained segmentations to be close to reality. An adversarial loss (Goodfellow et al., 2014) encourages the predicted segmentations $\tilde{\mathbf{y}}$ to be realistic even when no manual annotation is available for the input image. Such a training signal is provided by a mask discriminator $\Delta(\cdot)$ that learns to say apart real from predicted segmentation masks.

Encouraging Disentanglement SDNet uses two specific biases to disentangle anatomical from modality factors of an image. In particular, SDNet models the anatomical factors S as discrete multi-channel binary maps, and the modality factors z as continuous variables which affect image appearance at a global level. S is obtained using a channel-wise softmax activation function to force each pixel in the multi-channel output of the anatomy encoder to have activations that sum up to one. During a forward pass, this multi-channel output is thresholded as $S \mapsto \lfloor S + 0.5 \rfloor$, while the training gradients are simply propagated through the thresholding operation during backpropagation, as in the straight-through operator (Bengio, Léonard, and Courville, 2013). Since S is binarized, it cannot easily encode continuous modality-dependent characteristics. As a result, SDNet must encode any modality information in z . To ensure that z does not also encode anatomical information, SDNet uses a very restrictive model to extract z , obtained through an information bottleneck (Kingma and Welling, 2014).

Limitations of SDNet and Proposed Approach

Compared to classical supervised approaches, SDNet obtains better segmentation performance in a scarcity of annotations. However, its ability to segment strictly relies on the extracted anatomical representations, and improving the anatomical factors makes the task easier for the segmentor. In this chapter, we introduce a spatio-temporal prior in SDNet, encouraging the model to learn the temporal dynamics of the anatomies. In particular, we use the temporal information that is intrinsically available in cardiac cine MRIs, and train SDNet to foresee future instants of the anatomical factors in a cardiac cycle. By encouraging the extraction of temporally correlated factors, we impose the association of similar images with similar representations, with beneficial effects on the subsequent segmentation task.

We aim to build upon SDNet limitations based on a simple hypothesis: small transformation in the input domain x should be associated with small changes in the anatomical representation. Specifically, the S factors of different cardiac phases should be similar within the same

cardiac cycle, while any difference should be consistent across subjects. Moreover, anatomical components that move together in time, such as the heart, should be separated from static ones.

We introduce such regularisation via an additional neural network in the SDNet framework which we term as ‘the transformer’ $\Theta(\cdot)$, whose role is to learn the temporal dynamics of the anatomical factors during a cardiac cycle. In particular, the transformer encourages S to have smooth and consistent temporal transformations, which acts as a regulariser on the disentangled representation.

We provide the block diagram of the extended model, which we name Spatial Decomposition and Transformation Network (SDTNet), in Figure 14. The anatomy and modality encoders, the decoder, the segmentor and discriminator architectures follow those proposed by Chartsias, Joyce, et al., 2019 in the original framework. We analyse the SDTNet and the transformer in more details below.

4.3.2 Spatial Decomposition and Transformation Network (SDTNet)

As illustrated in Figure 14, the transformer receives as inputs the anatomical factors S_t at the current time point t , and the information about a temporal gap dt . Assuming that the image appearance \underline{z} is constant throughout the cardiac cycle – the imaging modality does not change – and that the only variations regard the patient anatomy, in SDTNet the transformer learns to deform the input S_t and predict the anatomical transformation.

In the following, we first describe the global framework and the training objectives. Then, we describe the transformer architecture and the optimisation strategy.

Cost Function and Training

We optimise SDTNet using a multi-task learning formulation, where the semi-supervised training objective is the sum:

$$\mathcal{L}_{oss} = a_0 \cdot \mathcal{L}_S + a_1 \cdot \mathcal{L}_{US} + a_2 \cdot \mathcal{L}_{ADV} + a_3 \cdot \mathcal{L}_{TR}, \quad (4.1)$$

where we use the scaling parameters $a_0 = 10$, $a_1 = 1$ and $a_2 = 10$ as in Chartsias, Joyce, et al., [2019], and a_3 determined experimentally. In the specific, \mathcal{L}_S is the supervised segmentation cost. \mathcal{L}_{US} is an unsupervised objective containing an image reconstruction term and a regulariser on the modality representation \underline{z} . \mathcal{L}_{ADV} is an adversarial cost obtained through a mask discriminator. \mathcal{L}_{TR} is the cost associated with the training of the temporal transformer network.

As discussed by Chartsias, Joyce, et al., [2019], separating the anatomy into segmentation masks is challenging because the image reconstruction process encourages parts having similar colour intensities to appear in the same channels. For this reason, it is crucial to give more importance to the segmentation losses and use higher values for a_0 and a_2 . We chose instead $a_3 = 0.8$ to scale \mathcal{L}_{TR} approximately to the same amplitude of $a_1 \cdot \mathcal{L}_{US}$ and thus obtain similar unsupervised contributions.

Supervised Objective \mathcal{L}_S is the cost associated to the segmentation task, when labels are provided. It consists in the differentiable Dice loss (Milletari, Navab, and Ahmadi, [2016]), defined as: $\mathcal{L}_S = 1 - \frac{2|\tilde{\mathbf{y}} \cdot \mathbf{y}|}{|\tilde{\mathbf{y}}| + |\mathbf{y}|}$, where \mathbf{y} is the ground-truth segmentation and the $\tilde{\mathbf{y}}$ is the one predicted by the segmentor. We evaluate \mathcal{L}_S as the average Dice loss obtained on every region to segment.

Unsupervised Objective The cost associated to the unsupervised task \mathcal{L}_{US} is the sum of contributions:

$$\mathcal{L}_{US} = |\mathbf{x} - \tilde{\mathbf{x}}| + a_{KL} \cdot D_{KL}[Q(\underline{z} | \mathbf{x}) \| N(\underline{0}, \underline{I})] - MI(\underline{z}, \tilde{\mathbf{x}}).$$

The first term in the formula is the mean absolute error between the input image \mathbf{x} and its reconstruction $\tilde{\mathbf{x}}$, where $\tilde{\mathbf{x}}$ is the output of the decoder $\Psi(\cdot)$. The second term $D_{KL}[\cdot]$ is the KL divergence between the distribution of the latent representation extracted by the modality encoder, $Q(\underline{z} | \mathbf{x})$, and a Multivariate Gaussian $N(\underline{0}, \underline{I})$ with zero mean and identity covariance matrix. As in (Chartsias, Joyce, et al., [2019]), we use $a_{KL} = 0.1$. Finally, the last term $MI(\cdot)$ is the mutual information between the latent code \underline{z} and the reconstruction $\tilde{\mathbf{x}}$ and it is approximated using

an additional neural network, as suggested by Chen, Duan, et al., [2016]. Maximising the mutual information term helps to build a meaningful latent space for \underline{z} and prevent the posterior collapse, thus encouraging the decoder $\Psi(\cdot)$ to use the modality factors.

Adversarial Objective We use a Least-Squares mask discriminator (Mao et al., [2018]) to introduce an adversarial term \mathcal{L}_{ADV} in the loss function. We use unpaired data to train the discriminator so that it can distinguish the ground-truth segmentation masks from those that are predicted by the segmentor. The adversarial term introduces a shape prior-based contribution in the model, which encourages the segmentor to output plausible segmentation masks even for unlabelled images.

Self-supervised Consistency Objective The term \mathcal{L}_{TR} is the self-supervised cost provided by an anatomy transformer $\Theta(\cdot)$. As previously discussed, we use \mathcal{L}_{TR} to regularise the anatomical factors through spatio-temporal constraints. During training, the transformer gradually learns to change the input tensor S_t such that it can match S_{t+dt} . Since the anatomical space is binary, we propose to train the transformer using the Dice loss between predicted and future binary anatomical factors:

$$\mathcal{L}_{TR} = 1 - \frac{2|\tilde{S}_{t_0+dt} \cdot S_{t_0+dt}|}{|\tilde{S}_{t_0+dt}| + |S_{t_0+dt}|}.$$

It is possible to give a distance-based interpretation to \mathcal{L}_{TR} . In particular, minimising \mathcal{L}_{TR} is a form of contrastive loss minimisation (Hadsell, Chopra, and LeCun, [2006]), where the learned representation favours small distances between pairs of *similar* examples and large distances for *dissimilar* pairs. Analogously, we constrain the representation of temporally close cardiac phases to be encoded closer (in terms of Dice distance) through the consistency objective \mathcal{L}_{TR} . However, to avoid the collapse of the *attract-only* force introduced by \mathcal{L}_{TR} , we use the weighted sum of \mathcal{L}_S , \mathcal{L}_{US} and \mathcal{L}_{ADV} as a *repulse-only* force for modelling dissimilar points.

After describing the optimisation strategy, we detail the transformer architecture in more details below.

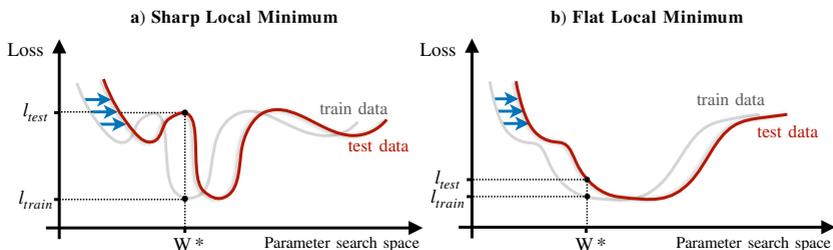


Figure 15: Effect of dataset shift on a model residing in a sharp or a flat local minimum. We plot the loss landscapes of train data in grey colour and test data in red colour. Given an optimised model with parameters W^* , the same dataset shift from train to test data (blue horizontal arrows) has an increased performance impact if the model resides in a sharp minimum (dashed vertical lines).

Optimisation strategy Training a neural network to perform many different tasks at the same time while using limited annotations is a challenging problem. In fact, the loss landscape can be noisy, and the training becomes unstable. On top of that, learning can be subject to additional noise whenever we must necessarily use small batch sizes, e.g. = 4, to cope with memory constraints. Thus, models that perform well on the training set may perform poorly on the validation and test samples.

We reduce this problem optimising the model toward solutions residing in flat local minima of the training loss landscape, which can generalise better (a concept depicted in Figure 15). In particular, we make this possible using two approaches: the Exponential Moving Average (EMA) and adopting a cyclical learning rate scheduling (Smith, 2017).

EMA consists of maintaining a moving average of the trained parameters and in using the averaged ones for inference. During the test, this is equivalent to perform an ensembling of the model in correspondence of the last iterations.

A cyclical learning rate scheduling, instead, consists of a periodic ranging of the learning rate between a minimum and a maximum values. When the learning rate has a high value, it can help the model escaping sharp local minima. On the contrary, when it has a smaller amplitude,

it helps to settle in the bottom of flat loss valleys. In our case, we used a triangular wave linearly ranging between 10^{-4} and 10^{-5} within a period of 20 epochs. We chose the cycle length according to the guidelines provided in (Smith, 2017).

Both EMA and the learning rate scheduling considerably facilitate comparisons with the baselines, by reducing loss fluctuations at the end of the training and thus reducing the effect of the chosen early stopping criterion. We used Adam optimiser (Kingma and Ba, 2015) and stopped training based on the segmentation loss on a validation set, as in (Chartias, Joyce, et al., 2019).

Transformer Architecture

The transformer is a modified UNet (Ronneberger, Fischer, and Brox, 2015), adapted to work in the binary anatomical space. We also include a long residual connection between the UNet input and its output, which allows initialising the transformer to operate an identity mapping (plus noise, as we randomly initialise the network weights). To ensure that the output of the transformer \tilde{S}_{t+dt} resides in the binary anatomical space, we process it with a softmax operator and then binarise it again with the thresholding operator $\tilde{S}_{t+dt} \mapsto \lfloor \tilde{S}_{t+dt} + 0.5 \rfloor$.

We introduce the temporal information in the transformer bottleneck through a conditioning mechanism which modulates the extracted features maps to operate the temporal transformation. We use a scalar value of dt to represent the time-gap between the current cardiac phase and the time frame we want to predict. The value dt is the input to an MLP with three fully connected layers having 128, 128 and 7744 units, respectively. The MLP prediction is first reshaped to a dimension of $22 \times 22 \times 16$ and then concatenated with the anatomical features maps extracted by the contracting encoder of the UNet. To encourage the use of the temporal features maps we bounded both the MLP output and the anatomical features in the range $[0, 1]$ using a sigmoid activation function, resulting in signals of comparable amplitude.

4.4 Experiments

4.4.1 Data

For the experiments, we used the cardiac datasets: ACDC (described in Section 3.8.1) and LVSC (Section 3.8.2). Both datasets contain 2-dimensional cine-MR images acquired from a variety of 1.5T and 3T MR scanners and imaging parameters. Images cover the whole heart in a short-axis view and have different temporal resolutions, ranging between 19 and 40 frames for the patient’s cardiac cycle.

For both datasets, we performed the experiments using 3-fold cross-validation. We randomly divided the total number of MRI scans to use 70% of patients for training, 15% for validation and 15% for the test sets.

4.4.2 Temporal Axis

Since our goal is to introduce temporal smoothness in the anatomical factors rather than learning to predict the whole cardiac cycle, we split the cine sequences into two halves: frames in the ED-ES interval and frames from ES to the end of the cardiac cycle. Then, we reversed the latter frames in their temporal order, to mimic once again the cardiac contraction. As a result, we could avoid dealing with the inherent uncertainty of the temporal instants in the middle of the cardiac cycle, where predicting if the heart will contract or dilate in the next frame is not possible by relying only on the current image.

To account for a temporal resolution changing across patients, we normalized the frame indexes on the total number of frames in the sequence. Thus, temporal distances between two consecutive frames were always considered relative to the whole contraction time and $t \in [0, 1]$.

4.4.3 Baselines and Evaluation

We compare our model with the fully supervised training of a UNet (Ronneberger, Fischer, and Brox, 2015) and the semi-supervised training of SDNet (Chartsias, Joyce, et al., 2019). We analyse the performance obtained using different fractions of annotations in the training set. In these

ACDC - Dice Score

Labels	UNet				SDNet				SDTNet (ours)			
	RV	MYO	LV	Average	RV	MYO	LV	Average	RV	MYO	LV	Average
100 %	81.5 ₀₅	84.5 ₀₃	89.2 ₀₄	85.0₀₄	78.4 ₀₆	83.5 ₀₃	89.2 ₀₄	83.7 ₀₄	77.8 ₀₆	83.7 ₀₃	88.1 ₀₄	83.2 ₀₄
25 %	76.3 ₀₆	83.5 ₀₃	87.2 ₀₅	82.3 ₀₅	73.6 ₀₇	79.7 ₀₄	86.4 ₀₅	79.9 ₀₅	77.3 ₀₆	84.5 ₀₃	87.5 ₀₄	83.1₀₄
12 %	66.8 ₀₇	76.9 ₀₄	82.4 ₀₆	75.3 ₀₆	68.8 ₀₇	79.0 ₀₄	83.5 ₀₅	77.1 ₀₅	67.8 ₀₈	82.1 ₀₃	86.0 ₀₄	78.6₀₅
6 %	46.5 ₀₈	61.1 ₀₆	73.2 ₀₈	60.3 ₀₇	54.3 ₀₈	66.2 ₀₅	75.3 ₀₇	65.3 ₀₇	52.6 ₀₉	70.6 ₀₅	76.0 ₀₇	66.4₀₇
3 %	34.6 ₀₈	46.6 ₀₇	56.9 ₀₉	46.0 ₀₈	42.0 ₀₉	60.1 ₀₆	71.8 ₀₇	57.9 ₀₇	45.0 ₀₉	60.3 ₀₇	69.1 ₀₇	58.1₀₈

Table 2: Dice Score average and standard deviation (subscript) for the segmentation of myocardium (MYO), left ventricle (LV) and right ventricle (RV), on ACDC dataset. We compare models at various proportions of training annotations. Results are the average of three-fold cross-validation. Best results in **bold**.

experiments, if a model can use the extra unlabelled images, we employ them for optimising the unsupervised, adversarial or self-supervised objectives.

We measure performance using Dice Score and Hausdorff Distance between predicted \tilde{y} and ground-truth segmentation masks y , for each cardiac structure.

4.5 Results and Discussion

In the following, we first analyse the advantages of introducing temporal consistency in the learned disentangled representations (Section 4.5.1). Then, we investigate the anatomical factors extracted by our model and how the Transformer modifies them, subject to the temporal signals (Section 4.5.2).

4.5.1 Semi-supervised segmentation

We compare our method (SDTNet) with the baselines qualitatively in Figure 16 and quantitatively in Table 2, 3 and 4. We provide below an analysis of the results using questions before each paragraph to help guide the reader.

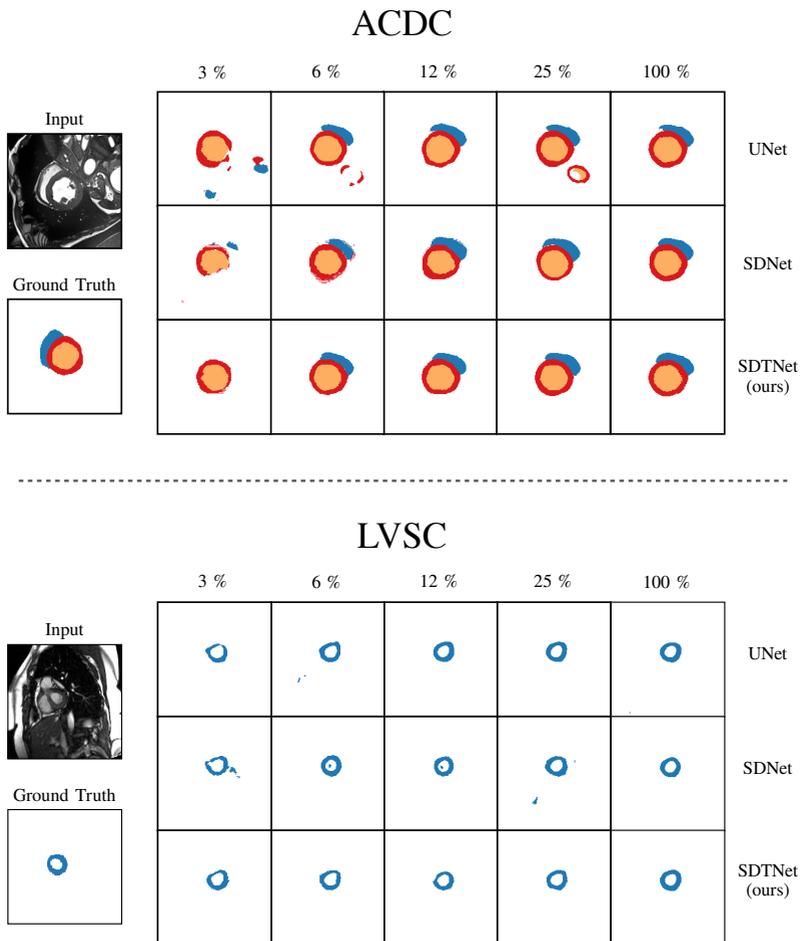


Figure 16: Segmentation masks predicted by the considered models at various levels of training annotations on ACDC (top) and LVSC (bottom) datasets. As can be seen, using temporal consistency to regularise disentanglement (SDTNet) leads to the best performance, especially when annotations are scarce.

ACDC - Hausdorff Distance

Labels	UNet				SDNet				SDTNet (ours)			
	RV	MYO	LV	Average	RV	MYO	LV	Average	RV	MYO	LV	Average
100 %	10.5 ₀₆	6.9 ₀₃	4.7 ₀₂	7.4 ₀₇	14.5 ₁₀	5.0 ₀₂	3.3 ₀₁	7.6 ₀₇	13.6 ₀₉	5.4 ₀₂	9.8 ₀₂	9.6 ₀₄
25 %	11.7 ₀₈	13.1 ₀₆	6.9 ₀₄	10.6 ₀₆	20.5 ₁₂	12.0 ₀₄	9.6 ₀₄	14.0 ₀₇	14.5 ₁₁	4.7 ₀₁	3.6 ₀₁	7.6 ₀₄
12 %	27.8 ₁₀	31.0 ₀₈	17.3 ₀₈	25.4 ₀₉	22.8 ₁₁	16.3 ₀₅	11.6 ₀₇	16.9 ₀₈	25.6 ₁₀	8.4 ₀₄	11.1 ₀₇	15.0 ₀₇
6 %	69.8 ₁₁	49.0 ₀₈	34.6 ₁₂	51.1 ₁₀	42.5 ₀₉	50.7 ₀₇	41.4 ₀₈	44.9 ₀₈	47.3 ₁₅	32.0 ₀₈	43.1 ₀₉	40.8 ₁₁
3 %	76.7 ₁₀	66.8 ₀₇	59.7 ₁₀	67.7 ₀₉	58.6 ₁₄	45.6 ₀₈	35.3 ₁₁	46.5 ₁₁	51.1 ₁₄	37.2 ₀₉	35.2 ₁₂	41.2 ₃₄

Table 3: Hausdorff Distance average and standard deviation (subscript) for the segmentation of myocardium (MYO), left ventricle (LV) and right ventricle (RV), on ACDC dataset. We compare models at various proportions of training annotations. Results are the average of three-fold cross-validation. Best results in **bold**.

Does using temporal information help? As can be seen from the tables, learning temporal dynamics regularises the training of SDTNet, especially when dealing with a limited number of annotations. Our model improves the performance of SDNet for almost every percentage of available annotations both in ACDC and LVSC datasets, increasing the Dice Score and decreasing the Hausdorff Distance. Moreover, we observe that the fully-supervised UNet has a consistent performance deterioration when the number of annotations reduces below 25% of the training set. In LVSC, the UNet is always the worst model, while in ACDC it performs worse than both the disentanglement frameworks when using less than 100% of annotation. This behaviour can be justified observing that when the number of annotated data decreases, the training set is not sufficiently representative of the data distribution and methods that *only* rely on supervision fail. On the other hand, using the unlabelled data, both SDNet and SDTNet learn more robust representations and perform well even with fewer annotations. These results show the advantage of disentangled representations and temporal priors in the absence of enough labels, which is important when dealing with rare pathologies or anatomical variants.

What happens when we have lots of annotations? When using all of the available annotations, SDNet, SDTNet and the fully supervised UNet

perform similarly on LVSC data. Instead, on ACDC, the UNet performs best. These observations are coherent with recent findings (Chartsias, Joyce, et al., [2019]; Liu*, Thermos*, et al., [2021]) reporting that disentanglement is most effective when there are not strong supervisory signals. When the number of annotations increases, the UNet is simple to optimise because it does not require the simultaneous minimisation of multiple objectives, nor to find a trade-off between supervised and unsupervised/adversarial losses. This points to the need to have dynamic mechanisms to balance different costs' contribution in a disentanglement framework. It is possible to close the performance gap between UNet, SDNet and SDTNet in ACDC, by using a higher weight on a_0 to give more importance to the supervised cost in Equation [4.1].

Can we double up self-supervision by leveraging the cardiac cycle?

Inspired by Wang, Jabri, and Efros, [2019] we also experimented introducing *cycle consistency* across the cardiac cycle for learning visual correspondences. In other words, given the prediction of the transformer $\tilde{S}_{t+dt} = \Theta(S_t, dt)$, we trained the model to learn to go back in time and estimate: $\tilde{S}_t = \Theta(\tilde{S}_{t+dt}, -dt)$ where $\tilde{S}_t \approx S_t$. However, the cycle consistency did not improve segmentation, and in some cases the transformer even collapsed, predicting constant anatomical channels for any time point. In the collapsed transformer, the output $\tilde{S}_{t\pm dt}$ corresponded to an average cardiac phase $S_{\bar{t}}$, rather than changing according to time gaps. We hypothesise that this behaviour originates because the cycle consistency adds additional constraints and makes predicting future representations harder. Since we use a small scaling factor to multiply the transformer loss, the transformer only has a small incentive to learn the required task, and it collapses to output average predictions because they don't increase the global loss (Eq. [4.1]) too much (a form of underfitting).

4.5.2 What does the model learn?

How do anatomy factors look? As in Chartsias, Joyce, et al., [2019] we use 8 anatomical channels to represent the patient anatomy. We show an

LVSC - Dice Score				LVSC - Hausdorff Distance			
Labels	UNet	SDNet	SDTNet (ours)	Labels	UNet	SDNet	SDTNet (ours)
100%	68.3 ₀₇	69.8₀₇	69.6 ₀₇	100%	22.7 ₁₀	12.2₀₄	15.9 ₀₈
25%	66.1 ₀₉	67.0 ₀₉	67.3₀₈	25%	22.9 ₁₁	12.1₀₅	12.1₀₅
12%	58.8 ₁₂	65.0 ₁₂	66.1₁₀	12%	33.2 ₀₉	28.3 ₁₃	18.8₀₉
6%	49.0 ₁₅	53.1 ₁₄	54.5₁₅	6%	53.2 ₀₉	42.9₁₂	51.8 ₁₁
3%	34.7 ₁₆	46.1 ₁₃	48.6₁₄	3%	64.2 ₁₂	48.7 ₀₈	36.3₁₀

Table 4: Average and standard deviation (subscript) performance for myocardium segmentation, on LVSC dataset. We report Dice Score on the left table, Hausdorff Distance on the right table. We compare models at various proportions of training annotations, reporting the average of three-fold cross-validation. Best results in **bold**.

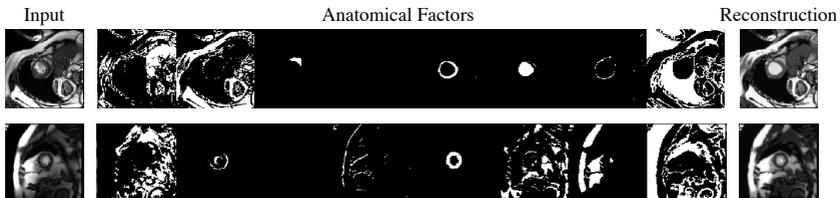


Figure 17: Example of anatomical factors extracted from an input image for ACDC (top row) and LVSC (bottom row) datasets. Anatomies are represented as multi-channel binary maps and can contain well defined anatomical components, such as left/right ventricle and myocardium, or other geometrical content needed for the image reconstruct through the image decoder (rightmost image).

example of the multi-channel anatomical factors learned by our model in Figure 17. Some of the binary channels contain well defined anatomical parts, such as the cardiac structures, and others the remaining spatial information, which is necessary to reconstruct the input image through the decoder.

How do predicted images look in time? Observe that, if we assume that the modality factor z remains constant throughout the whole cardiac cycle, given an image at $t = 0$ it is possible to predict future frames of the temporal sequence using the following steps: i) extract S_0 and z_0 from

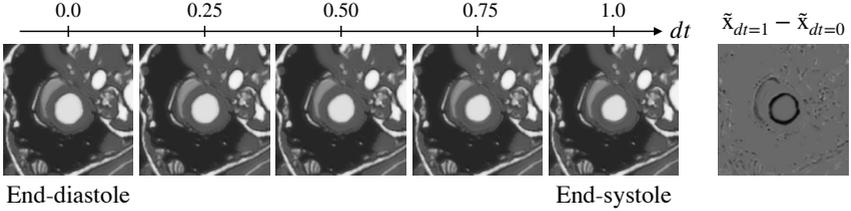


Figure 18: Example of temporal interpolation from ED to ES cardiac phases. Images were obtained by keeping fixed the anatomical factors S_t at time $t = 0$ and ranging dt in $[0, 1]$.

the input image x_0 ; ii) keep z_0 fixed and predict the future frames with the transformer, as $\tilde{S}_{dt} = \Theta(S_0, dt)$ while changing dt in the range $[0, 1]$; iii) use the decoder to reconstruct the future frame $\tilde{x}_{t>0} = G(\tilde{S}_{dt}, z_0)$. We report an example of the procedure in Figure 18. As can be seen from the figure, the model can predict the cardiac contraction. Interestingly, the image colours appear flat, whose reason can be found in the design of the decoder $\Psi(\cdot)$. In fact, as (Chartsias, Joyce, et al., 2019), we use a decoder architecture based on FiLM (Perez et al., 2018), which reintroduces the modality-specific “colours” into the anatomical channels by simply scaling and multiplying the whole binary maps. Because of its design, FiLM cannot introduce texture-related information, and the image reconstruction shows flat colours. An alternative to FiLM layers is using a decoder based on SPADE (Park, Liu, et al., 2019) which is less restrictive and it allows to reproduce textures in the reconstructed image, rather than just intensity values. Contrarily to FiLM, SPADE uses the anatomical channels to modulate the modality-dependent features maps, which can also contain textures, and the reconstructed images become more realistic (Chartsias, Papanastasiou, et al., 2020). An in-depth comparison in terms of disentanglement and segmentation performance between SPADE and FiLM-based decoders can be found in Liu*, Thermos*, et al., 2021.

What does the transformer learn? In Figure 19, we show images of features maps extracted at the transformer bottleneck. In particular, we

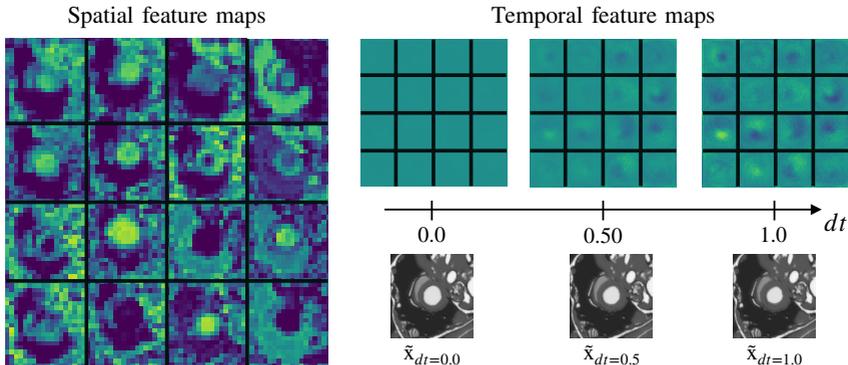


Figure 19: Features maps in the transformer bottleneck. On the left, we show 16 out of the 64 features maps extracted by the anatomical representations S_t . On the right, we show the features maps predicted by the MLP (top row) when ranging the value of dt from 0 (ED cardiac phase) to 1 (ES cardiac phase). Colour maps linearly range from 0 (dark blue) to 1 (yellow).

show examples of features extracted by the anatomical channels (left), and the 16 temporal features maps predicted by the MLP used to condition the transformer (right) when it receives as input dt in $\{0.0, 0.5, 1.0\}$. As can be seen, the MLP outputs globally larger signals in correspondence of the complete cardiac contraction, that is when we go from ED to ES cardiac phase.

4.6 Conclusion

This chapter discussed how disentangled representations aid semi-supervised learning by decomposing a medical image into anatomical and imaging modality-specific factors. The presence of a reconstruction cost and a segmentation loss render disentangled representations suitable for semi-supervised learning by taking advantage of the semantic information residing in the image content. More broadly, disentanglement allows intuitive factorisation of the image into spatial and non-spatial factors. Such a factorisation increases model interpretability, which is a key advantage in healthcare. Furthermore, it allows intuitive image manip-

ulations by combining factors across patients and modalities (Chartsias, Joyce, et al., 2019), it is well suited for multi-modal learning (Yang et al., 2019; Chartsias, Papanastasiou, et al., 2020), and has considerable potential to automatically detect artefacts and pathologies (Jiang et al., 2020; Xia, Chartsias, and Tsaftaris, 2020; Liao et al., 2019).

In this chapter, we built on a recent disentanglement framework that produces interpretable representations (Chartsias, Joyce, et al., 2019). We presented a novel strategy to regularise the disentangled representation based on temporal transitions of the image components. We motivated and demonstrated that by conditioning the anatomical factors to undergo smooth temporal changes, it is possible to increase model performance on a post hoc task, such as semantic segmentation. We introduced the temporal information using a self-supervised objective and a transformer neural network, reporting increased performance in a lack of annotations. Lastly, we showed that the transformer model could potentially work for video prediction tasks and cardiac temporal synthesis.

In the future, it would be interesting to explore other forms of anatomical factor consistency, e.g. between adjacent slices on the third spatial dimension, rather than in time. Finally, it would be exciting to explore entirely unsupervised settings where no annotated images are needed to decouple an image’s anatomical components.

4.7 Summary

For disentanglement frameworks, such as SDNet and SDTNet, it is common to balance several unsupervised training objectives. As discussed in Chapter 3, these objectives may include features-level constraints and output-level constraints. In the models considered in this chapter, the former derives from imposing Gaussianity constraints on the modality-dependent factors, while we can divide the latter into self-reconstruction and adversarial losses.

We highlight that imposing constraints at the features level may considerably limit the model flexibility. For example, setting a Gaussian latent space in SDNet and SDTNet leads to an information bottleneck in

the modality encoder (Higgins, Matthey, et al., 2017), making the latent space most suitable to model Gaussian distributions. At the same time, it is sometimes better to represent the modality with a multi-modal – or even more complex – distribution (for an overview of methods attempting to model data using complex prior distributions, interested readers may refer to Bond-Taylor et al., 2021). Consequently, the self-reconstruction cost may not efficiently leverage the modality factors and lead the model to “hide” information inside the anatomical factors, making them worse.

On the other hand, constraining the model at the output level, i.e. encouraging the prediction of realistic segmentation masks, allows for more freedom in the latent space. Consequently, models may generally be more stable and easier to optimise.

SDNet and SDTNet penalise the segmentor at the output level via an adversarial cost, imposing a shape-prior on the predicted masks. There are several possible ways of introducing such form of prior knowledge. In the next chapter (Chapter 5), we will give a broader overview of the possible mask discriminators we can use. Then, we show that we can use adversarial shape priors to learn multi-scale consistent predictions in Chapter 6 and in Chapter 8 to improve performance under test-time data distribution shifts.

Chapter 5

Learning Adversarial Shape Priors with GANs

As discussed in the previous chapter, encouraging a segmentor to produce realistic masks in a lack of annotations is useful to regularise the model and improve test-time performance in semi-supervised learning. Moreover, contrary to direct approaches imposing constraints on the features space, output level constraints leave more flexibility to the model optimisation. As a result, models can more easily adapt to the training objectives and are more stable and easier to optimise in practice.

Output level regularisation can take different forms, including data-driven shape priors. Data-driven priors are advantageous because we can learn them from data without requiring the design of handcrafted loss functions for a specific problem. As a result, these priors are often part of simple frameworks, such as conditional GANs for semantic segmentation, and more complex methods, such as SDNet and SDTNet, discussed in the previous chapter.

In the following, we carry out an empirical analysis of adversarial techniques to learn data-driven shape priors for regularising semi-supervised learning of semantic segmentation. In Section 5.1, we briefly review recent work learning shape priors from unpaired masks and regularising the training of a segmentor in a lack of annotations. In

Section 5.2, we describe the unsupervised and semi-supervised training of GANs for semantic segmentation and advanced techniques to train GANs effectively. We also compare several GAN variants for semi-supervised learning. In Section 5.3, we present a novel approach for GAN regularisation, exploiting synthetic textures to train better mask discriminators. Lastly, in Section 5.4, we draw general considerations about GANs.

5.1 GANs and Adversarial Shape Priors

In supervised learning, a model learns a semantic segmentation task by minimising a cost function which is subject to the availability of input-output pairs. For example, provided a pair of data (x, y) , where x is the input image and y its binary ground-truth segmentation, and given a segmentor $\Sigma(\cdot)$, one can learn the correct mapping by minimising the pixel-wise cross-entropy loss: $\mathcal{L} = -y \log(\Sigma(x))$. Unfortunately, as discussed in Chapter 3, this approach is limited by the availability of annotated data, which can be hard to acquire. On the contrary, unlabelled data are usually more abundant, and we would like to use them to improve our models.

Recent literature on image segmentation (Yi, Walia, and Babyn, 2019; Cheplygina, de Bruijne, and Pluim, 2019) has reported an increasing interest in using GANs to learn data-driven losses and mitigate the need for supervised objectives. In particular, it is possible to replace the GAN generator with a segmentor $\Sigma(\cdot)$, and encourage the prediction of realistic segmentation masks when we don't have ground truth labels to evaluate the model. In these cases, an adversarial discriminator $\Delta(\cdot)$ learns to assess whether an input mask belongs to the distribution of hand-made annotations or is a generated mask, and penalises the segmentor when its predictions are not realistic.

In the classical formulation (Goodfellow et al., 2014), GANs are trained using a classification cost, where we assign a label "1" to real images, and "0" to fake, or predicted, images. For semantic segmentations, we can model the adversarial game between a segmentor $\Sigma(\cdot)$ and

a mask discriminator $\Delta(\cdot)$ with the training objective:

$$\min_{\Sigma} \max_{\Delta} V(\Delta, \Sigma) = \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} [\log \Delta(\mathbf{y})] + \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} [\log(1 - \Delta(\Sigma(\mathbf{x})))] \quad (5.1)$$

where \mathbf{x} is an input image for the segmentor Σ , $\Sigma(\mathbf{x})$ is the predicted mask, and \mathbf{y} is a real (unpaired) mask.

In practice, training a GAN with Equation 5.1 often does not provide sufficient gradient for $\Sigma(\cdot)$ to learn well. This is common when, in the early stages of training, the segmentor performs poorly and the discriminator can reject its predictions with a high confidence, because they are clearly different from the training data. In this case, the term $\log(1 - \Delta(\Sigma(\mathbf{x})))$ saturates and the gradients go to zero.

Rather than training the discriminator to minimise $\log(1 - \Delta(\Sigma(\mathbf{x})))$, Goodfellow et al., 2014 proposed to train it to minimise $-\log(\Sigma(\mathbf{x}))$, which can provide much stronger gradients early in learning. This formulation of the adversarial setup is usually named Non-saturating GAN (NSGAN), and it improves the adversarial training. However, training GANs is challenging also because of other problems, such as training instability and mode collapse.

To address these problems, many authors suggested to use different variants of Equation 5.1. The most common formulations are those of the Least-square GAN (Mao, Li, et al., 2017; Mao et al., 2018) and the Wasserstein GAN (Arjovsky, Chintala, and Bottou, 2017), which we describe below.

Instead of optimising a classification cost, the Least-square GAN (LSGAN) proposes to minimise the Pearson divergence between real and fake data distributions, using the objective:

$$\begin{aligned} \min_{\Delta} \mathcal{V}_{LS}(\Delta) &= \frac{1}{2} \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} [(\Delta(\mathbf{y}) - b)^2] + \frac{1}{2} \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} [(\Delta(\Sigma(\mathbf{x})) - a)^2] \\ \min_{\Sigma} \mathcal{V}_{LS}(\Sigma) &= \frac{1}{2} \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} [(\Delta(\Sigma(\mathbf{x})) - c)^2], \end{aligned} \quad (5.2)$$

with the advantage of preventing saturating gradients for any value predicted by the discriminator. Such a formulation also has the advantage of penalising more the generated masks that fall far away from the de-

cision boundary, while less the closest ones. A common choice for the parameters in Equation 5.2 is using $a = -1$, $b = 1$ and $c = 1$ or $c = 0$.

As an alternative approach, a Wasserstein GAN (WGAN) makes the model convergence easier by minimising the Wasserstein distance (Arjovsky, Chintala, and Bottou, 2017; Gulrajani et al., 2017) between real and fake data distributions, obtained through the learning objective:

$$\begin{aligned} \min_{\Delta} \mathcal{V}_{LS}(\Delta) &= - \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} [(\Delta(\mathbf{y}))] + \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} [(\Delta(\Sigma(\mathbf{x}))) \\ \min_{\Sigma} \mathcal{V}_{LS}(\Sigma) &= - \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} [(\Delta(\Sigma(\mathbf{x})))]. \end{aligned} \tag{5.3}$$

Similar to LSGANs, the WGANs prevent gradient saturation and offer a training signal for every prediction of the discriminator. An important assumption of WGANs is that the function approximated by the discriminator satisfies 1-Lipschitz constraints, i.e. it is a smooth function, which is enforced during the training of $\Delta(\cdot)$ by clipping its weights to a small absolute value. A softer alternative to weight clipping is using a regularising term to encourage a unit-norm gradient on all the data points generated by linear interpolation between real and generated samples (Gulrajani et al., 2017). The trained discriminator becomes piecewise linear around the data manifold, and provides higher quality training gradients to the generator. Enforcing 1-Lipschitz constraints in $\Delta(\cdot)$ proved to be a successful strategy in the broader population of GAN variants, for example improving the training of vanilla GANs and NSGANs (Fedus et al., 2017; Chu, Minami, and Fukumizu, 2020; Kodali et al., 2017).

5.1.1 Techniques to Improve GAN Training

There are several strategies to regularise GANs' training and reach better results. Among these, gradient penalty (Gulrajani et al., 2017) encourages similar data samples to be associated with close predictions of the discriminator. Similarly, spectral normalization (Miyato et al., 2018) constrains the discriminator to be Lipschitz continuous. Recently, Chu, Minami, and Fukumizu, 2020 highlighted that training powerful generators requires discriminators modelling smooth functions. For this reason,

regularisation techniques such as gradient penalty and spectral normalization – both of which indirectly encourage discriminator smoothness – proved to be successful.

There are other techniques to improve GANs, unrelated to the discriminator smoothness, but helping to prevent mode collapse and overfitting. Derived from reinforcement learning, experience replay (Lin, 1992) consists of presenting older generated images to the discriminator, to prevent forgetting (Schaul et al., 2016; Wu et al., 2018). Label smoothing (Szegedy, Vanhoucke, et al., 2016; Müller, Kornblith, and Hinton, 2019) limits overfitting by hindering the training of overconfident discriminators. Lastly, of great importance is the use of data augmentation on both real and generated data, which prevents the overfitting of both generator and discriminator (Karras, Aittala, et al., 2020; Zhao, Liu, et al., 2020). Among data augmentation techniques, we also mention instance noise (Sønderby et al., 2017), which improves one-sided label smoothing (Salimans et al., 2016) and limits the risk of discriminator overfitting by adding small perturbations on the real and generated images, making the data distribution denser.

5.1.2 Popular GAN Variants

Most recently, the literature has seen a plethora of proposed variants for GANs, which add up to the previously discussed NSGAN, LSGAN and WGAN. Having a well-trained discriminator is crucial for having high-quality training signals for the generator. Hence, recent work has mainly focused on changing the discriminator output, or designing it using more sophisticated architectures.

For example, EBGAN (Zhao, Mathieu, and LeCun, 2017) and BEGAN (Berthelot, Schumm, and Metz, 2017) use autoencoder-like discriminators, and minimise the divergence between the reconstruction losses obtained autoencoding real and generated images. Alternatively, both the ALI (Dumoulin, Belghazi, et al., 2017) and the BiGAN variants (Donahue, Krähenbühl, and Darrell, 2017) suggest using an additional encoder together with the discriminator. Such an encoder learns the data

distribution and limits the posterior collapse problem in the generator. Relativistic GANs (Jolicoeur-Martineau, 2019; Jolicoeur-Martineau, 2020) optimise generator and discriminator using batches of data having half real and half fake samples, and using the relative realness (and fakeness) of the images as training signal. The RealnessGAN (Xiangli et al., 2020) trains the discriminator to output a distribution as the measure of realness of the input images.

In the context of image-to-image translation, PatchGAN (Isola et al., 2017) only focuses on local properties of the generated images, penalising the generator at the scale of image patches. Motivated by the idea of “starting small”, multi-scale GANs (Denton, Chintala, Fergus, et al., 2015; Karras, Aila, et al., 2017; Luo, Zheng, et al., 2018) focus on low-resolution images first, and then gradually introduce higher-resolution and more complex data distributions. As a result, GANs first learn global aspects of an image and then start to focus on fine-grained image details. To model both local and global scales of the image, Schonfeld, Schiele, and Khoreva, 2020 have recently proposed to penalise the generator using a UNet-like discriminator (Ronneberger, Fischer, and Brox, 2015). In Chapter 6, we also introduce a GAN variant that interconnects generator and discriminator at multiple resolution levels, providing a multi-scale formulation of the adversarial game.

5.2 GANs for Unsupervised and Semi-supervised Image Segmentation

Despite GANs have been proposed in the context of unsupervised image generation, learning to segment images without supervision is challenging. In fact, without any constraint, the generator can learn an arbitrary mapping from the image space to the segmentation space. Thus, training GANs with purely unsupervised loss, such as those presented in equations 5.2 and 5.3, cannot guarantee that the predicted segmentation overlaps with the input image.

To encourage the learning of the right mapping, it is possible to introduce self-supervised objectives during training. In particular, we can im-

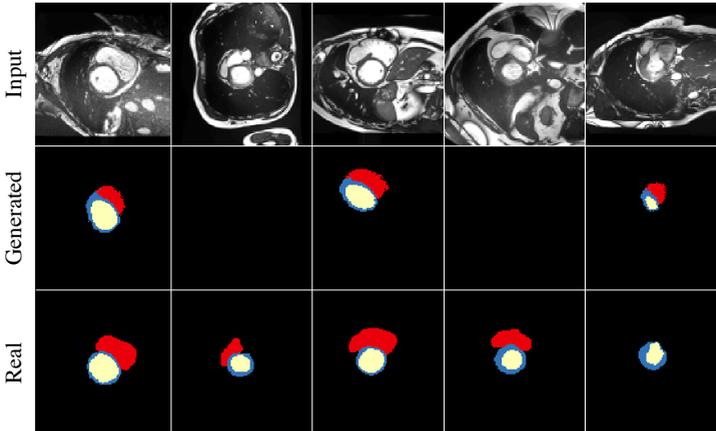


Figure 20: Example of segmentation masks generated by an unsupervised conditional GAN trained using images from the ACDC dataset (Section 3.8.1). The generator receives an input image and produces realistic segmentation masks. To prevent posterior collapse, we used a self-supervised consistency loss between the transformed images and their associated predicted segmentation. However, without using more stringent pixel-level constraints, there are no guarantees that the generated masks overlap with the regions of interest.

pose consistency constraints on the generator, to predict consistent outputs before and after applying known transformations to the images. As an example, we should expect that after translating or rotating an input image, the predicted mask is also translated, or rotated, accordingly. This self-supervised regularisation has the advantage to encourage the generator to learn a smooth representation manifold, increases the robustness to nuisance factors, and stabilises the adversarial training, preventing the mode collapse problem. However, it is still not enough to learn the correct mapping, as we show in Figure 20.

A different self-supervised objective which, instead, might have the potential to help to learn per-pixel correspondences between an image and the output segmentation mask, is a reconstruction cost. Image decoders can provide powerful image priors to learn the correct mapping. As we depict in Figure 21, adopting an auto-encoding strategy can en-

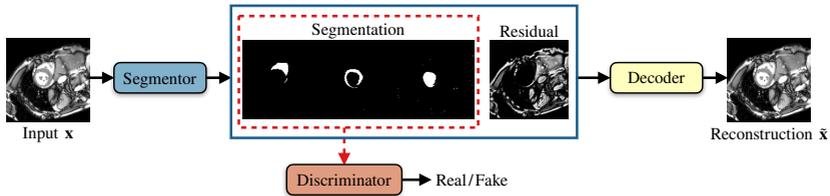


Figure 21: A reconstruction cost can help unsupervised GANs to generate realistic masks that also overlap with the anatomical regions. Given an input image x , the generator produces realistic segmentation masks and a residual representation using a *softmax* activation function. A mask discriminator encourages the predicted segmentations to look realistic, while a decoder combines them with the residual representations to reconstruct the input image and obtain $\hat{x} \approx x$.

courage the generator to produce realistic segmentations that align with the anatomical regions of interest. For example, we can use the generator to predict both the masks and a complementary residual representation. A decoder network can then combine the segmentation with its residuals and reconstruct the input image (making the generator role similar to that of an encoder). The key components to make the system work are: i) a mask discriminator that judges the predicted segmentation; and ii) a softmax activation function that allows obtaining complementary per-pixel information between the mask and the residuals. A very similar framework has been recently proposed in the context of cardiac segmentation by Joyce, Chartsias, and Tsaftaris, [2018].

Unfortunately, fully unsupervised approaches often exhibit unstable behaviour and are difficult to train or have a high variance. On the contrary, the most effective way to regularise GANs for semantic segmentation is to use supervisory signals at least on a subset of training images (Yi, Walia, and Babyn, [2019]). In these cases, the generator is usually trained in an alternate fashion, being optimised with an adversarial objective on a batch of unlabelled images at first, then using a supervised cost function on a batch of annotated images (Fig. 22).¹

¹This way of training the model maintains a 1:1 ratio between the supervised and the unsupervised training. More in general, it is possible to use a different balance between the

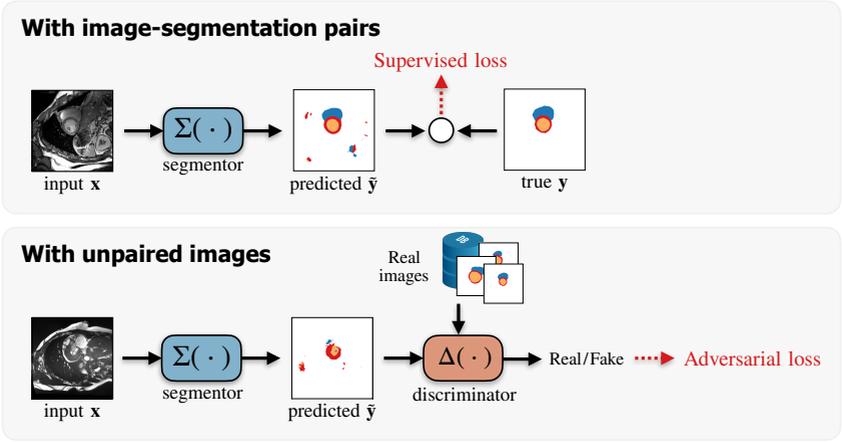


Figure 22: Training a segmentor $\Sigma(\cdot)$ with paired and unpaired data. When annotations are available, we optimise $\Sigma(\cdot)$ with a supervised cost function. For unlabelled images, we train $\Sigma(\cdot)$ using a data-driven adversarial loss, evaluating if the predicted mask belongs to the manifold of real segmentation masks.

In Figure 23 we report a comparison between a vanilla UNet segmentor and several GAN variants trained in a semi-supervised setting. We use the same UNet as segmentor and we keep the discriminator the same across GAN variants. We report a detailed description of the architectures and hyperparameters used for training in the Appendices A.1.1 and A.1.2.

As shown in the figure, GANs benefit from unlabelled data and generally increase performance when annotations are scarce. However, there are exceptions to this rule. When there are minimal annotation levels *and* limited training data, several GAN variants underperform and sometimes are even worse than a simple UNet segmentor. For example, we find that the UNet has competitive performance on the CHAOS dataset when we have only 5% of labels. CHAOS is a small dataset, and the ad-

optimisation steps on the labelled and unlabelled data. Similarly, it is not uncommon to find solutions optimising the discriminator more steps before updating the generator (Arjovsky, Chintala, and Bottou, 2017).

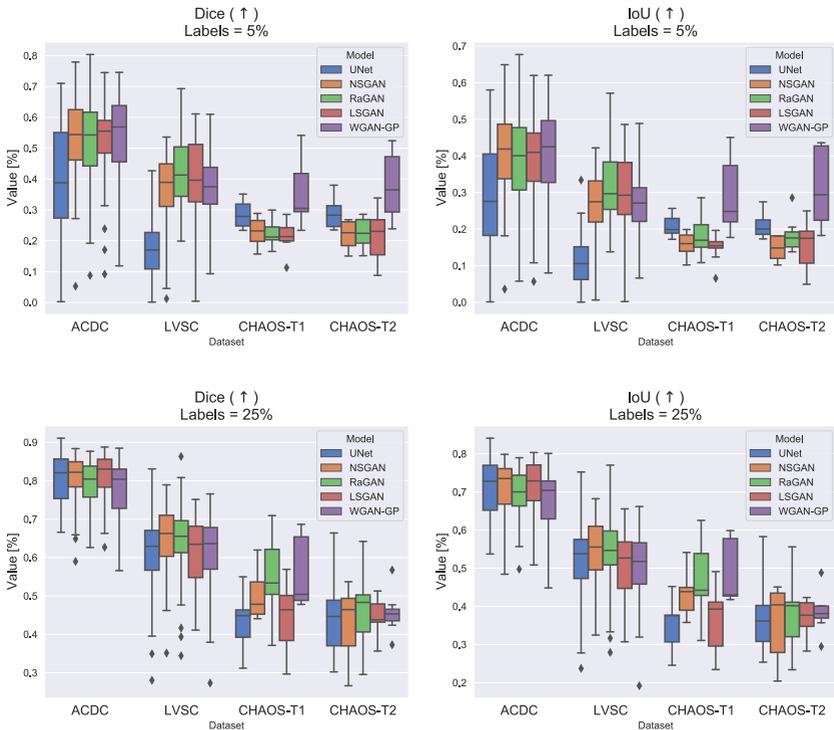


Figure 23: Performance of popular GAN variants in the task of semantic segmentation. Performance is measured in terms of Dice (\uparrow) and IoU (\uparrow) scores, where arrows show the metric improvement direction. The box plots report median and inter-quartile range (IQR) of each metric on a given test set, considering as outliers those values falling outside two times the IQR. We compare the performance of a standard segmentor (UNet) and the following GAN variants: Non-saturating GAN (NSGAN), Relativistic-average GAN (RaGAN), Least-square GAN (LSGAN), Wasserstein GAN with gradient penalty (WGAN-GP). We consider the following medical datasets: ACDC (Section [3.8.1](#)), LVSC (Section [3.8.2](#)), and CHAOS (T1 and T2 images, Section [3.8.4](#)). The top row shows Dice and IoU scores on the test set after training the segmentor using only 5% of labelled training samples. The bottom row reports the performance when using 25% of annotated training samples. We observe that GANs improve the segmentor training especially when training annotations are scarce.

versarial framework cannot fully exploit its potential of using unpaired images to train better segmentors. Moreover, 5% of annotated data consists of just one annotated patient in CHAOS (while in ACDC and LVSC, there are 4 and 3 patients, respectively). The annotations are too scarce, and training a GAN is challenging in CHAOS. Consequently, most GAN variants perform worse than a UNet. Nevertheless, using WGAN with gradient penalty leads to the best results on this dataset (also better than the UNet), proving that – if well optimised – GANs can still increase performance in these settings.

Instead, on ACDC and LVSC, we find that GANs are generally better: both at 5% and 25% of labels. GANs are also better than the simple UNet on CHAOS data at 25% of available annotations (4 annotated patients).

Finally, we observe that some GAN variants perform better than others on specific datasets, but none of them is the best always. These observations are coherent with a recent large-scale study analysing GANs for unconditional image generation and reporting that, overall, most variants can reach similar scores (Lucic et al., 2017). We also did not find improvements in terms of epochs needed to converge, although training GANs with gradient penalty adds extra computational load, making the optimisation of WGAN-GP slower.

5.3 GAN Regularisation: Exploiting Texture Biases in Mask Discriminators

In the remainder of this chapter, we present regularisation techniques to stabilise the training of GANs for semantic segmentation. Based on recent findings that CNNs are intrinsically biased to focus on the image textural information, we hypothesise that it is possible to regularise mask discriminators by adding textures on top of the flat segmentation masks.

5.3.1 Introduction

Until recently, it was widely believed that convolutional neural networks recognize objects because they learn increasingly complex and higher-

level spatial features. However, Geirhos et al., [2018] demonstrated that CNNs are heavily biased towards image texture, which they exploit for their predictions. From a neuroscientific perspective, these results indicate a significant divergence from the primate visual processing, whose bias toward shapes is instead well documented (Landau, Smith, and Jones, [1988])²

Moreover, in a recent study, Kayhan and Gemert, [2020] have shown that convolutional layers can exploit the absolute spatial location of an object. Such behaviour is made possible by learning filters that respond exclusively to specific spatial coordinates and image boundary effects, which exposes CNNs to statistical biases when dealing with finitely sampled data (Kayhan and Gemert, [2020]).

Based on these observations, we would like to address the following questions. Since segmentation masks lack textures, *is it possible that mask discriminators learn to distinguish real from fake image mostly focusing on object size and position?* If that is the case, *can mask discriminators learn a better shape prior by enriching the binary segmentations with synthetic textures?*

In the following, we study the possibility to regularise mask discriminators by adding synthetic textures on top of the masks. Our results suggest that, in some cases, textures can be a useful and non-computationally expensive regulariser.

5.3.2 Related Work

After the work of Geirhos et al., [2018], several papers reported the texture bias of convolutional neural networks. Recently, Brendel and Bethge, [2019] showed that to perform an image classification task it is sufficient to use small image patches, without taking into account their spatial ordering. In particular, after splitting an image into tiny unordered regions, a CNN classifier can correctly classify the image content, without any consideration of the global spatial relationships between the patches. In other words, CNNs can solve the classification task by only using the textural information inside the image patches.

²For completeness, we mention that part of the human visual cortex also responds to textural information (Schwartz and Simoncelli, [2001]; Freeman et al., [2013]).

Also Malhotra and Bowers, [2019] have observed that contrarily to humans, CNNs do not have a shape-bias, but they rely on whichever features allows them to perform the best prediction. Wang, Wu, et al., [2020] have reported that CNNs exploit high-frequency image components that are not perceivable by humans, offering also a possible explanation for adversarial attacks. Hermann and Kornblith, [2020] have recently shown that also generative models exhibit a texture bias. Additionally, they have discussed that training objective, model architecture, data pre-processing, and hyperparameter choices all make distinct contributions to the level of texture bias in a model.

Although texture bias can help in standard image classification tasks, Ringer et al., [2019] have shown that it significantly harms few-shot learning, where the distribution shift is a crucial problem and a focus on the object shape can make models more robust. For this reason, Azad et al., [2021] have proposed to reduce the texture bias in few-shot image segmentation, integrating a set of Difference of Gaussians (DoG) (Lowe, [2004]) into the learned feature space. The role of the DoG is to attenuate high-frequency local components, which the authors hypothesise are associated with textures. Zaech et al., [2019] have proposed a training procedure that facilitates texture underfitting to improve domain adaptation. Similarly, recent work has suggested that textures may reduce the accuracy in object recognition tasks and that one should remove the texture bias to learn a more object-oriented image classification and segmentation (Zhang, Zhang, Xu, et al., [2020]; Kim and Byun, [2020]; Chai, Rueckert, and Fetit, [2020]). Differently from these methods, we do not want to learn to perform a task on the input *images*. On the contrary, we would like to add textures on top of flat segmentation masks to exploit the CNN textural bias and provide additional input signal to a mask discriminator, while ensuring we do not incur in intensity distribution shift problems.

Most recently, Sinha, Garg, and Larochelle, [2020] have introduced a curriculum-based scheme to improve CNNs' ability to represent both the shape and textural information. In detail, they performed the training procedure by controlling the amount of textural information that is present in the data. To adjust the textures level, the authors suggest to

convolve the output of a CNN layer with a low-pass Gaussian filter. As the training proceeds, they gradually re-introduce textures by annealing the standard deviation of the Gaussian kernels.

While most of these methods attempt to remove the texture bias, in this chapter, we suggest going in the opposite direction and exploiting the texture bias to train better mask discriminators. We observe that introducing the *same* textural statistics on all the binary segmentation masks should not incur in the distribution shift problem observed by Ringer et al., [2019]. Instead, using different textures for different object classes provides a dense signal throughout the entire mask, rather than just sparse signals in correspondence of the object boundaries.

5.3.3 Method

We consider a semi-supervised GAN formulation, where we jointly train a conditional mask generator, or segmentor, and an adversarial mask discriminator. The architectures of the two models are the same as in Section [5.2], and they are described in the Appendix [A.1]. As standard practice, the discriminator alternately receives batches of generated and batches of real segmentation masks, learning to say one distribution apart from the other.

The peculiarity of the proposed approach consists in using synthetic textures to enrich the segmentation masks analysed by the adversarial discriminator. We artificially generate the textures as sinusoidal patterns modulated by learnable frequency and phase parameters, which we detail below.

Textures Generation

We model textures as a smooth sinusoidal “grid” pattern \mathbf{G} , parameterised using frequency and phase components on the orthogonal \underline{x} and \underline{y} axis of the image. We align \underline{x} parallel to the image width and \underline{y} parallel to the height. The grid values along the axis are described by the tuples

$(\underline{g}_x, \underline{g}_y)$, defined as:

$$\begin{aligned} \underline{g}_x &= \sin(2\pi f_x \cdot \underline{x} + b_x) \\ \underline{g}_y &= \sin(2\pi f_y \cdot \underline{y} + b_y), \end{aligned} \tag{5.4}$$

where f_x and f_y are the grid frequencies, \underline{x} and \underline{y} the pixel coordinates, and b_x and b_y sinusoidal phases. Once generated the textures grid, we add it to the segmentation mask \mathbf{y} to obtain an augmented version \mathbf{y}' , as:

$$\mathbf{y}' = \mathbf{y} \cdot (1 + m\mathbf{G}), \tag{5.5}$$

where m is the desired textural amplitude. We consider the introduced textures as a form of structured noise, and thus we choose a small value for m . Since the segmentation masks are one-hot encoded, the maximum possible value for a pixels in \mathbf{y} is 1. For this reason, the amplitude m should preferably be smaller than 1, to have a good signal to noise ratio (SNR). In our case, we consider textures whose signal power is decreased by 20dB compared to the binary pixels, and thus set $m = 0.1$.

We learn the grid parameters via gradient descent optimisation. We first initialise the values $f_{x_0}, f_{y_0}, b_{x_0}$ and b_{y_0} with a normal distribution with zero mean and 1 standard deviation. Then, we map these variables to a suitable range, computing:

$$\begin{aligned} f_x &= 0.5 \cdot \text{sigmoid}(f_{x_0}) \\ f_y &= 0.5 \cdot \text{sigmoid}(f_{y_0}) \\ b_x &= \pi \cdot \text{tanh}(b_{x_0}) \\ b_y &= \pi \cdot \text{tanh}(b_{y_0}), \end{aligned}$$

Finally, we use f_x, f_y, b_x and b_y to parameterise the textures grid \mathbf{G} with Equation 5.4. Notice that we scale f_x and f_y to have a maximum value of 0.5, and thus satisfy the Nyquist theorem, which is necessary to prevent aliasing artifacts. Similarly, we bound b_x and b_y in the $0 \div 2\pi$ range to consider any possible angle while ensuring a bijective mapping (i.e., once fixed f , any b maps to a different sinusoidal amplitude).

We account for the possibility of having constructive and destructive interference inside the textures grid, and thus normalise \mathbf{G} in the

0÷1 range, before obtaining the augmented mask as described in Equation 5.5. Lastly, we let the model generate a separate grid of textures for each class of the segmentation mask, independently of the input image. As a result, the mask associated with each class may be enriched with a class-specific texture.

Training Objectives and Optimisation

We train the segmentor in a semi-supervised fashion, using only a small portion of annotated data and many unlabelled images. For the unpaired images, we optimise the discriminator $\Delta(\cdot)$ and the segmentor $\Sigma(\cdot)$ according to the two objectives: $\min_{\Delta} \mathcal{V}_{LS}(\Delta)$ and $\min_{\Sigma} \mathcal{V}_{LS}(\Sigma)$. In particular, we define the training losses:

$$\begin{aligned}\mathcal{V}_{LS}(\Delta) &= \frac{1}{2} E_{\mathbf{y} \sim p(\mathbf{y})} [\Phi(\Delta(\mathbf{y}), \ell_{real})] + \frac{1}{2} E_{\mathbf{x} \sim p(\mathbf{x})} [\Phi(\Delta(\Sigma(\mathbf{x})), \ell_{fake})] \\ \mathcal{V}_{LS}(\Sigma) &= \frac{1}{2} E_{\mathbf{x} \sim p(\mathbf{x})} [\Phi(\Delta(\Sigma(\mathbf{x})), \ell_{real})],\end{aligned}$$

where: $\mathbf{x} \sim p(\mathbf{x})$ is an unlabelled image, $\mathbf{y} \sim p(\mathbf{y})$ an unpaired segmentation mask, ℓ_{real} and ℓ_{fake} are the labels for real and generated masks, respectively, and $\Phi(\cdot)$ is a metric defined according to the GAN type. We consider three possible formulations of $\Phi(\cdot)$, according to the popular variants of: Non-saturating GAN (NSGAN, Goodfellow et al., 2014), Least-square GAN (LSGAN, Mao, Li, et al., 2017), and Wasserstein GAN with gradient penalty (WGAN-GP, Gulrajani et al., 2017).

When annotations are available, we optimise the segmentor with the supervised Dice loss, proposed by Milletari, Navab, and Ahmadi, 2016, between the ground truth segmentation masks and the mask predicted by the segmentor.

As a result, the training objective of the segmentor \mathcal{L}_{Σ} follows a multi-task learning formulation, containing supervised and unsupervised components:

$$\mathcal{L}_{\Sigma} = \mathcal{L}_{SUP} + a \cdot \mathcal{L}_{ADV}, \quad (5.6)$$

with $\mathcal{L}_{ADV} = \mathcal{V}_{LS}(\Sigma)$. Since the adversarial discriminator can only judge the predicted segmentation from a general point of view (*real* vs. *fake*),

but it does not ensure a correct mapping $p(\mathbf{x}) \rightarrow p(\mathbf{y})$, we give more importance to the supervised loss, setting $a = 0.1$.

We minimize the training losses using Adam optimiser (Kingma and Ba, 2015), a learning rate of 0.0001 and a batch size of 12.

5.3.4 Experimental Setup

Below, we first describe the datasets used for our experiments. Then, we describe the adopted baselines, benchmark methods and evaluation protocol.

Data

We test our model on data from the cardiac dataset ACDC (Section 3.8.1), and the abdominal organ CHAOS dataset (T1 and T2 images, Section 3.8.4). In both cases, we considered a semi-supervised training scenario, where only a portion of the available data is annotated. The training datasets also contain a subset of unpaired masks, which may be obtained from a different modality or acquisition protocol (Larrazabal et al., 2020; Painchaud et al., 2020). Differently from the experiments in Section 5.2, we do not consider LVSC data. In fact, the LVSC dataset only contains segmentation of the left myocardium, which is a thin structure and thus, adding textures would not be useful.

We split the considered datasets into groups of 70% of patients for training, 15% for validation, and 15% for the test set, respectively. To test the model in a challenging supervision setup, we consider only one-tenth of annotated patients out of the 70% training MRI scans (i.e. 7 patients on ACDC, 2 patients on CHAOS T1 and CHAOS T2).

Baselines, Benchmark Methods and Evaluation Protocol

We compare the performance of a GAN regularised with synthetic textures, a vanilla UNet segmentor, and the non-regularised GAN variants: NSGAN, LSGAN, and WGAN-GP. In addition, we also consider each GAN variant including an additional regularisation at the discriminator

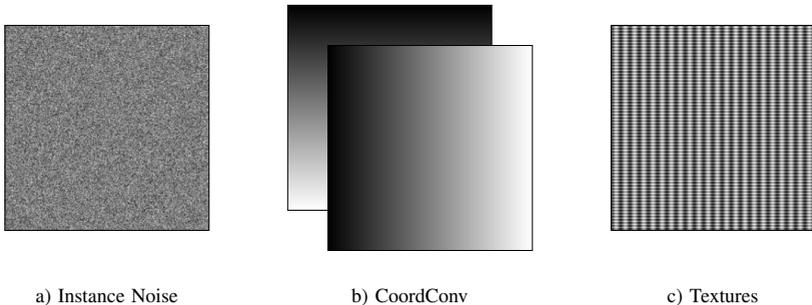


Figure 24: Comparison of different regularisation techniques: **a)** Instance Noise adds a small random perturbation to the discriminator input; **b)** CoordConv introduces continuous spatial information by adding a spatial coordinate grid to the data; **c)** Texture layer learns to introduce continuous information in the form of a sinusoidal grid.

level, detailed below. The goal of this regularisation is to introduce a continuous signal on top of the flat segmentation masks.

We consider the techniques:

- **Instance Noise.** We can see textures as a form of structured noise. For this reason, we compare the proposed regulariser to another method introduced to stabilize GAN training: instance noise, a form of *unstructured textures*. Recently introduced by Sønderby et al., [2017], instance noise consists of adding a small random perturbation to the discriminator input to enlarge the data distribution. Consequently, the real and the generated data are more likely to fall in dense regions of the data manifold, and it becomes easier to have overlapping support between the two distributions. Thus, the discriminator can more easily find a unique decision boundary to classify real and generated data, and adversarial training is more stable. In our experiments, we use noise with the same amplitude as that of our structured textures.
- **CoordConv.** From another perspective, adding textures on top of the segmentation masks is a way of transforming the mask val-

ues from binary to continuous. An alternative approach for making the binary masks continuous is using a CoordConv layer (Liu, Lehman, et al., 2018). The CoordConv solution explicitly introduces spatial coordinates in a convolutional layer, obtained by concatenating a hard-coded coordinate grid to the extracted features maps. As summation and concatenation have similar practical effect in CNNs³ (Dumoulin, Perez, et al., 2018), we consider CoordConv as another possible way of generating continuous segmentations from the binary ones.

We report a visual comparison of the aforementioned techniques and our method in Figure 24.

Assuming that using a better discriminator improves the training signals for the segmentor, we evaluate the quality of each regulariser in terms of segmentation performance on the test data. We measure performance in a 3-fold cross-validation, using the Dice and Intersection over Union (IoU) scores.

5.3.5 Results

We show visual examples the textures introduced on top of the segmentation masks by our model in Figure 25, while we report results of our experiments in the box plots in Figure 26. The box plots report median and inter-quartile range (IQR) of each metric on a given test set, indicating values outside 1.5 times the IQR as outliers.

As Figure 26 shows, the NSGAN variant is the one benefitting the most from regularisation. NSGAN shows improved performance for all datasets when a regularisation technique is applied, confirming that using continuous segmentation masks is beneficial to the model. In particular, our method is the best on CHAOS datasets, where the discriminator has the best compromise between performance median and spread.

³To see this, consider the operation $W[x, y] = W_0x + W_1y$, where $[x, y]$ denotes the concatenation of the features maps x and y , and W is a weight matrix that we can split horizontally into W_0 and W_1 . Comparing this to $W(x + y) = Wx + Wy$, we observe that it is sufficient to constrain $W_0 = W_1$ to make summation and concatenation equivalent.

	Mask	Textures	Mask + Textures	Textures Parameters by Class				
ACDC				f_x	f_y	b_x	b_y	
				Background	+0.251	+0.249	+0.012	+0.013
				Right Ventricle	+0.248	+0.249	-0.020	-0.010
				Left Ventricle	+0.249	+0.250	-0.016	+0.011
				Myocardium	+0.250	+0.250	-0.003	-0.001
CHAOS - T1				f_x	f_y	b_x	b_y	
				Background	+0.250	+0.249	+0.000	-0.001
				Liver	+0.249	+0.250	-0.001	+0.005
				Kidney 1	+0.250	+0.249	-0.001	-0.001
				Kidney 2	+0.250	+0.250	-0.001	-0.008
			Spleen	+0.250	+0.249	-0.001	-0.003	
CHAOS - T2				f_x	f_y	b_x	b_y	
				Background	+0.250	+0.250	-0.001	+0.004
				Liver	+0.250	+0.251	-0.001	+0.001
				Kidney 1	+0.249	+0.250	-0.004	-0.001
				Kidney 2	+0.249	+0.249	-0.002	-0.002
			Spleen	+0.251	+0.250	+0.000	-0.001	

Figure 25: Examples of textures added by the discriminator on top of the segmentation masks. We show examples on ACDC, CHAOS-T1 and CHAOS-T2 test sets. The textures appear as a small amplitude sinusoidal grid pattern, having different phases and oscillation frequencies for each class (including the background). To easy visualization, all images are cropped around the object of interest. On the right, we report the values of the textural parameters learned by the mask discriminator, in radians.

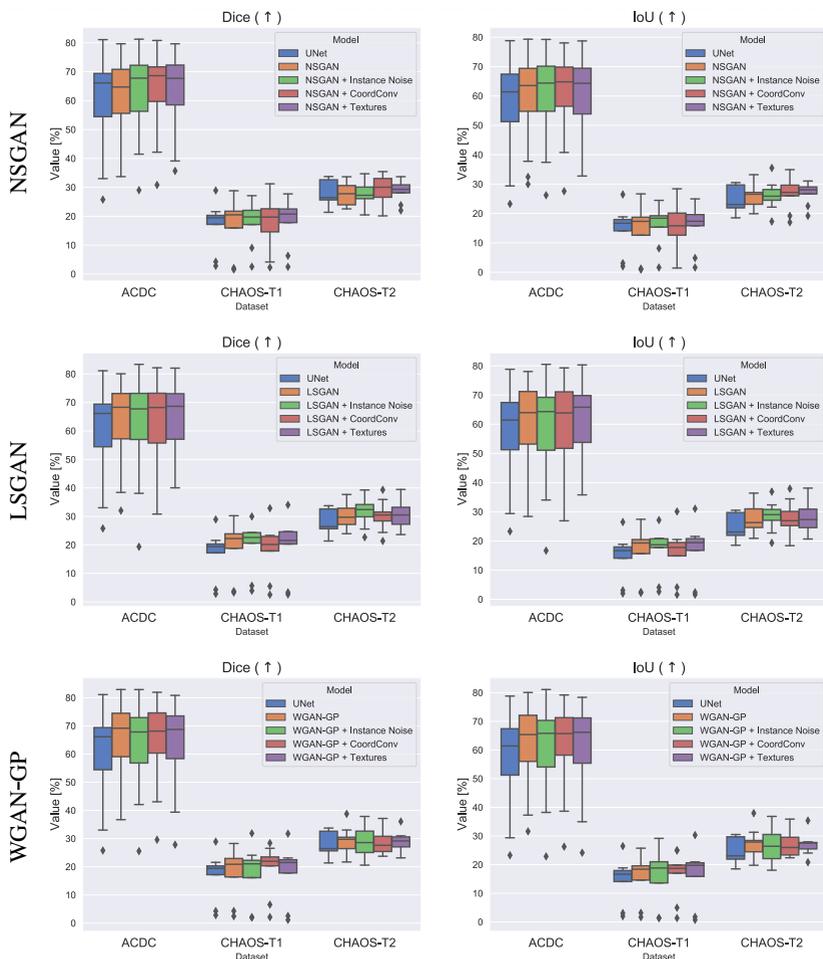


Figure 26: Segmentation performance of a vanilla UNet segmentor and the analysed GAN variants when regularised with different techniques. Performance is measured in terms of Dice (\uparrow) and IoU (\uparrow) scores, where arrows show the metric improvement direction. The box plots report median and inter-quartile range (IQR) of each metric on a given test set, considering outliers those values falling outside $1.5 \times \text{IQR}$. Overall, we observe that regularisation leads to larger performance gains on NSGAN, while LSGAN and WGAN-GP have smaller gains.

On the other hand, we find LSGAN and WGAN-GP generally more stable, with regularisation showing a small effect.

We confirm that optimising a segmentor with an adversarial shape prior always improves the segmentation with respect to a vanilla UNet. In fact, the semi-supervised training of the GAN segmentor allows to introduce additional information via unpaired images, useful to generalise better when the annotated data are scarce.

5.4 Summary

This chapter demonstrated that Generative Adversarial Networks can learn powerful shape priors to be used to regularise training. However, in the previous sections' setup, the adversarial discriminator could only penalise the segmentor globally. In other terms, the segmentor is not encouraged to learn that shapes have a hierarchical structure and must satisfy both short-range and long-range pixel dependencies in the image.

Recently, multi-scale adversarial frameworks have shown to be effective to learn better shape priors (Luo, Zheng, et al., 2018). Standard multi-scale GANs consider a generator and multiple discriminators. Each discriminator learns an independent shape prior at a given resolution level and drives the generator training to produce realistic predictions at various scales. However, these approaches do not enforce any consistency between the different resolution levels. On the other hand, it would be beneficial to introduce multi-scale shape consistency in the segmentor. Moreover, using a single discriminator would reduce the computational load of training a multi-scale GAN. Improving these two aspects would provide a simple and powerful regulariser for semi-supervised and weakly-supervised learning, where object shape information is often missing from the training data. To address these limitations, the next chapter introduces a novel multi-scale GAN formulation that biases the segmentor toward more structured predictions.

Lastly, we observe that standard mask discriminators bias the segmentor penalising unrealistic predictions during training. However, this behaviour remains tied to the *training stage*, and discriminators are com-

monly discarded at inference. We argue that properly trained discriminators can still provide a useful shape prior after training. We find exciting the idea of re-using mask discriminators at test-time to detect, and ideally correct, unrealistic predictions of the segmentor. Toward this goal, Chapter 8 will present a novel approach showing that adversarial shape priors improve test-time segmentation under data distribution shifts.

Chapter 6

Multi-scale Adversarial Shape Priors for Weak Supervision

▣

In the previous chapters, we have thoroughly discussed that obtaining large-scale datasets with pixel-level annotations is challenging, particularly in medical imaging, where annotating segmentation masks is time-expensive and requires expert knowledge. For this reason, shape priors can help to include additional information when labels are missing or partial.

In Chapter 5, we presented several different methods for introducing data-driven shape priors in a segmentor. This type of priors is helpful for semi-supervised learning (as shown in Chapter 4 and 5), but also for weakly-supervised approaches which must rely on imperfect forms of annotations for training. Below, we offer a novel approach to improve standard adversarial training in the presence of weak supervision. We

This chapter is based on:

- Valvano, Gabriele, Andrea Leo, and Sotirios A. Tsaftaris (2021c). "Learning to Segment From Scribbles Using Multi-Scale Adversarial Attention Gates". In: *IEEE Transactions on Medical Imaging*. DOI: [10.1109/TMI.2021.3069634](https://doi.org/10.1109/TMI.2021.3069634)

also report results of the proposed method in a semi-supervised setting.

In the following, we optimise a novel multi-scale GAN with unpaired segmentation masks while ensuring a low computational training cost. Conditioned by an input image, the GAN segmentor learns to generate realistic predictions at multiple scales, using scribble supervision to learn the mapping to the correct spatial object location. Central to the model’s success is a novel attention gating mechanism that we condition with adversarial signals to act as a shape prior, resulting in better object localisation at multiple scales. Subject to adversarial conditioning, the segmentor learns attention maps that are semantic, suppress the noisy activations outside the objects, and reduce the vanishing gradient problem in the deeper convolutional layers.

We evaluate our model on several medical (ACDC, LVSC, CHAOS) and non-medical (PPSS) datasets. We report performance levels matching those achieved by models trained with fully annotated segmentation masks. We also demonstrate extensions in a variety of settings: semi-supervised learning, combining multiple scribble sources (a crowdsourcing scenario), and multi-task learning (combining scribble and mask supervision).

6.1 Introduction

Convolutional Neural Networks (CNNs) have obtained impressive results in computer vision. However, their ability to generalize on new examples is strongly dependent on the amount of training data, thus limiting their applicability when annotations are scarce. There has been a considerable effort to exploit semi-supervised and weakly-supervised strategies. For semantic segmentation, semi-supervised learning (SSL) aims to use unlabeled images, generally easier to collect, together with some fully annotated image-segmentation pairs (Chapelle, Scholkopf, and Zien, [2009](#); Cheplygina, de Bruijne, and Pluim, [2019](#)). However, the information inside unlabeled data can improve CNNs only under specific assumptions (Chapelle, Scholkopf, and Zien, [2009](#)), and SSL requires representative image-segmentation pairs being available.

Alternatively, weakly-supervised approaches (Khoreva et al., 2017; Souly, Spampinato, and Shah, 2017; Can et al., 2018; Zhou, Li, et al., 2019) attempt to train models relying only on weak annotations (e.g., image-level labels, sparse pixel annotations, or noisy annotations (Tajbakhsh et al., 2020)), that should be considerably easier to obtain. Thus, building large-scale annotated datasets becomes feasible and the generalization capability of the model per annotation effort can dramatically increase: e.g., 15 times more bounding boxes can be annotated within the same time compared to segmentation masks (Lin, Maire, et al., 2014). Among weak annotations, scribbles are of particular interest for medical image segmentation, because they are easier to generate and well suited for annotating nested structures (Can et al., 2018). Unfortunately, learning from weak annotations does not provide a supervisory signal as strong as one obtained from fine-grained per-pixel segmentation masks, and training CNNs is harder. Thus, improved training strategies can enable remarkable gains with weaker forms of annotations.

6.1.1 Overview of the proposed approach

In this paper, we introduce a novel training strategy in the context of weakly supervised learning for multi-part segmentation. We train a model for semantic segmentation using scribbles, shaping the training procedure as an adversarial game (Goodfellow et al., 2014) between a conditional mask generator (the segmentor) and a discriminator. We obtain segmentation performance comparable to when training the segmentor with full segmentation masks. We demonstrate this for the segmentation of the heart, abdominal organs, and human pose parts.

Our uniqueness is that we use adversarial feedback at all scales, coupling the generator with a multi-scale discriminator. But, differently from other multi-scale GANs (Denton, Chintala, Fergus, et al., 2015; Karras, Aila, et al., 2017; Luo, Zheng, et al., 2018), our generator includes customized attention gates, i.e. modules that automatically produce soft region proposals in the feature maps, highlighting the salient information inside of them. Differently from the attention gates presented in

(Schlemper et al., 2019) ours are conditioned by the adversarial signals, which enforce a stronger object localization in the image. Moreover, differently from other multi-scale GANs (Denton, Chintala, Fergus, et al., 2015; Karras, Aila, et al., 2017; Luo, Zheng, et al., 2018) we use a single discriminator rather than multiple ones, thus reducing the computational cost whilst retaining their advantages in semantic segmentation.

The discriminator, acting as a learned shape prior, is trained on a set of segmentation masks, obtained from a different data source¹ and is thus unpaired. We drive the segmentor to generate accurate segmentations from the input images, while satisfying the multi-scale shape prior learned by the discriminator. We encourage a tight multi-level interaction between segmentor and discriminator introducing *Adversarial Attention Gating*, an effective attention strategy that, subject to adversarial conditioning, i) encourages the segmentor to predict masks satisfying multi-resolution shape priors; and ii) forces the segmentor to train deeper layers better. Finally, we also penalize the segmentor when it predicts segmentations that do not overlap with the available scribbles, pushing it to learn the correct mapping from images to label maps.

We summarise the proposed approach in Figure 27.

6.1.2 Contributions

We summarize the contributions of this work as follows:

- We use scribble annotations to learn semantic segmentation during a multi-scale adversarial game.
- We introduce Adversarial Attention Gates (AAGs): effective prior-driven attention gates that force the segmentor to localize objects in the image. Subject to adversarial gradients, AAGs also encourage a better training of deeper layers in the segmentor.
- We obtain state-of-the-art performance compared to other scribble-supervised models on several popular medical datasets (ACDC,

¹We simulate a realistic clinical setting, where the unpaired masks can be obtained from a different modality or acquisition protocol (Larrazabal et al., 2020; Painchaud et al., 2020).

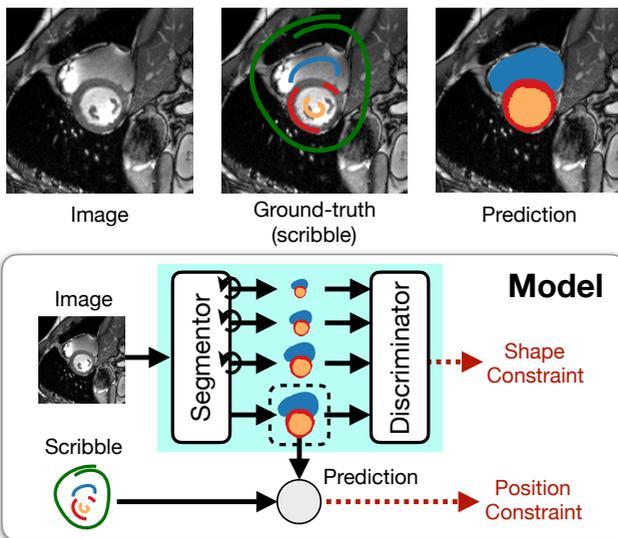


Figure 27: In an adversarial game, our model learns to generate segmentation masks that look realistic at multiple scales and overlap with the available scribble annotations. Loopy arrows in the figure, on the segmentor, represent the proposed attention gates, which under adversarial conditioning suppress irrelevant information in the extracted features maps.

LVSC and CHAOS, described in Section [3.8.1](#), [3.8.2](#), and [3.8.4](#) and computer vision data (PPSS, Section [3.8.5](#)).

- We investigate diverse learning scenarios, such as: learning from different extents of weak annotations (i.e., semi-supervised learning); learning from multiple scribbles per image (and thus simulating a crowdsourcing setting); and finally learning also with few strong supervision pairs of segmentation masks and images (i.e., multi-task learning).
- Lastly, we compare our model, trained on scribbles, with a method designed for few-shot learning, which we train with densely annotated segmentation masks. With this experiment, we show the advantage of collecting large-scale, weakly annotated datasets.

We release expert-made scribble annotations for the ACDC dataset and the code used for the experiments. Both are available on our project page, at: <https://vios-s.github.io/multiscale-adversarial-attention-gates>,

6.2 Related Work

A large body of research aimed at developing learning algorithms that rely less on high-quality annotations (Cheplygina, de Bruijne, and Pluim, 2019; Tajbakhsh et al., 2020). Below, we briefly review recent weakly supervised methods that use scribbles to learn image segmentation. Then, we discuss what are the advantages of our adversarial setup compared to other multi-scale GANs. Finally, we discuss the difference between the attention gates that are an integral part of our segmentor and other canonical attention modules.

6.2.1 Learning from Scribbles

Scribbles are sparse annotations that have been successfully used for semantic segmentation, reporting near full-supervision accuracy in computer vision and medical image analysis. However, scribbles lack information on the object structure, and they are limited by the uncertainty of unlabelled pixels, which makes training CNNs harder, especially in boundary regions (Lin, Dai, et al., 2016). For this reason, many approaches have tried to expand scribble annotations by assigning the same class to pixels with similar intensity and nearby position (Lin, Dai, et al., 2016; Ji et al., 2019). At first, these approaches relabel the training set propagating annotations from the scribbles to the adjacent pixels using graph-based methods. Then, they train a CNN on the new label maps. A recent variant has been introduced by Can et al., 2018, who suggest estimating the class of unlabelled pixels via a learned two-step procedure. In the first step, they train a CNN directly with scribbles. Subsequently, they relabel the training set by refining the CNN predictions with Conditional Random Fields (CRF), and they retrain the CNN

on the new annotations.

The major limitation of the aforementioned approaches is relying on dataset relabeling, which can be time-consuming and is prone to errors that can be propagated to the models during training. Thus, many authors (Can et al., [2018]; Tang, Perazzi, et al., [2018]) have investigated alternatives that avoid this step, post-processing the model predictions with CRF (Chen, Papandreou, et al., [2017]) or introducing CRF as a trainable layer (Zheng et al., [2015]). Tang, Perazzi, et al., [2018] have also demonstrated the possibility to substitute the CRF-based refining step, directly training a segmentor with a CRF-based loss regulariser.

Similarly, here we propose a method that avoids the data relabeling step. We train our model to directly learn a mapping from images to segmentation masks, and we remove expensive CRF-based post-processing. We cope with unlabelled regions of the image introducing a multi-scale adversarial loss which, differently from the loss introduced by Tang, Perazzi, et al., [2018], does not rely on CRF, and can handle both long-range and short-range inconsistencies in the predicted masks.

Concurrent to our work, Zhang, Zhong, and Li, [2020] recently introduced a method that learns to segment images from scribbles using an adversarial shape prior. However, they suggest using a PatchGAN (Isola et al., [2017]) discriminator, which only focuses on *local* properties of the generated segmentations, while we introduce a method that focuses on both *local* and *global* aspects.

6.2.2 Shape Priors in Deep Learning for Medical Imaging

In semantic segmentation, there has been considerable interest in incorporating prior knowledge about organ shapes to obtain more accurate and plausible results (Nosrati and Hamarneh, [2016]). Below, we summarise recent work on shape priors in Deep Learning.

Recently, Clough et al., [2020] used Persistent Homology to enforce shape priors in medical image segmentation. Oktay, Ferrante, et al., [2017] demonstrated that we can learn a data-driven shape prior with a convolutional autoencoder trained on unpaired segmentation masks, and

it can be used as regulariser to train a segmentor. Dalca, Guttag, and Sabuncu, [2018] suggested learning the shape prior with a variational autoencoder (VAE) (Kingma and Welling, [2014]), and then share part of the VAE weights with a segmentor. Other approaches included shape priors as post-processing, regularising the training (Yue et al., [2019]), or adjusting predictions at inference, using VAEs (Painchaud et al., [2019]) or Denoising Autoencoders (Larrazabal et al., [2020]). Kervadec, Dolz, Tang, et al., [2019] suggested introducing size information as a differentiable penalty, during training. Alternatively, Dalca, Yu, et al., [2019] proposed to learn to warp a segmentation atlas. Other methods (Kohl et al., [2018]; Baumgartner et al., [2019]) proved that image segmentation has intrinsic uncertainty, which can be reflected in the learned shape prior. Finally, a body of literature showed that decoupling (disentangling) object shapes and appearance is beneficial in a lack of data (Chartsias, Joyce, et al., [2019]; Yang et al., [2019]), as well as using temporal consistency constraints on the object shapes dynamics (Valvano, Chartsias, et al., [2019]).

Herein, we will focus on a particular type of shape prior, learned by a multi-scale GAN from unpaired segmentation masks. Particularly, we use an adversarial loss during training and avoid expensive post-processing of the predicted masks.

6.2.3 Multi-scale GANs

Herein, we use the generator as a segmentor, which we train to predict realistic segmentation masks at multiple scales. Recently, other methods introduced multi-scale adversarial losses for segmentation. For example, Xue et al., [2018] proposed to use the discriminator as a critic, measuring the ℓ_1 -distance between *real* and *fake* inputs in features space, at multiple resolution levels. In particular, pairs of real and fake inputs consist in the Hadamard product between an image and the associated ground truth or predicted segmentation mask, respectively. Also Luo, Zheng, et al., [2018] separated *real* from *fake* input pairs at multiple scales, using two separate discriminators (one working at high, one at low resolution) to distinguish the image concatenation with the associated ground truth or

predicted segmentation, respectively.

Unfortunately, these approaches rely on image-segmentation pairs to train the discriminator. Thus, training the segmentor with unlabelled, or weakly annotated data is not possible. Instead, we train a discriminator using *only* masks, making the model suitable for semi- and weakly-supervised learning. Also, contrarily to Luo, Zheng, et al., [2018], we use a single multi-scale discriminator rather than two, keeping the computational cost lower.

Finally, while previous approaches use multi-scale GANs with strong annotations, this is, to the best of our knowledge, the first work to explore their use in weakly-supervised learning. Furthermore, we alter the canonical interplay between discriminator and segmentor to improve the object localization in the image, that we obtain with a novel adversarial conditioning of the attention maps learned by the segmentor.

6.2.4 Attention Gates

Due to the ability to suppress irrelevant and ambiguous information, attention gates have become an integral part of many sequence modeling (Vaswani et al., [2017]) and image classification (Jetley et al., [2018]) frameworks. Recently, they have also been successfully employed for segmentation (Schlemper et al., [2019]; Oktay, Schlemper, et al., [2018]; Wang, Deng, et al., [2018]; Sinha and Dolz, [2020]; Fu et al., [2019]), along with the claim that gating helps to detect desired objects. However, standard approaches don't incorporate any explicit constraint in the learned attention maps, which are generally predicted by the neural network autonomously. On the contrary, we show that conditioning the attention maps to be semantic, i.e. able to localize and distinguish separate objects, considerably boosts the segmentation performance. Herein, we introduce a novel attention module named Adversarial Attention Gate (AAG), whose learning is conditioned by a discriminator.

6.3 Proposed Approach

In this section, we present a general overview of the proposed method. Then, we detail model architectures and training objectives.

We will assume a weakly supervised setting, where we have access to: i) image-scribble pairs (x, y_s) , being x the image and y_s the associated scribble; ii) unlabelled images; and iii) a set of segmentation masks y unrelated to any of the images²

6.3.1 Method Overview

We formulate the training of a CNN with weak supervision (i.e., scribbles) as an adversarial game. Particularly, we use an adversarial discriminator to learn a multi-resolution shape prior, and we enforce a mask generator, or segmentor, to satisfy it, supported by the purposely designed adversarial attention gates. Critically, AAGs localize the objects to segment at multiple resolution levels and suppress noisy activations in the remaining parts of the image (see Figure 28).

In detail, we jointly train a multi-scale segmentor $\Sigma(\cdot)$ and a multi-scale adversarial discriminator $\Delta(\cdot)$. $\Sigma(\cdot)$ is supervisedly trained to predict segmentation masks $\tilde{y} = \Sigma(x)$ that overlap with the scribble annotations, when available. Meanwhile, $\Delta(\cdot)$ learns to distinguish real segmentation masks from those (fake) predicted by the segmentor (i.e., $\Delta(y)$ vs $\Delta(\tilde{y})$) (Goodfellow et al., 2014), at multiple scales. We model both $\Sigma(\cdot)$ and $\Delta(\cdot)$ as CNNs.

In principle, other models can be used to learn multi-scale shape priors, as multi-scale VAEs (Baumgartner et al., 2019; Vahdat and Kautz, 2020). We use GANs because they can be trained together with the segmentor in an adversarial game. The potential of using multi-scale VAEs in weakly supervised segmentation learning is an open research problem, which we leave for future work.

²In Section 6.5.6 we will also investigate a mixed setting, where we additionally have: iv) pairs of image-segmentation masks (x, y) .

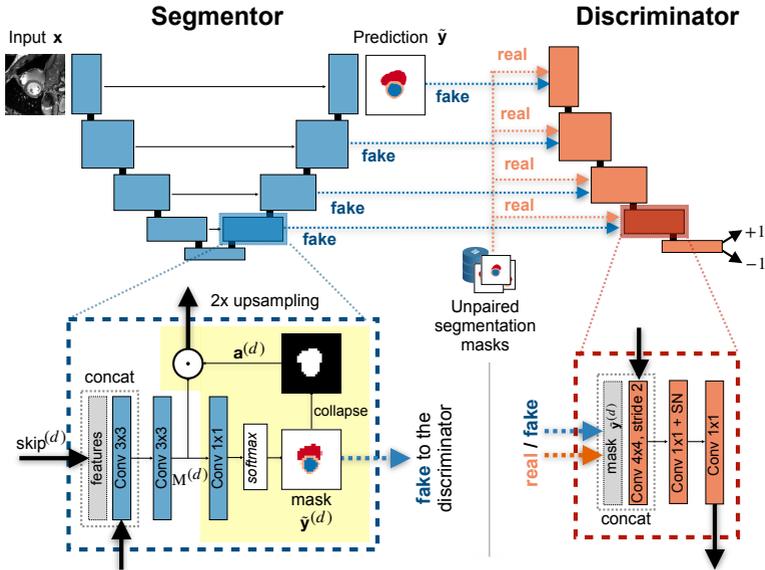


Figure 28: Model architectures. Top: segmentor and discriminator interact at multiple scales. Bottom: convolutional blocks detail. In yellow background, the Adversarial Attention Gate (AAG).

6.3.2 Architectures

We now describe the architectural details of our model.

Segmentor $\Sigma(\cdot)$ We modify a UNet (Ronneberger, Fischer, and Brox, 2015) to include AAG modules in the decoder and to allow collaborative training between segmentor and discriminator at multiple scales (Fig. 28). We leave the UNet encoder as in the original framework, allowing to extract feature maps at multiple depth levels and propagate them to the decoder via skip connections and concatenation (Ronneberger, Fischer, and Brox, 2015). Instead, we alter the decoder such that, for every depth level d , after the two convolutional layers, an AAG first produces an attention map as the probabilistic prediction of a classifier (detailed below), then uses it to filter out activations from the input features map. Particularly, we use convolutional layers with $3 \times 3 \times k$ filters, being k

the number of input channels, and produce the features map $M^{(d)}$. Then, the AAG classifier uses $M^{(d)}$ to predict a segmentation $\tilde{y}^{(d)}$ at the given resolution level d . As a classifier, we use a convolutional layer with $c \times 1 \times 1 \times k$ filters (where c is the number of possible classes, including the background). We do not apply any *argmax* operation on its prediction, while we use a pixel-wise *softmax* to give a probabilistic interpretation of the output: as a result, every pixel is associated to a probability of belonging to every considered class, which is important to have smoother gradients on the learned attention maps. We then slice the predicted array removing the channel associated to the background, and we use the multi-channel soft segmentation: i) as input to the discriminator at the same depth level; and ii) to produce an attention map, obtained by summing up the remaining channels into a 2D probabilistic map $a^{(d)}$, localizing object positions in the image (Fig. 28). To force the segmentor to use $a^{(d)}$, we multiply the extracted features $M^{(d)}$ with $a^{(d)}$ using the Hadamard product (gating process). The resulting features maps are upsampled to the next resolution level via a nearest-neighbor interpolation. After each convolutional layer, we use batch normalization (Ioffe and Szegedy, 2015) and *ReLU* activation function.

Discriminator $\Delta(\cdot)$ We design an encoding architecture receiving *real* or *fake* inputs at multiple scales. This allows a multi-level interaction between $\Sigma(\cdot)$ and $\Delta(\cdot)$, and the *direct* propagation of adversarial gradients into the AAGs. We refer to this multi-level interaction as *Adversarial Deep Supervision* (ADS), as it regularises the output of AAG classifiers similarly to deep supervision, but using adversarial gradients (Fig. 29).

The *real samples* $\{y^{(d)}\}_{d=1}^4$ consist of expert-made segmentations, that we supply at full or downsampled resolution at multiple discriminator depths, while *fake samples* $\{\tilde{y}^{(d)}\}_{d=1}^4$ are the multi-scale predictions of the segmentor. In both cases, the lower-resolution inputs ($d > 1$) are supplied to the discriminator by simply concatenating them to the features maps it extracts at each depth d (Fig. 28, right).

The discriminator is a convolutional encoder adapted from (Chartias, Joyce, et al., 2019). At every depth d , at first, we process and down-

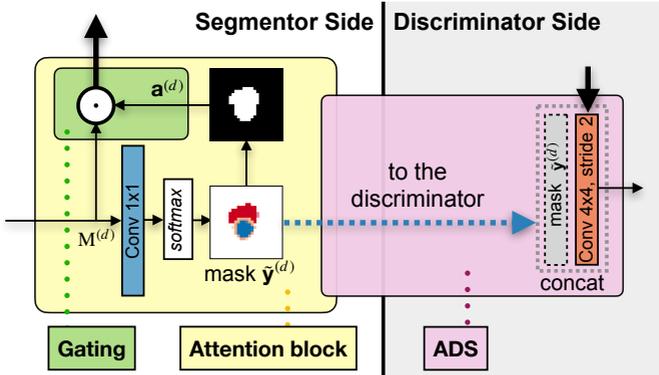


Figure 29: Adversarial Attention Gates consist of an attention block (yellow background in the figure) pairing Adversarial Deep Supervision (ADS, obtained via the connection in pink background) and a multiplicative gating operation (green background).

sample the features maps using a convolutional layer with $4 \times 4 \times k$ kernels and stride of 2. The number of filters follows that of the segmentor encoder (e.g. 32, 64, 128, 256, 512). We also use spectral normalization (Miyato et al., 2018) to improve training. Obtained feature maps are then compressed with a second convolutional layer using $12 \times 1 \times k$ filters. Both layers use *tanh* activations.

To improve the learning process and avoid overfitting, we make the adversarial game harder for the discriminator, using *label noise* (Salimans et al., 2016) and *instance noise* (Sønderby et al., 2017). In particular, we obtain label noise by a random flip of the discriminator labels (*real* vs *fake*) with a 10% probability, while we apply instance noise as a Gaussian noise with zero mean and standard deviation of 0.2, that we add to the highest resolution input.

Lastly, we compute the final prediction of the discriminator using a fully connected layer with scalar output ($\Delta(\mathbf{y}), \Delta(\tilde{\mathbf{y}})$).

6.3.3 Loss Functions and Training Details

We train the model minimizing supervised and adversarial objectives. In particular, we consider both contributions when scribble annotations are available for the input image, but only use the latter when dealing with unlabeled data.

Supervised Cost

When scribbles are available, we train the segmentor to minimize a pixel-wise classification cost on the annotated pixels of the image-scribble pair $(\mathbf{x}, \mathbf{y}_s)$, while, most importantly, we don't propagate any loss gradient through the unlabeled pixels. Crucially, we use the pixel-wise cross-entropy because it is shape-independent, and, to resolve the class imbalance problem, we multiply the per-class loss contribution by a scaling factor that accounts for the class cardinality. We can write the supervised cost as:

$$\mathcal{L}_{SUP} = \mathbb{1}(\mathbf{y}_s) * \left[- \sum_{i=1}^c w_i \cdot \mathbf{y}_{s_i} \log(\tilde{\mathbf{y}}_i) \right], \quad (6.1)$$

where i refers to each class and c is the number of classes. We choose the class scaling factor $w_i = 1 - n_i/n_{tot}$, being n_i the number of pixels with label i within \mathbf{y}_s , and n_{tot} the total number of annotated pixels. To avoid loss contribution on unlabeled pixels, we multiply the result by the masking function $\mathbb{1}(\mathbf{y}_s)$, which returns 1 for annotated pixels, 0 otherwise. A similar formulation was suggested in (Tang, Djelouah, et al., 2018) termed as Partial Cross-Entropy (PCE) loss but without the class balancing. Thus, we term our formulation as Weighted-PCE (WPCE).

Adversarial Cost

Adversarial objectives are the result of a minimax game (Goodfellow et al., 2014) between segmentor and discriminator, where $\Delta(\cdot)$ is trained to maximize its capability of differentiating between real and generated segmentations, $\Sigma(\cdot)$ to predict segmentation masks that are good enough to trick the discriminator and minimize its performance.

To address the difficulties of training GANs, that can lead to training instability (Mao et al., 2018), we adopt the Least Square GAN objective (Mao et al., 2018) which penalizes prediction errors of the discriminator based on their distances from the decision boundary.

Given an image \mathbf{x} and an unpaired mask \mathbf{y} , we optimize Δ and Σ according to: $\min_{\Delta} \mathcal{V}_{LS}(\Delta)$ and $\min_{\Sigma} \mathcal{V}_{LS}(\Sigma)$, where:

$$\begin{aligned}\mathcal{V}_{LS}(\Delta) &= \frac{1}{2} E_{\mathbf{y} \sim p(\mathbf{y})} [(\Delta(\mathbf{y}) - 1)^2] + \frac{1}{2} E_{\mathbf{x} \sim p(\mathbf{x})} [(\Delta(\Sigma(\mathbf{x})) + 1)^2] \\ \mathcal{V}_{LS}(\Sigma) &= \frac{1}{2} E_{\mathbf{x} \sim p(\mathbf{x})} [(\Delta(\Sigma(\mathbf{x})) - 1)^2].\end{aligned}\tag{6.2}$$

Training Strategy

We iterate the training of the model over two steps: i) optimization over a batch of weakly annotated images, and ii) optimization over a batch of unlabeled images.

When scribble annotations are available, we minimize $\mathcal{L} = a_0 \mathcal{L}_{SUP} + a_1 \mathcal{V}_{LS}(\Sigma)$. In particular, we compute a_0 *dynamically*, so that we don't need to tune it. We define: $a_0 = \frac{\|\mathcal{V}_{LS}(\Sigma)\|}{\|\mathcal{L}_{SUP}\|}$ to maintain a fixed ratio between the amplitude of supervised and adversarial costs throughout the entire training process, preventing one factor to prevail over the other. We report a study of the dynamic weighting effect in Appendix B.3

When dealing with a batch of unlabeled images, we alternately optimize the model. First, we compute the discriminator loss, $a_2 \mathcal{V}_{LS}(\Delta)$, and update discriminator's weights to reduce it. Then, with the updated discriminator, we estimate the generator loss, $a_3 \mathcal{V}_{LS}(\Sigma)$, and optimize the generator's weights.

We give more importance to the supervised objective rather than the adversarial loss because the discriminator only evaluates if the predicted masks look realistic, while it does not say anything about their accuracy. Besides, the supervised cost requires the segmentor to learn the correct mapping from images to segmentation masks, which is what we are interested into. Thus, we scale the adversarial contribution to be one order of magnitude smaller, setting $a_1 = 0.1$ for training with weak supervision. Similarly, we use $a_2 = a_3 = 0.2$ to train generator and discriminator

equally on the unlabeled data.

We minimize the loss function using Adam (Kingma and Ba, 2015) and a batch size of 12. Most importantly, learning from limited annotations can easily trap the model in sharp, bad, local minima because the training data poorly represents the actual data distribution. Thus, we promote the search of flat and more generalizable solutions using a cyclical learning rate (Smith, 2017) with a period of 20 epochs, that we oscillate between 10^{-4} and 10^{-5} . As a result, we observed a smoother loss function and more stable performance between subsequent epochs, diminishing the early stopping criterion effects (as also observed in (Valvano, Chartsias, et al., 2019)). Similarly to previous work with weak annotations (Lin, Dai, et al., 2016; Dai, He, and Sun, 2015), we train the model until an early stopping criterion is met, and we arrest the training when the loss between predicted and real segmentations stops decreasing on a validation set.

6.4 Experimental Setup

6.4.1 Data

For the experiments, we adopted the medical datasets: ACDC (described in Section 3.8.1), LVSC (Section 3.8.2), and both T1 and T2 images from CHAOS, separately (Section 3.8.4). To demonstrate the broad utility of our method, we also test performance on the (non-medical) PPSS (Section 3.8.5) dataset, which focuses on human pose parts segmentation.

Below, we first detail the procedure used to generate scribble annotations; then, we define how we construct train, validation, and test set.

Scribble Generation

To obtain scribbles with these datasets we follow different processes. Examples of those scribbles are shown in Figure 30. Experts draw scribbles in a certain way (e.g., away from border regions). A dataset containing manual scribbles helps test a method more realistically than using

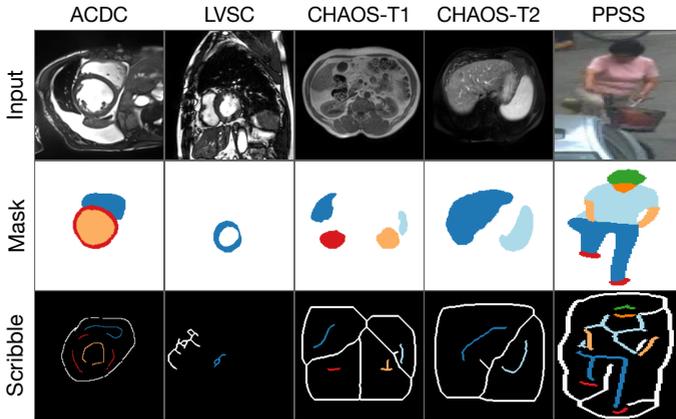


Figure 30: Example of generated scribbles for each dataset. In ACDC, scribbles were manually annotated inside the available segmentation masks. In CHAOS and PPSS, we obtained scribbles for each class via binary erosion of the associated segmentation mask, as in (Rajchl et al., 2017). In LVSC, the binary erosion would result in a very good approximation of the myocardium: thus, we generated scribbles with a random walk inside of each class. Please, refer to Section 6.4.1 for additional details.

simulated data from automatic procedures. Thus, in ACDC, we use ITK-SNAP (Yushkevich et al., 2006) to manually draw scribbles for ES and ED phases within the available segmentation masks. We obtained separate scribbles for RV, LV, and MYO, enabling us to test against ground truth segmentations. To identify pixels belonging to the background class (BGD), we draw an ulterior scribble approximately around the heart, while leaving the rest of the pixels unlabeled. Scribbles for RV, MYO, LV, BGD had an average (standard deviation) image coverage of 0.1 (0.1)%, 0.2 (0.1)%, 0.1 (0.1)% and 10.4 (8.4)%, respectively.

For CHAOS and PPSS, we obtained scribbles by eroding the available segmentation masks (Rajchl et al., 2017). For each object, we followed standard skeletonisation by iterative identification and removal of border pixels, until connectivity is lost. Resulting scribbles are deterministic, typically falling along the object’s midline (as with manual ones (Lin, Dai, et al., 2016)).

For LVSC, since MYO is thin, a skeleton is already too good of an approximation of the full mask. Thus, we generate scribbles with random walks. For every object, we first initialize an “empty” scribble, and define the 2D coordinates of a random pixel $P \equiv (x_P, y_P)$ inside the segmentation mask. Then, we iterate 2500 times the steps: i) assign P to the scribble; ii) randomly “move” in the image, adding or subtracting 1 to the coordinates of P ; iii) if the new point belongs to the segmentation mask, assign the new coordinates to P . Scribbles for MYO and BGD had an average (standard deviation) image coverage of 0.2 (0.1) % and 1.9 (0.5) %, respectively.

Train, Validation, Test

We divided ACDC, LVSC, CHAOS-T1 and CHAOS-T2 datasets in groups of 70%, 15% and 15% of patients for train, validation, and test set, respectively. Following seminal semi-supervised learning approaches (Salimans et al., [2016]; Chartsias, Joyce, et al., [2019]), we additionally split the 70% of training data into two halves, the first of which is used to train the segmentor $\Sigma(\cdot)$ with weak labels (image-scribble pairs), while we use *only* the masks of the second half to train the discriminator $\Delta(\cdot)$. Correlations between groups are limited by: i) splitting the data by patient, rather than by images (limiting intra-subject leakage, as masks come from different subjects (Chartsias, Joyce, et al., [2019])); and ii) discarding images associated to masks used to train the discriminator (thus, $\Sigma(\cdot)$ never sees images used to train $\Delta(\cdot)$).

For PPSS, following Luo, Wang, and Tang, [2013], we use the video scenes from the last 71 cameras as test set, while we split images from the first 100 cameras to train (90% of images) and validate (10% of images) the model. As with the medical datasets, we further divide the training volumes into two halves, and we use one of them to exclusively train the discriminator, using the segmentation masks and discarding the associated images.

6.4.2 Baseline, Benchmark Methods and Upper Bounds

We evaluate the robustness of our method in terms of segmentation performance compared with methods using different prior assumptions to regularise training with scribbles, summarized in Table 5. In particular, we consider:

- **UNet_{PCE}** and **UNet_{WPCE}** (Tang, Djelouah, et al., 2018): The UNet (Ronneberger, Fischer, and Brox, 2015) is one of the most common choices for training with fully annotated segmentation masks. We evaluate its behavior when trained with the PCE loss proposed for scribble supervision in (Tang, Djelouah, et al., 2018), or the WPCE loss introduced in Equation 6.1
- **UNet_{CRF}**: We also consider the previous UNet_{WPCE} whose prediction is further processed by CRF as RNN layer (Chen, Papandreou, et al., 2017; Zheng et al., 2015), (Monteiro, Figueiredo, and Oliveira, 2018). CRF as RNN models Conditional Random Fields as a recurrent neural network (RNN), incorporating the prior that nearby pixels with similar color intensities should be classified similarly in the segmentation mask. This layer can be trained end-to-end and does not require relabeling the training set. For ACDC and LVSC, we train such a layer with the same hyperparameters used for cardiac segmentation in (Can et al., 2018): $\sigma_\alpha = 160$, $\sigma_\beta = 3$ and $\sigma_\gamma = 10$. These parameters model the pairwise potentials of CRF as weighted Gaussians (Zheng et al., 2015). As in (Can et al., 2018), we use 5 iterations for the RNN. For the other datasets, we set $\sigma_\gamma = 3$, as suggested in (Zheng et al., 2015).
- **TS-UNet_{CRF}**: We compare our model to the two-steps procedure in (Can et al., 2018), using the variant modeling CRF as an RNN rather than a separate post-processing step, because no relevant difference was observed between the two, and this is simpler to use at inference. For the CRF as RNN, we used the same hyper-parameter setting of UNet_{CRF}.

The above approaches do not exploit unpaired data during training.

Thus, we also compare with two models that, despite not being proposed for weakly supervised learning, can exploit the extra unpaired data and learn data-driven shape priors:

- **PostDAE** (Larrazabal et al., 2020): this method trains a Denoising Autoencoder (DAE) on unpaired masks, and then uses it to post-processes the predictions of a pre-trained UNet. To train the UNet on scribbles and directly compare with our method, we use the WPCE loss.
- **UNet_D**: as in vanilla GANs, we train a UNet segmentor and a mask discriminator. The latter has the same architecture as ours (same capacity), but it receives inputs only at the highest resolution.

Lastly, we compare with the method of Zhang et al.:

- **ACCL** (Zhang, Zhong, and Li, 2020): similar to UNet_D, ACCL trains with scribbles using a PatchGAN discriminator (Isola et al., 2017).

Finally, we consider two **upper bounds**, based on training with fully annotated segmentation masks:

- **UNet^{UB}**: UNet trained with strong annotations. In this case, we train the UNet in a fully-supervised way using image-segmentation pairs and a weighted cross-entropy loss (with per-class weights defined as in Equation 6.1).
- **UNet_D^{UB}**: UNet as before, but with an additional vanilla mask discriminator, used to train on the unlabeled images. The discriminator is the same as that of our model, but it receives an input only at the highest resolution.

To compare methods, we always use same UNet segmentor, learning rate, batch size, and early stopping criterion. If a method does not use a discriminator, we simply discard the data we would have used to train $\Delta(\cdot)$. As Can et al., 2018, we train the CRF as RNN layer of TS-UNet_{CRF} with a learning rate 10^4 times smaller than that used for the UNet training, and we update the RNN weights only every 10 iterations.

Model	Uses Prior	Type of Prior
UNet _{PCE}	✗	–
UNet _{WPCE}	✗	–
UNet _{CRF}	✓	Mean Field Assumption (Zheng et al., 2015)
TS-UNet _{CRF}	✓	Mean Field Assumption (Zheng et al., 2015)
PostDAE	✓	Shape, via DAE
UNet _D	✓	Shape, via Discriminator
ACCL	✓	Shape, via Patch Discriminator (Isola et al., 2017)
Ours	✓	Multi-scale Shape, via AAGs

Table 5: Type of prior used by each model.

Evaluation

We measure performance with the multi-class Dice score: $Dice = \frac{2|\tilde{y} \cdot y|}{|\tilde{y}| + |y|}$, where \tilde{y} and y are the multi-channel predicted and true segmentation, respectively. To assess if improvements are statistically significant we use the non-parametric Wilcoxon test, and we denote statistical significance with $p \leq 0.05$ or $p \leq 0.01$ using one (*) or two (**) asterisks, respectively. We avoid multiple comparisons comparing our method only with the best benchmark model.

6.5 Experiments and Discussion

We present and discuss the performance of our method in various experimental scenarios. Our primary question is: *Can scribbles replace per-pixel annotations* (Section 6.5.1, 6.5.2); *and what happens when we have fewer scribble annotations, or less unpaired data* (Section 6.5.3, 6.5.4)? Then, we consider two natural questions that extend the applicability of our approach: *Can we learn from multiple scribbles per training image* (Section 6.5.5)? *Can we mix per-pixel annotations with scribbles during training* (Section 6.5.6)? Finally, we ask: *Why does Adversarial Attention Gating work* (Section 6.5.7)?

6.5.1 Learning from Scribbles

A prime contribution of our work is to close the performance gap between the most common strongly supervised models and weakly supervised approaches. Thus, we compare our method with other benchmarks and upper bounds quantitatively, in Table 6 and qualitatively, in Figure 31.

In particular, Table 6 reports average and standard deviation of the Dice score on test data for each dataset³. We clarify that, as discussed in Section 6.4.1, these results refer to training the segmentors with half of the annotated training images. We report Dice scores and the Hausdorff distances for each anatomical region of the medical datasets in Appendix B.1.

Our method matches and sometimes even improves the performance of approaches trained only with strong supervision. As an example, we improve the Dice score of UNet^{UB} on both ACDC and PPSS. A result that further confirms the potential of weakly supervised approaches that use annotations which are much easier to collect than segmentation masks.

Moreover, as can be seen from the upper part of the table (methods trained with scribble supervision), we consistently improve segmentation results⁴. When compared to the 2nd best model, we obtain up to ~8.5% of improvement on CHAOS-T1. As our ablation study shows in Section 6.5.7, such performance gains originate from the multi-scale interaction between adversarial signals and attention modules, which regularises the segmentor to predict both locally and globally consistent masks. In particular, our training strategy enforces multi-scale shape constraints, discouraging the appearance of isolated pixels and unrealistic spatial relationships between the object parts (Fig. 31).

Interestingly, we observe that weighting the loss contribution of each class based on their numerosity (UNet_{PCE} vs UNet_{WPCE}) is not always beneficial to the model, probably because, being sparse, scribble super-

³For ACDC, we also evaluated our model using the challenge server. After training our method on scribbles, we obtained an average (over the anatomical regions) Dice of 86.5%. We report the full results in Table 14, Appendix B.2.

⁴The only exception is on LVSC, where we have same results as ACCL.

		Dataset					
		Model	ACDC	LVSC	CHAOS-T1	CHAOS-T2	PPSS
Supervision Type	Scribble	UNet _{PCE}	79.0 ₀₆	62.3 ₀₉	34.4 ₀₆	37.5 ₀₆	71.9 ₀₄
		UNet _{WPCE}	69.4 ₀₇	59.1 ₀₇	40.0 ₀₅	<u>52.1₀₅</u>	69.3 ₀₄
		UNet _{CRF}	69.6 ₀₇	60.4 ₀₈	40.5 ₀₅	<u>44.7₀₆</u>	68.8 ₀₄
		TS-UNet _{CRF}	37.3 ₀₈	50.5 ₀₇	29.3 ₀₅	27.6 ₀₅	67.1 ₀₄
		PostDAE	69.0 ₀₆	58.6 ₀₇	29.1 ₀₆	35.5 ₀₅	67.5 ₀₄
		UNet _D	61.8 ₀₈	31.7 ₀₉	44.0 ₀₃	46.3 ₀₁	73.1 ₀₄
		ACCL	82.6 ₀₅	65.9₀₈	48.3 ₀₇	49.7 ₀₅	73.2 ₀₄
		Ours	**84.3₀₄	<u>65.5₀₈</u>	*56.8₀₅	57.8₀₄	**74.6₀₄
Mask	UNet ^{UB}	82.0 ₀₅	67.2 ₀₇	60.8 ₀₆	58.6 ₀₁	72.8 ₀₄	
	UNet _D ^{UB}	83.9 ₀₅	67.9 ₀₉	63.9 ₀₅	60.8 ₀₁	77.2 ₀₄	

Table 6: Dice average and standard deviation (subscript) obtained from each method on the test set, for medical and vision datasets. Leftmost column indicates if the learning algorithm has been trained with full mask or scribble annotations. The best method is in bold characters, while the second best is underlined; asterisks denote if their difference has statistical significance ($* p \leq 0.05$, $** p \leq 0.01$).

vision suffers less than mask supervision from the class unbalance problem. However, when the class imbalance increases, e.g. with CHAOS-T1 and T2, weighting the PCE seems to be beneficial. We also did not find evident performance boost in using CRF as RNN to post-process the UNet predictions (UNet_{WPCE} vs UNet_{CRF}).

The two-step paradigm of TS-UNet_{CRF} is one of the worst. We observed that errors reinforce themselves in self-learning schemes (Chapelle, Scholkopf, and Zien, 2009), and unreliable proposals in the relabeled training set lead the retrained model to fit to errors.⁵

Lastly, we discuss the performance of the methods that learn a shape prior from the unpaired masks. As Table 6 shows, post-processing the

⁵In this experiment, we explore the learning capability of the model and compare with benchmarks on the same ground. Thus, we did not enlarge scribbles as suggested by Can et al., 2018 (Grady, 2006). With the enlarged scribbles, TS-UNet_{CRF} improved from 37.3% to 53.6%, on ACDC. Doing the same for our method, gave no improvement (83.5% vs 84.3% from Table II). This illustrates that such additional training signal is useful for TS-UNet_{CRF} but it is not necessary for our method. While we are not certain about the origins of this, we hypothesise that it is the adversarial discriminator that provides a similar training signal as those provided by the enlarged scribbles.

segmentor output with a DAE does not improve performance (PostDAE). As discussed by the PostDAE authors (Larrazabal et al., 2020), a reason could be the poor performance of the segmentor which, when trained on scribbles, produces out-of-distribution segmentation masks for the DAE (i.e., the corrupted data used for training the DAE are not representative of the test-time segmentation errors). Sometimes, we even observed degenerate cases where the PostDAE always produces empty masks (CHAOS dataset and PPSS), or it completely omits some classes (ACDC). See Appendix B.5 for visual examples of these and other models' failures.

Instead, mask discriminators are an effective choice (see UNet_D and ACCL). In fact, the discriminator can recover missing label information from the scribble-annotated data, and the model has competitive performance. However, our model generalises better across datasets.

6.5.2 Segmentation Masks vs Scribbles

To understand the trade-off between the time-to-segment and the type-of-annotations, we evaluate if it's better to collect many scribble annotations instead of few fully annotated images. Assuming that similar to bounding boxes (Lin, Maire, et al., 2014), scribbles can be collected about $15\times$ faster than segmentations, annotating 35 images with scribbles on ACDC would require a similar time as two densely labelled masks. Some authors suggest the possibility to learn to segment using a few or even one single annotated sample (Tajbakhsh et al., 2020; Shaban et al., 2017; Zhao, Balakrishnan, et al., 2019; Chaitanya, Karani, et al., 2019; Liu, Lee, et al., 2019; Feyjie et al., 2020). Thus, we want to compare the performance of our model using 35 scribble-annotations (Dice of 84.3%) with that obtainable using two full masks and the Task-driven and Semi-supervised Data Augmentation (TSDA) method (Chaitanya, Karani, et al., 2019).⁶ TSDA uses a GAN to learn realistic deformations and intensity transformations to apply on the annotated images and uses the augmented training set to optimise a UNet-like segmentor. We perform

⁶We used the code provided by the authors at https://github.com/krishnabits/001/task_driven_data_augmentation.

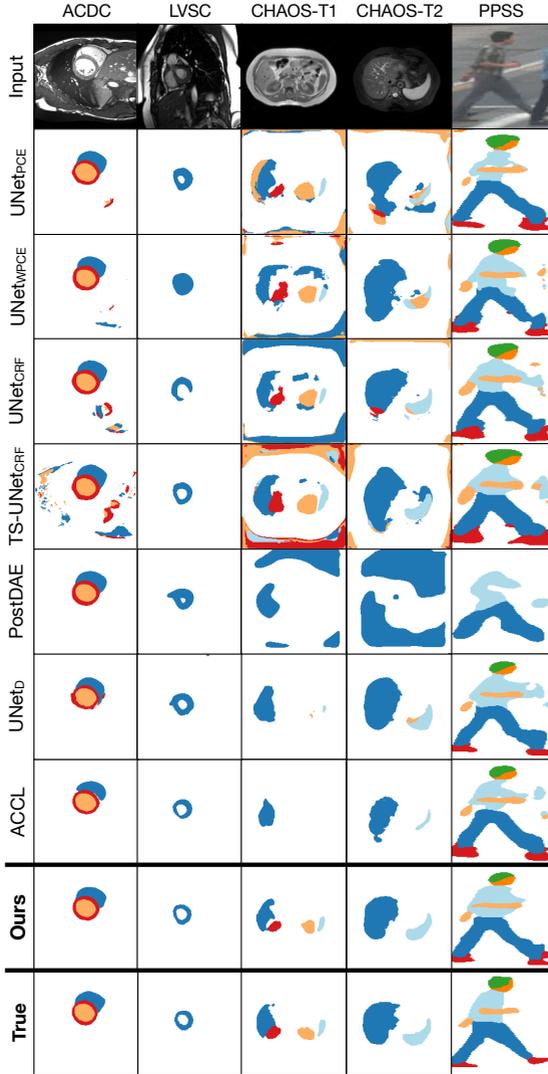


Figure 31: Example of predicted segmentation masks for the considered methods on each task. Observe that our approach (bottom row) learns spatial relationships in the image, thus preventing the prediction of isolated pixels in the mask, as well as unrealistic spatial relationship among the object parts.

3-fold cross-validation, using the same validation and test sets as before. We randomly selected two fully-annotated patients among the training subjects, and we learned the augmentation GAN with the unpaired images we assumed available (35 patients). With TSDA, we obtained an average Dice (standard deviation) of 56.8% (13.5%), which is considerably better than the standard training of a segmentor (Dice of 24.9% (14.1%)) but worse than other models trained with all the 35 scribble-annotated data (ACDC column, Table 6).

Our results confirm recent findings (Asano, Rupprecht, and Vedaldi, 2020) observing that despite a single image can be enough to train the first few layers of a CNN, deeper layers require additional labels.

Lastly, notice that TSDA data augmentation can be potentially integrated into our model, too.

6.5.3 Model Robustness to Limited Annotations

We analyze model robustness with a scarcity of annotations in Figure 32. In particular, we compare with methods that don't employ shape priors during training. In the experiments, we always use 50% of training data to exclusively train the discriminator, if present in the method. The remaining 50% is used to train the segmentor $\Sigma(\cdot)$, with varying amount of labels: e.g. "5%" means we train $\Sigma(\cdot)$ with 5% of labeled and 45% of unlabeled images (adversarial setup). As upper bound, we consider the results obtained by UNet_D^{UB} after training it with all the available image-segmentation pairs.

As shown in Figure 32, our model can rapidly approach the upper bound and, overall, it shows the best performance for almost every percentage of training annotations. With 5% of weakly annotated data, our method performs slightly worse than other models in LVSC and CHAOS: however, the performance gap is not statistically significant.

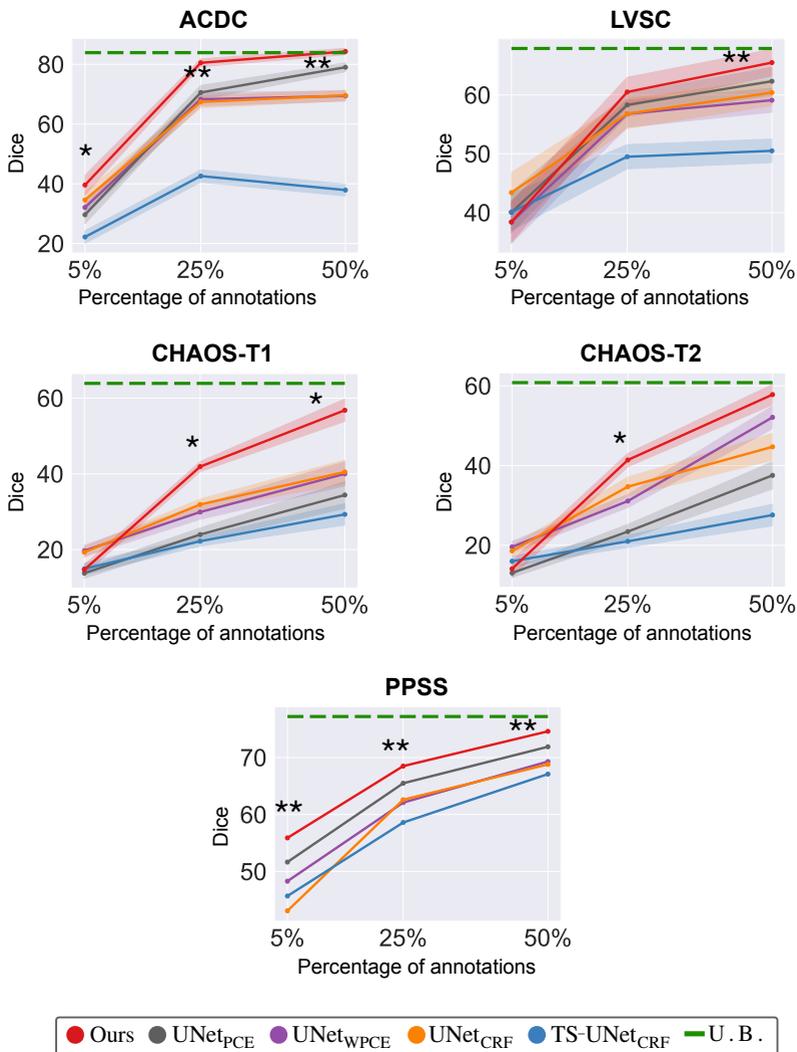


Figure 32: Dice score obtained on the test data by our and methods that don't use shape priors when changing the percentage of available labels in the training set (shaded bands show standard errors instead of deviation for clarity). As upper bound (U.B.) we consider UNet_D^{UB}, trained using all the densely annotated masks. Asterisks denote if differences between first and second best has statistical significance (* $p \leq 0.05$, ** $p \leq 0.01$).

6.5.4 How Much Does the Model Rely on the Unpaired Data?

Here, we investigate how much the model relies on the unpaired data by reducing the number of unpaired masks first, then the unpaired images. In the first case, we trained the discriminator using only 5% of the unpaired masks (3 ACDC patients) and the segmentor using all the scribbles. Despite training $\Delta(\cdot)$ with less masks, thanks to data augmentation (random roto-translations and instance noise), the model learned a robust shape prior and got a Dice of 83.7% (5%), i.e. less than 1% decrease. Thus, the adversarial conditioning of the attention gates was still strong enough to correctly bias the segmentor to learn multi-scale relationships in the objects.⁷ Secondly, we repeated the experiments in Section 6.5.3 training our model without the additional unpaired images, and by varying the number of annotated data from 5% to 50%. At 5% of annotations, we obtained an average (standard deviation in parenthesis) Dice of 22.5% (10%); with 25% of scribble-annotations, a Dice of 75.0% (8%); and with 50% of labels, we got 84.3% (4%). As can be seen, the model dependence on the number of unpaired images decreases when the number of scribble-annotated images (that are easy to collect) increases.

Based on these experiments, we conclude that the model performs well even when the unpaired data are scarce, provided that enough scribble-annotations are available.

6.5.5 Combining Multiple Scribbles: Simulating Crowdsourcing

Here we investigate the possibility to train our model using multiple scribbles per training image. This scenario simulates crowdsourcing applications, which are useful for annotating rare classes and to exploit dif-

⁷We conducted experiments also using more than 5% of masks. Overall, we observed similar performance, with some fluctuations in Dice score due to the optimisation process. Such fluctuations originate from several factors: weight initialisation, training data, stochastic order of the batches presented to the network during training, etc. Minimising the performance gap between best- and worst-case scenario is a well-known problem of weakly supervised learning, and an active area of research (Guo, Li, et al., 2019).

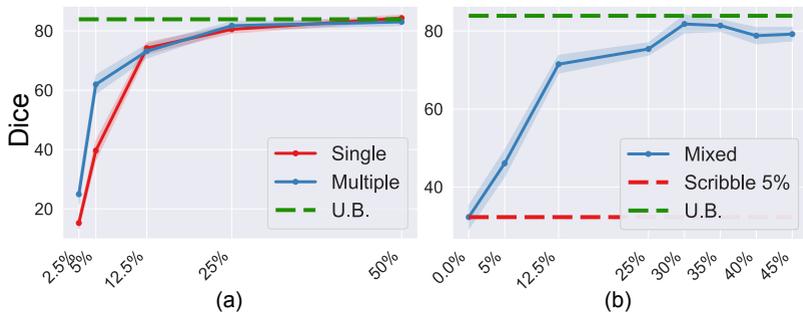


Figure 33: (a) Effect of training with labels from multiple annotators; and (b) performance in presence of mixed supervision (i.e. using masks and scribbles) on ACDC. The upper bound (U.B.) is the $\text{UNet}_D^{\text{UB}}$, trained with all the dense segmentation masks.

ferent levels of expertise in annotators (Lin, Maire, et al., 2014; Ørting et al., 2019). We mimic the scribble annotations collected by three different “sources”, using: i) expert-made scribbles; ii) scribbles approximated by segmentation masks skeletonization; iii) scribbles approximation by a random walk in the masks (see Section 6.4.1 for a description of the approaches used to generate ii) and iii)).

For every training image, we combine multiple scribbles summing up the supervised loss Equation 6.1 obtained for each of them: $\mathcal{L}_{SUP} = \sum_{i=1}^3 \mathcal{L}_{SUP}^i$. Thus, we consider multiple times pixels that are labeled across annotators, while considering ‘once’ pixels labeled only from one annotator. Other ways of combining annotations are also possible (e.g., considering the union of the scribbles, or weighting differently each annotator (Ørting et al., 2019)), but they are out of the scope of this chapter.

In Figure 33a, we compare the Dice score of our method trained in a “single” vs a “multiple” annotator scenario. As can be seen, multiple scribbles have a regularising effect when the number of annotated data is scarce.

6.5.6 Multitask Learning: Combining Masks and Scribbles

Collecting homogeneous large-scale datasets can be difficult, but often we have access to multiple data sources, that can have different types of annotations. Here, we relax the assumption of using only scribble annotations, and investigate if we can train models that also leverage extra fully annotated data. For simplicity, we assume to have 5% of scribble annotations, and we gradually introduce from 0% to 45% of fully-annotated images (for a maximum total of 50% annotated data). We train the model using as loss: Equation 6.1 for scribble-annotated data, Equation 6.2 for unlabeled data, and the weighted cross-entropy for fully annotated images. We report results on ACDC in Figure 33b, showing that mixing scribble and mask supervision is feasible, and it can increase model performance. Although training only with masks is beyond the scope of this chapter, we also investigated training in a fully supervised full mask setting. As expected, results show that training using only masks further improves segmentation performance (we report numbers in Appendix B.4).

6.5.7 Why does Adversarial Attention Gating work?

Prior-conditioned Attention Maps are Object Localizers We will now show that – contrary to canonical attention gates – AAGs act as object localizers at multiple scales. In detail, we consider our attention mechanism with or without the adversarial conditioning (ADS). In both cases, the probability attention map is obtained as in Section 6.3.2, and results from a 1×1 convolutional layer with softmax activation (that can be interpreted as a classifier), and a sum operation on all but one channel (see a summary in Figure 29). In Figure 34 we illustrate: i) the most active channels in the classifier output, and ii) the predicted attention maps, at multiple depth levels d . As the attentions maps show (Fig. 34, top), the adversarial conditioning of the attention gates encourages the segmentor at multiple scales to i) learn to localize objects of interest; and ii) suppress activations outside of them. Thus, scattered false positives (see

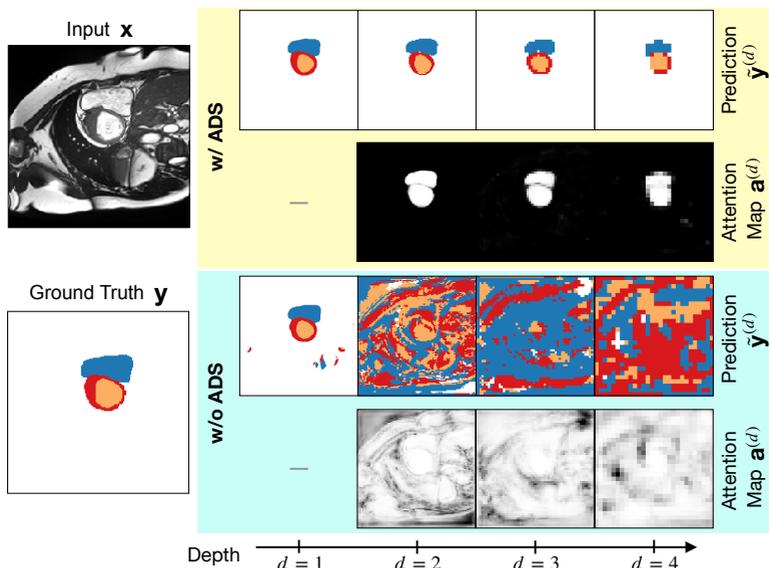


Figure 34: UNet-like segmentor with (top) vs without (bottom) adversarial conditioning of the attention gates in its decoder. Conditioned by an adversarial shape prior (w/ ADS), the model learns semantic attention maps able to localize the object to segment at multiple scales. Also, the shape prior encourages the segmentor to learn multi-scale relationships in the objects.

UNet’s prediction for $d = 1$ in Figure 34 are prevented, and the model performance improves (see also Figure 31).

Adversarial Attention Gating Trains Deep Layers Better We qualitatively show that AAGs increase the training of the segmentor deepest layers. In Figure 35, we show the distribution of weights values in the convolutional layers at depth $d = 4$ in absence vs presence of adversarial conditioning (ADS) of the attention gates. As shown, attention gates with ADS force the segmentor to update its weights also in deeper layers, which would otherwise suffer from vanishing gradients (Szegedy, Liu, et al., 2015; Lee, Xie, et al., 2015).

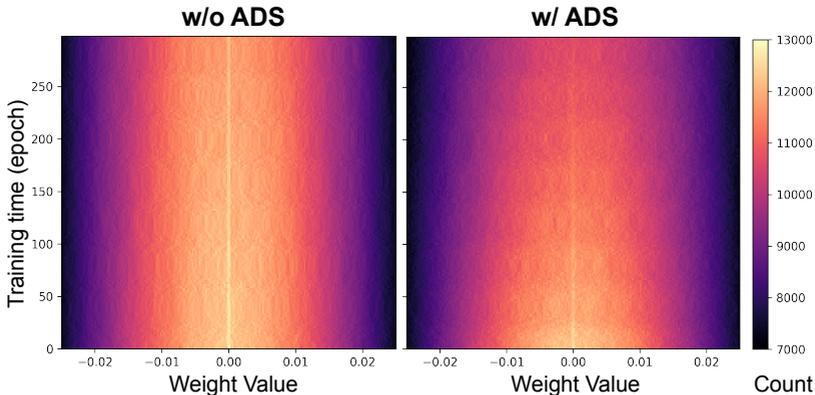


Figure 35: Weight distribution for the convolutional layers at depth $d=4$ of the segmentor. We compare how the weight distribution changes during training, with and without the use of ADS on the segmentor. Notice that ADS helps the layer training, and the initially narrow distribution becomes broader in time.

Ablation Study We show ablations on ACDC in Table 7. Removing ADS from the model, we leave the discriminator as a vanilla one, receiving inputs only at the highest resolution (classic GAN), while the segmentor remains unchanged. Unless otherwise stated, removing ADS we leave the attention gates in the segmentor, but without the adversarial conditioning (i.e. the segmentor is a UNet with classical self-attention; see Figure 29). When we completely remove the discriminator, the segmentor is trained just with scribble supervision and no adversarial signals. As Table 7 shows, each model component contributes to the final performance.

In particular, Table 7 highlights that our model’s success is not merely due to the use of additional unpaired images. In fact, if we compare with a classic GAN that also uses extra unpaired images, we Dice increases of 23% when enough scribbles are available (compare “Ours” vs “#3” at 25% and 50% of labels).

From Table 7, we further observe that both ADS and the multiplicative gating are important aspects of the model, and they increase the seg-

	Attention	Discriminator		5%	25%	50%
		single	multi			
Ours	✓		✓	40.7 ₀₉	80.6 ₀₆	84.3 ₀₅
#1	✓		✓	38.4 ₁₃	79.1 ₀₆	83.8 ₀₄
#2	✓	✓		39.4 ₁₀	77.3 ₀₇	84.0 ₀₅
#3		✓		55.8 ₁₀	60.2 ₀₇	61.8 ₀₈
#4	✓			34.8 ₀₉	71.6 ₀₈	71.0 ₀₈
#5				32.1 ₀₉	68.3 ₀₉	69.4 ₀₇

Table 7: Our ablations, as the name states, start with our model but remove: #1: Only gating; #2: Only ADS; #3: Both Gating and ADS; #4: Both ADS and the Discriminator; and finally #5: ADS, the Discriminator and Gating.

mentation quality of a similar amount (e.g., going from the ablation “#3” to “#2”, or to “#1”, we obtain similar performance gains). This is not surprising: in fact, both the approaches enforce an attention process inside the segmentor. Specifically, the gating does so because it acts as an information bottleneck on what gets transmitted to the next convolutional block (i.e. it zeroes out unimportant information in the features maps). The ADS also enforces attention since it forces the segmentor to extract the information needed to predict realistic segmentations at every resolution. However, it is evident that ADS and the gating mechanism bring complementary advantages to the model, and it is when we combine *both* of them that we reach the best results, at every percentage of labels (“Ours” vs “#2”, “Ours” vs “#3”).

Finally, we compared the use of the PCE vs WPCE loss to train the full model. With PCE, we obtained a Dice of: 25.2 (11), 74.0 (7), 83.4 (5) for 5%, 25% and 50% of labels, respectively. With WPCE, our method performs better. We believe that this happens because PCE is intrinsically biased to penalize more the errors of the class having more annotated pixels. On the contrary, the WPCE loss is invariant to the number of annotated pixels. Thus, with WPCE, the discriminator can more easily bias the segmentor to predict masks which reflect the expected ratio between the organs/parts sizes and make them look realistic, ultimately improving segmentation performance.

6.6 Conclusion

We introduce a novel strategy to learn object segmentation using scribble supervision and a learned multi-scale shape prior. In an adversarial game, we force a segmentor to predict masks that satisfy short- and long-range dependencies in the image, narrowing down or eliminating the performance gap from strongly supervised models on medical and non-medical datasets. Fundamental to the success of our method are the proposed generalization of deep supervision and the novel adversarial conditioning of attention modules in the segmentor.

We show the robustness of our approach in diverse training scenarios, including: a varying number of scribble annotations in the training set, multiple annotators for an image (crowdsourcing), and the possibility to include fully annotated images during training. In the future, it would be interesting to explore the introduction of other types of multi-scale shape priors, such as those obtained by multi-scale VAEs, which can take into account also segmentation uncertainty. Furthermore, it would be exciting to study other variants of the proposed attention gates, without relying on multiplicative gating operations and thus on background/foreground object segmentation tasks. It would also be interesting to explore the application of these gates for other tasks which could benefit from multi-scale adversarial signals, such as image registration (Krebs et al., 2019), conditional image generation (Azadi et al., 2019) and localised style transfer (Kurzman, Vazquez, and Laradji, 2019).

6.7 Summary

This chapter showed that GANs could learn powerful shape priors to regularise the learning driven by weak supervision. However, the presented approach has the limit of requiring a set of compatible segmentation masks for training. These masks must contain annotations for the exact same classes included in the weak labels. Moreover, to limit the covariate shift risk, we must ensure that the segmented structures are similar across datasets. For example, suppose we train the segmentor on

weakly annotated data obtained from an elderly population, and we use unpaired masks from a pediatric database to train the discriminator. In that case, we risk introducing a harmful data bias which may lead the model to under-segment the input images.

The next chapter shows that it is possible to introduce multi-scale relationships without requiring unpaired masks. In particular, we suggest using a self-supervised consistency objective to regularise the training with scribble annotations. Such a framework exhibits good performance levels while being more general and flexible than multi-scale GANs.

Chapter 7

Self-supervised Multi-scale Consistency for Weak Supervision

▣

We showed that it is possible to learn shape priors directly from data in an adversarial framework. Yet, we also discussed that there might be differences between the population used to train the segmentor and the one used to optimise the mask discriminator. Such differences are not necessarily a problem in semi-supervised learning, where paired image-segmentation masks regularise the segmentor training. However, they could introduce harmful biases in weakly supervised learning, where object shape and size information is not available within the labelled data.

This chapter presents a more general framework that extends the previously introduced multi-scale GAN and makes the multi-scale segmentor independent from unpaired masks' availability. In particular, we

This chapter is based on:

- Valvano, Gabriele, Andrea Leo, and Sotirios A Tsafaris (2021a). "Self-supervised Multi-scale Consistency for Weakly Supervised Segmentation Learning". In: *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Springer, pp. 14–24

adapt the Adversarial Attention Gates to learn multi-scale relationships using self-supervised training objectives. We show that the resulting method has a competitive performance compared to frameworks requiring unpaired masks to function, and it has the advantage of being mask annotation independent.

7.1 Introduction

To reduce the need for expensive annotation procedures, researchers have recently explored the use of weak annotations, such as scribbles (Can et al., 2018; Valvano, Leo, and Tsaftaris, 2021c), extreme points (Roth et al., 2020), and bounding boxes (Kervadec, Dolz, Wang, et al., 2020), to supervise model training. In these cases, we must limit the risk of overfitting and use regularisation. For example, we can constrain a segmentor to produce similar predictions when it receives similar inputs (Ouali, Hudelot, and Tami, 2020; Valvano, Chartsias, et al., 2019), or we can use prior knowledge about object shape (Kervadec, Dolz, Tang, et al., 2019; Zhou, Li, et al., 2019), intensity (Nosrati and Hamarneh, 2016), and position in the image (Kayhan and Gemert, 2020). As discussed in the previous chapters, GANs are a popular approach to regularise segmentors when lacking high-quality labels. They also proved to be effective in weakly supervised learning (Zhang, Zhong, and Li, 2020), where multi-scale adversarial shape priors led to state-of-the-art performance in various settings (Valvano, Leo, and Tsaftaris, 2021c).

However, GANs can be hard to optimise (Saxena and Cao, 2020), and they require a set of compatible segmentation masks for training. Annotated on images from a different data source, these masks must contain annotations for the exact same classes used to train the segmentor. Moreover, the structures to segment must be similar across datasets to limit the risk of covariate shift. For example, there are no guarantees that optimising a GAN using unpaired masks from a paediatric dataset will not introduce a harmful bias in a weakly supervised segmentor meant to segment elderly images.

As a result, multi-scale GANs are not always a feasible option. In

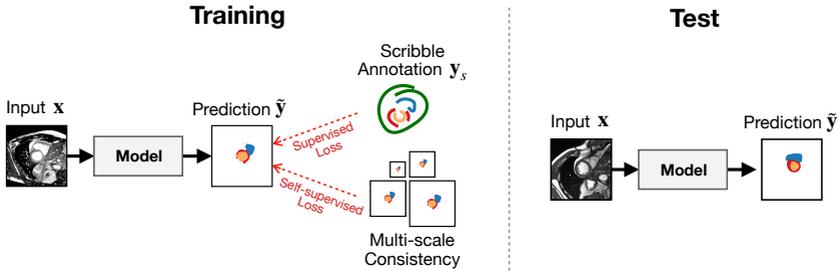


Figure 36: We train a segmentor using only scribble annotations as supervision. To regularise the model to produce realistic predictions, we introduce a self-supervised multi-scale consistency objective. Coupled with a customised attention gate, this objective biases the segmentor toward predicting masks satisfying short-range and long-range dependencies in the image, ultimately improving segmentation performance.

these cases, it would be helpful to introduce multi-scale relationships without relying on unpaired masks. This chapter shows that it is possible to do so without renouncing to obtain competitive performance levels. We summarise our main idea in Figure 36 and list our major contributions below.

7.1.1 Contributions

We present a novel self-supervised approach to introduce multi-scale shape consistency in a segmentor *without* relying on unpaired segmentation masks for training. Inspired by the success of the Adversarial Attention Gates introduced in the previous chapter, we train a shape-aware segmentor coupling multi-scale predictions and attention gates. However, we substitute the adversarial framework with a *mask-free* self-supervised objective that is simple to optimise. We show that our approach leads to comparable performance gains to that of GANs, but it removes the need for unpaired masks.

We release code used for the experiments at <https://vios-s.github.io/multiscale-pyqa/>.

7.2 Related Work

7.2.1 Weakly-supervised Learning for Image Segmentation

To help clinicians to annotate medical images more efficiently, recent research has explored the use of weak annotations to supervise models. Examples of weak annotations are: bounding boxes (Khoreva et al., 2017; Kervadec, Dolz, Wang, et al., 2020), image-level labels (Patel and Dolz, 2021), point clouds (Bearman et al., 2016; Roth et al., 2020; Qu et al., 2020), and scribbles (Lin, Dai, et al., 2016; Can et al., 2018; Ji et al., 2019; Dorent et al., 2020; Valvano, Leo, and Tsaftaris, 2021c). Although it is possible to extend the proposed approach to other types of weak annotations, herein, we focus on scribbles, which proved to be convenient to collect in medical imaging, especially when annotating nested structures Can et al., 2018.

A standard way to improve segmentation with scribbles is to rely on Conditional Random Fields (CRFs) to post-process model predictions (Lin, Dai, et al., 2016; Can et al., 2018; Ji et al., 2019). Recent work avoids the post-processing step and the need of tuning the CRF parameters by including learning constraints during training. For example Belharbi et al., 2020 uses a max-min uncertainty regulariser to limit the segmentor flexibility, while other approaches regularise training using global statistics, such as the size of the target region (Zhou, Li, et al., 2019; Kervadec, Dolz, Tang, et al., 2019; Kervadec, Dolz, Wang, et al., 2020) or topological priors (Kervadec, Dolz, Wang, et al., 2020). Although they increase model performance, these constraints' applicability is limited to specific assumptions about the objects and usually requires prior knowledge about the structure to segment. As a result, these methods face difficulty when dealing with pathology or uncommon anatomical variants. On the contrary, we do not make any strong assumption. We use a general self-supervised regularisation loss, optimising the segmentor to maintain multi-scale structural consistency in the predicted masks.

7.2.2 Multi-scale Consistency and Attention

Multi-scale consistency is not new to medical image segmentation. However, most deep learning methods either need strong annotations to supervise the segmentor at multiple levels (Dou, Yu, et al., 2017) or require training GANs using a set of compatible segmentation masks for training the discriminator (Zhang, Zhong, and Li, 2020; Valvano, Leo, and Tsaftaris, 2021c). In this work, we remove the necessity of having full masks for training. Instead, we impose multi-scale consistent predictions through an architectural bias localised at the level of attention gates within the segmentor.

Attention has been widely adopted in deep learning (Vaswani et al., 2017; Jetley et al., 2018) as it suppresses the irrelevant or ambiguous information in the features maps. Recently, attention was also successfully used in image segmentation (Li, Xiong, et al., 2018; Oktay, Schlemper, et al., 2018; Wang, Deng, et al., 2018; Schlemper et al., 2019; Fu et al., 2019; Sinha and Dolz, 2020). While standard approaches do not explicitly constrain the learned attention maps, Valvano, Leo, and Tsaftaris, 2021c have recently shown that conditioning the attention maps to be semantic increases model performance. In particular, they condition the attention maps through an adversarial mask discriminator, which requires a set of unpaired masks to work. Herein, we replace the mask discriminator with a more straightforward and general self-supervised consistency objective, obtaining attention maps coherent with the segmentor predictions at multiple scales.

7.2.3 Self-supervised Learning for Medical Image Segmentation

Self-supervised learning studies how to create supervisory signals from the data using pretext tasks. Pretext tasks are cheap surrogate objectives aimed at reducing human intervention requirements. Several tasks have been proposed for network pre-training, including image in/out-painting (Zhou, Sodha, et al., 2019), superpixel segmentation (Ouyang et al., 2020), coordinate prediction (Bai, Chen, et al., 2019), context restora-

tion (Chen, Bentley, et al., 2019) and contrastive learning (Chaitanya, Erdil, et al., 2020). After a self-supervised training phase, these models need a second-stage fine-tuning on the segmentation task. Unfortunately, choosing a proper pretext task is not trivial, and pre-trained features may not generalise well if unrelated to the final objective (Zamir et al., 2018). For this reason, our method is more similar to those using self-supervision to regularise training, at first modifying an input image and then encouraging feature prediction (Valvano, Chartsias, et al., 2019) or transformation consistency of the segmentor output (Xie et al., 2020).

7.3 Proposed Approach

Below, we first present an overview of the proposed approach. Then, we detail model architecture and training strategy.

7.3.1 Method Overview

We assume to have access to pairs of images x and their weak annotations y_s (in our case, y_s are scribbles), which we denote with the tuples (x, y_s) . We present a segmentor incorporating a multi-scale prior learned in a self-supervised manner. We introduce the shape-prior through a specialised attention gate residing at several abstraction levels of the segmentor. These gates produce segmentation masks as an auxiliary task, allowing them to construct semantic attention maps used to suppress background activations in the extracted features. As our model predicts and refines the segmentation at multiple scales, we refer to these attention modules as Pyramid Attention Gates (PyAG).

7.3.2 Model Architecture and Training

The segmentor $\Sigma(\cdot)$ is a modified UNet (Ronneberger, Fischer, and Brox, 2015) with batch normalisation (Ioffe and Szegedy, 2015). Encoder and decoder of the UNet are interconnected through skip connections, which propagate features across convolutional blocks at multiple depth levels d . We leave the encoder as in the original framework while we modify the

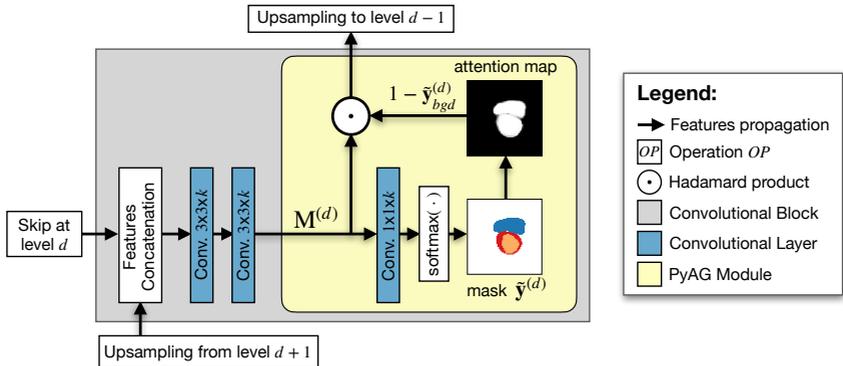


Figure 37: Detail of a decoding block at depth level d . The convolutional block processes the input features and predicts a low-resolution version of the segmentation mask $y^{(d)}$ as part of a PyAG attention module (represented in light yellow background). To ensure that the mask $\tilde{y}^{(d)}$ is consistent with the final prediction \tilde{y} , we use the self-supervised multi-scale loss described in Equation 7.1 and graphically represented in Figure 38. Using the predicted mask, we compute the probability of pixels belonging to the background and then suppress their activations in the features map $M^{(d)}$ according to Equation 7.2

decoder at each level, as illustrated in Figure 37. In particular, we first process the extracted features with two convolutional layers, as in the standard UNet. Next, we refine them with the introduced PyAG module, which we represent in light yellow background in Figure 37. Each PyAG module consists of: classifier, background extraction, and multiplicative gating operation. As classifier, we use a convolutional layer with c filters having size $1 \times 1 \times k$, with c the number of segmentation classes including the background, and k the number of input channels. Obtained an input features map $M^{(d)}$ at depth d , the classifier predicts a multi-channel score map that we post-process with a *softmax* operation. The resulting tensor assigns a probabilistic value between 0 and 1 to each spatial location. We make this tensor a lower-resolution version of the predicted segmentation mask using the self-supervised consistency constraint:

$$\mathcal{L}_{Self} = - \sum_{d=1}^n \sum_{i=1}^c \tilde{y}_i^{(0)} \log(\tilde{y}_i^{(d)}), \quad (7.1)$$

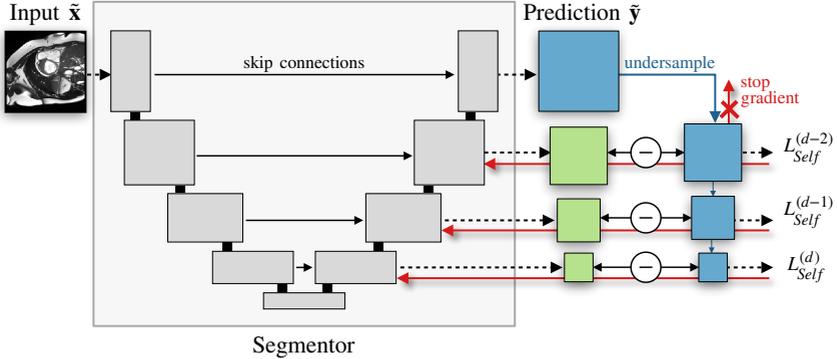


Figure 38: Self-supervised training of the segmentor. Thanks to PyAG modules, the model produces segmentation masks at multiple scales. We compare (\ominus symbol) the lower resolution masks (green squares) to those obtained undersampling the full resolution prediction \tilde{y} (blue squares). At each level, we compute a self-supervised loss contribution $\mathcal{L}_{Self}^{(i)}$ that we use as a regulariser. To prevent trivial solutions, we stop (\times symbol) gradients (red arrows) from propagating through the highest resolution stream.

where d is the depth level, i is an index denoting the class, $\tilde{y}^{(d)}$ is the prediction at depth d , and $\tilde{y}^{(0)} = \tilde{y}$ is the final prediction of the model.

To prevent hampering the final prediction, we propagate the self-supervised training gradients only through the attention gates and the segmentor encoder, as we graphically show in Figure 38. We further constrain the segmentor to reuse the extracted information by suppressing the activations in the spatial locations of the features map $M^{(d)}$ which can be associated with the background (Fig. 37). This multiplicative gating operation can be formally defined as:

$$M^{(d)} \leftarrow M^{(d)} \cdot (1 - \tilde{y}_{bkd}^{(d)}), \quad (7.2)$$

where $\tilde{y}_{bkd}^{(d)}$ is the background channel of the predicted mask at the depth level d . The extracted features are finally upsampled to the new resolution level $d - 1$ and processed by the next convolutional block.

To supervise the model considering the weak annotations, we use the Partial Cross-Entropy loss (Tang, Djelouah, et al., 2018) on the final pre-

diction $\tilde{\mathbf{y}}$. This formulation avoids loss contribution on the unlabelled pixels by multiplying the cross-entropy with a labelled pixel identifier $\mathbb{1}(\mathbf{y}_s)$. The role of the masking function $\mathbb{1}(\mathbf{y}_s)$ is to return 1 for annotated pixels, 0 otherwise. Mathematically, we formulate the weakly-supervised loss as:

$$\mathcal{L}_{PCE} = \mathbb{1}(\mathbf{y}_s) \cdot \left[- \sum_{i=1}^c \mathbf{y}_{s_i} \log(\tilde{\mathbf{y}}_i) \right], \quad (7.3)$$

with \mathbf{y}_s the ground truth scribble annotation.

Considering both weakly-supervised and self-supervised objectives, the overall cost function becomes: $\mathcal{L} = \mathcal{L}_{PCE} + a \cdot \mathcal{L}_{Self}$, where a is a scaling factor that balances training between the two costs. Similar to Valvano, Leo, and Tsaftaris, 2021c we find beneficial to use a dynamic value for a , which maintains a fixed ratio between supervised and regularisation cost. In particular, we set $a = a_0 \cdot \frac{\|\mathcal{L}_{Self}\|}{\|\mathcal{L}_{PCE}\|}$, where $a_0 = 0.1$ is meant to give more importance to the supervised objective. We minimise the loss using Adam optimiser (Kingma and Ba, 2015) with a learning rate of 0.0001, and a batch size of 12.

7.4 Experiments

Below, we first describe the used datasets. Then we present the benchmark models we compare with and report quantitative and qualitative results.

7.4.1 Data

We show the advantages of our method in the cardiac medical datasets: ACDC (described in Section 3.8.1) and LVSC (Section 3.8.2); in the abdominal organ dataset CHAOS (T1 images, Section 3.8.4); and finally, on the vision dataset on human part segmentation PPSS (Section 3.8.5).

The aforementioned datasets were released with fully-annotated segmentation masks. To test our approach’s advantages in weakly-supervised learning, we use the manual scribble annotations provided for the

ACDC dataset in Valvano, Leo, and Tsaftaris, [2021c]. For the remaining datasets, we follow the guidelines provided by Valvano, Leo, and Tsaftaris, [2021c] to emulate synthetic scribbles, using binary erosion operations (CHAOS and PPSS data) and random walks inside the available segmentation masks (LVSC).

Train, Validation, Test

We divided data from ACDC, LVSC, and CHAOS into groups of 70%, 15% and 15% of patients for train, validation, and test of the model, respectively. In PPSS, we follow the recommendation in Luo, Wang, and Tang, [2013] and use images from the first 100 cameras to train (90% of images) and validate (10% of images) our model. We use the remaining 71 cameras for testing it.

7.4.2 Evaluation Protocol

We evaluate the performance of our method – which we term $\text{UNet}_{\text{PyAG}}$ – in terms of segmentation performance. As benchmark models, we consider:

- **UNet**: The UNet (Ronneberger, Fischer, and Brox, [2015]) is one popular choice for fully-supervised training using segmentation masks. We evaluate its behaviour when trained on scribbles using the \mathcal{L}_{PCE} loss (Tang, Djelouah, et al., [2018]).
- **UNet_{Comp}**: We also consider the UNet segmentor whose training is regularised with the Compactness loss proposed by Liu, Dou, and Heng, [2020], which models a generic shape compactness prior. Such a prior is mathematically defined as: $\mathcal{L}_{\text{Comp.}} = \frac{P^2}{4\pi A}$, where P is the perimeter length and A is the area of the generated mask. The role of $\mathcal{L}_{\text{Comp.}}$ is to prevent the appearance of scattered false positives and negatives in the generated masks. As for our method, we dynamically rescale this regularisation term to be 10 times smaller than the supervised cost (Section [7.3]).

- **UNet_{CRF}**: Lastly, we consider post-processing the previous UNet predictions through CRF to better capture the object boundaries (Chen, Papandreou, et al., [2017]). The CRF uses weighted Gaussians to model the pairwise potentials between pixels, weighting the Gaussians with values ω_1 and ω_2 , and parametrising the distributions using appearance factors σ_α and σ_β , and smoothness factors σ_γ . For ACDC and LVSC datasets, we use the same parameters used for cardiac segmentation by Can et al., [2018], i.e.: $\sigma_\alpha = 2, \sigma_\beta = 0.1, \sigma_\gamma = 5, \omega_1 = 5, \omega_2 = 10$. For CHAOS data, we tuned them by setting $\omega_1 = 0.1$ and $\omega_2 = 0.2$. Finally, for the RGB images in PPSS, we tuned them to be: $\sigma_\alpha = 80, \sigma_\beta = 3, \sigma_\gamma = 3, \omega_1 = 3, \omega_2 = 3$.

While our method does not need a set of unpaired masks for training, we also compare with methods which learn the shape prior from masks:

- **UNet_{AAG}**: First, we consider the method developed in Chapter [6] upon which we build our model by substituting the multi-scale GAN with a self-supervised loss. The subscript AAG stands for Adversarial Attention Gates, which couple adversarial signals and attention blocks.
- **DCGAN**: We also consider a standard convolutional GAN, learning the shape prior from unpaired masks. This model is the same as UNet_{AAG}, but without attention gates and multi-scale connections between segmentor and discriminator.
- **ACCL**: Proposed by Zhang, Zhong, and Li, [2020] and similar to DCGAN, ACCL trains with scribbles using a PatchGAN discriminator (Isola et al., [2017]) to provide adversarial training signals, and with the \mathcal{L}_{PCE} loss on the annotated pixels.

We perform 3-fold cross-validation and report the performance as a distribution of values on the test-set samples. To measure segmentation quality, we use Dice and IoU scores, and the Hausdorff Distance.

7.4.3 Results

We report test results for each dataset in Figure 39. The box plots show the median and inter-quartile range (IQR) of each metric, considering outliers the values outside $2 \times \text{IQR}$. We show visual examples of predicted segmentation masks in Figure 40.

As can be observed, our method is the best one when we compare it to other approaches which do not require unpaired masks for training (Fig. 39, left column). In particular, using a simple UNet leads to unsatisfying performance, while regularisation considerably helps. Introducing the compactness loss aids more with compact objects, such as those in ACDC, CHAOS and PPSS, while it can be harmful when dealing with non-compact shapes, such as that of the myocardium (which has a doughnut-shape) in LVSC.

Post-processing the segmentor predictions with CRF can lead to performance increase when object boundaries are well defined. On the contrary, we could not make the performance increase on CHAOS data, where using CRF made segmentation worse with all the metrics.

On LVSC, the introduced multi-scale shape consistency prior tends to make the model a bit less conservative on the most apical and basal slices of the cardiac MRI. Unfortunately, whenever there is a predicted mask but the manual segmentation is empty, the Hausdorff distance peaks. In fact, by definition, the distance assumes the maximum possible value (i.e. the image dimension) whenever one of the masks is empty, which makes the performance distribution on the test samples broader (see the box plot of the Hausdorff distance for LVSC, in Figure 39, left column).

On CHAOS, the IoU and Dice scores are more skewed for methods not using unpaired masks for training (Fig. 39, left column). This happens because CHAOS is a small dataset, and optimising models using only scribble supervision is challenging. On the contrary, adding extra knowledge, such as unpaired masks, may help (Fig. 39, right column).

Finally, comparing our method with those using unpaired masks for training (Fig. 39, right column), we find that training with PyAG modules leads to competitive performance on all datasets. Although, in some

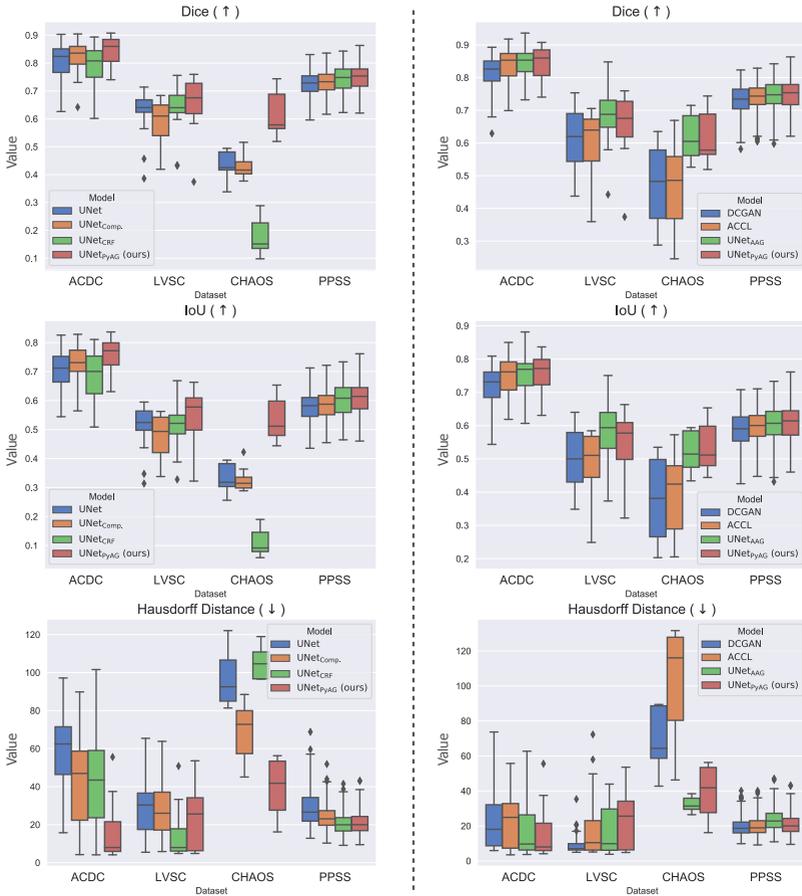


Figure 39: Segmentation performance in terms of Dice (\uparrow) and IoU scores (\uparrow) and Hausdorff distance (\downarrow), with arrows reporting metric improvement direction. The box plots report median and inter-quartile range (IQR) of each metric on a given test set, considering outliers those values falling outside $2 \times \text{IQR}$. **Left column:** our method vs baseline (UNet) and other methods regularising the prediction with a Compactness loss (UNet_{Comp.}), or CRF as post-processing (UNet_{CRF}). Observe how our method shows the best performance across datasets. **Right column:** our method vs methods regularising the predictions using a shape prior learned from unpaired masks (DCGAN, ACCL, and UNet_{AAG}). Our method has competitive performance with the best of the benchmark models, and it has the advantage of not requiring a set of unpaired masks for training.

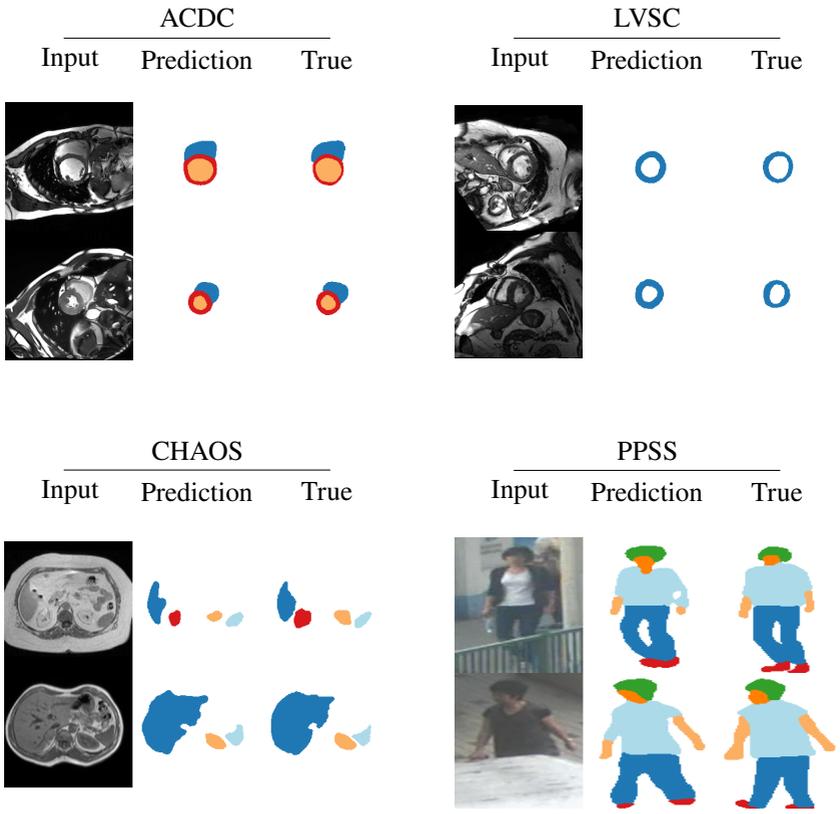


Figure 40: Example of segmentation masks predicted by our model on different datasets. In most cases, the model can effectively approximate the true segmentations.

cases, the UNet_{AAG} performs slightly better than UNet_{PyAG}, we emphasise that our approach is entirely mask-annotation free.

7.5 Conclusion

We introduced a novel self-supervised learning strategy for semantic segmentation. Our approach consists of predicting masks at multiple resolution levels and enforcing multi-scale segmentation consistency. We use these multi-scale predictions as part of attention gating operations, restricting the model to re-use the extracted information on the object shape and position. Our method performs considerably better than other scribble-supervised approaches while having comparable performance to approaches requiring additional unpaired masks to regularise their training.

7.6 Summary

This chapter showed that the multi-scale relationship constitutes an important shape prior and that it is possible to introduce it in a self-supervised manner, generalising the approach of Chapter 6 to the case where unpaired masks are not available. More broadly, we believe self-supervision is a promising research direction, improving model performance without relying on human annotation effort. However, we also experimented that – when it is possible to collect unpaired masks – adversarial learning can be a helpful training regulariser.

We highlight that, until now, our work focused on introducing shape priors to regularise model training. However, shape priors can be helpful during inference, too. In fact, segmentors may under-perform if a test image falls outside the learned training distribution. In these cases, detecting mistakes and possibly correcting model predictions is a challenging and exciting research avenue. Toward this goal, the next chapter introduces a novel strategy for re-using components already developed during training to act as a shape prior at inference. We show that it is possible to increase segmentation quality by adapting the model to each

test image independently, ultimately improving model robustness under distribution shifts.

Chapter 8

Re-using Adversarial Shape Priors for Test-Time Training

□

In the previous chapters, we discussed that thanks to their ability to learn data distributions without requiring paired data, Generative Adversarial Networks (GANs) are an integral part of many object segmentation methods. At inference, it is common practice to discard the adversarial discriminator and only use the segmentor to predict label maps on the test data. But should we discard the discriminator? In this chapter, we argue that the life cycle of adversarial discriminators should not end after training. On the contrary, training stable GANs produces powerful shape priors that we can use to *correct* segmentor mistakes at inference.

This chapter is based on:

- Valvano, Gabriele, Andrea Leo, and Sotirios A Tsaftaris (2021b). “Stop Throwing Away Discriminators! Re-using Adversaries for Test-Time Training”. In: *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Springer, pp. 68–78
- Valvano, Gabriele, Andrea Leo, and Sotirios A. Tsaftaris (2021d). “Re-using Adversarial Mask Discriminators for Test-time Training under Distribution Shifts”. In: *arXiv preprint arXiv:2108.11926*

To achieve this, we develop stable mask discriminators that do not overfit or catastrophically forget. At test time, we fine-tune the segmentor on each individual test instance until it satisfies the learned shape prior.

The proposed method is simple to implement and increases model performance. Moreover, it opens new directions for re-using mask discriminators at inference.

8.1 Introduction

Semi-supervised and weakly-supervised learning are emerging training paradigms for image segmentation (Cheplygina, de Bruijne, and Pluim, 2019; Tajbakhsh et al., 2020), often involving adversarial training (Goodfellow et al., 2014) when annotations are sparse or missing. Adversarial training involves two simultaneously trained networks: one focusing on an image generation task and the other learning to tell apart generated images from real ones. In the context of image segmentation, the generator is named segmentor and, conditioned on an input image, learns to predict a realistic segmentation mask. After training, the second network – named discriminator, or critic – is discarded, and the segmentor used for inference.

Unfortunately, segmentors may underperform and make prediction errors whenever the test data fall outside the training data distribution. Here we propose a simple mechanism to detect and correct these segmentation errors in an end-to-end fashion, re-using components already developed during training.

We embrace an emerging paradigm (Sun et al., 2020; Wang, Shelhamer, et al., 2021; Karani et al., 2021) where a model is fine-tuned on individual test instances without requiring access to other data nor labels. We propose strategies that permit *recycling* an adversarial mask discriminator during inference, thus introducing a data-driven shape prior to correct predictions. Motivated by recent findings of Asano, Rupprecht, and Vedaldi, 2020, reporting that we can effectively train the early layers’ weights of a CNN with just one image, we propose to tune them on a per-testing instance to minimise an adversarial loss.

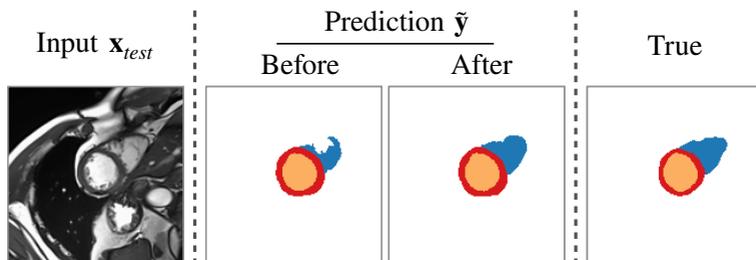


Figure 41: Whenever a test image falls outside the training data distribution, a segmentor may underperform and produce unrealistic predictions. Herein, we suggest re-using already optimised adversarial discriminators to tune the segmentor predictions on the individual test images until the predicted mask satisfies the learned shape prior.

8.1.1 Contributions

We summarise the contributions of this chapter as follows:

- This is the first attempt to use adversarial mask discriminators to *detect* and *correct* segmentation mistakes during inference.
- We define specific assumptions (and show how to satisfy them) to make the discriminators useful once training is complete.
- We explore several learning scenarios and report consistent performance increase on multiple medical datasets.

We report an example of the proposed method effect on initially erroneous predictions in Figure 41. We make code used for the experiments available at <https://vios-s.github.io/adversarial-test-time-training>.

8.2 Related Work

8.2.1 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 fine-tuning

a model on the test samples has recently been introduced by Sun et al., [2020] with the name of Test-time Training (TTT). Test-time Training optimises a model by jointly minimising a supervised and an auxiliary self-supervised loss on a training set, such as detecting the rotation angle of an input image. At inference, TTT fine-tunes the model to minimise the auxiliary loss on the individual test instances, thus adapting to potential distribution shifts. Although the model was successful for classification, the authors admit that designing a well-suited auxiliary task is non-trivial. For example, predicting a rotation angle may be less effective for medical image segmentation, where images have different acquisition geometries. Moreover, Sun et al. only test their model “simulating” domain shifts with hand-crafted image corruptions (e.g., noise and blurring) without investigating if TTT can improve segmentation performance in real-world settings.

Following this seminal work, Wang, Shelhamer, et al., [2021] suggested tuning an adaptor network to minimise the test prediction entropy. Unfortunately, CNNs usually make low-entropy overly-confident predictions (Guo, Pleiss, et al., [2017]), and entropy minimisation could be sub-optimal for segmentation. More crucially, Wang, Shelhamer, et al., [2021] rely on having access to the *entire* test-set to do the fine-tuning.

Karani et al., [2021] recently proposed Test-time Adaptable Neural Networks (TTANN) to extend TTT for image segmentation using a pre-trained mask Denoising Autoencoder (DAE). At inference, they compute a reconstruction error between the mask generated by a segmentor and its auto-encoded version predicted by the DAE. This error constitutes a test-time loss used to fine-tune a small adaptor CNN in front of the segmentor. Once tuned, the adaptor maps the individual test images onto a normalised space which overcomes domain shifts problems for the segmentor. A limitation of this approach is the need to train the mask DAE separately. On the contrary, GANs learn the shape prior and optimise the segmentor in an *end-to-end* fashion. Moreover, tuning the model with a convolutional encoder (e.g., a mask discriminator) rather than an autoencoder has advantages in terms of occupied memory and is faster at inference. In this work, we show that improving performance

using a discriminator is also possible and, at the same time, we open a new research direction toward learning re-usable discriminators.

8.2.2 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. 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 the training and test data (Joyce, Chartsias, and Tsaftaris, 2017; Li, Pan, et al., 2018; Dou, Castro, et al., 2019; Guan and Liu, 2021; Zhou, Liu, 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.

8.2.3 Shape Priors in Deep Learning for Medical Image Segmentation

Incorporating prior knowledge about organ shapes is not uncommon in medical imaging (Nosrati and Hamarneh, 2016; Jurdi et al., 2020). Several methods introduced shape priors to regularise the training of a segmentor using penalties (Kervadec, Dolz, Tang, et al., 2019; Clough et al., 2020; Jurdi et al., 2021), autoencoders (Oktay, Ferrante, et al., 2017; Dalca,

Guttag, and Sabuncu, [2018]), atlases (Dalca, Yu, et al., [2019]), and adversarial learning (Yi, Walia, and Babyn, [2019]; Valvano, Leo, and Tsaftaris, [2021c]). Others included shape priors for post-processing, fixing prediction mistakes (Painchaud et al., [2019]; Larrazabal et al., [2020]).

GANs have become a popular way of introducing shape priors for image segmentation (Yi, Walia, and Babyn, [2019]), with the advantage of: i) learning the prior directly from data; ii) having a simple model that works well for semi- and weakly-supervised learning; and iii) learning the prior while also training the segmentor, instead of in two separate steps (as it would happen for autoencoders).

8.2.4 Re-using Adversarial Discriminators

Re-using pre-trained discriminators has been proposed to obtain feature extractors for transfer learning (Radford, Metz, and Chintala, [2015]; Donahue, Krähenbühl, and Darrell, [2017]; Mao, Su, et al., [2019]), or anomaly detectors (Zenati et al., [2018]; Ngo et al., [2019]). To the best of our knowledge, their (re-)use to detect segmentor mistakes during inference remains unexplored. We are also not aware of previous use of discriminators for test-time tuning of a segmentor.

8.3 Method

In this section, we first provide an overview of the proposed approach. Then, we describe the challenges of re-using adversarial discriminators at inference and suggest possible solutions to address them. Finally, we detail model architectures and training objective and show how we re-use discriminators at test time.

8.3.1 Method Overview

As we summarise in Figure [42], we consider two stages: i) standard adversarial training; and ii) during inference, image-specific tuning of a small adaptor CNN $\Omega(\cdot)$ in front of the trained segmentor. In the first stage, we optimise adaptor and segmentor to minimise a supervised cost

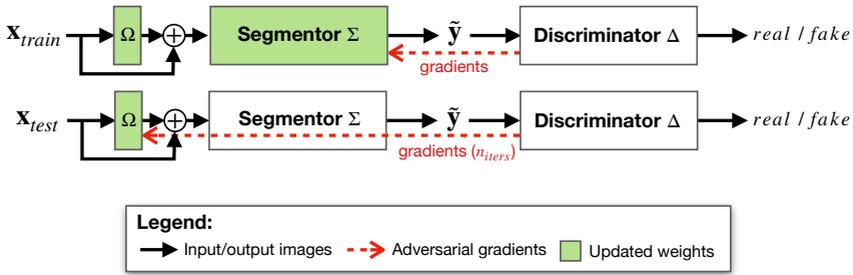


Figure 42: We re-use GAN discriminators to correct segmentation predictions at inference. The key to our success is training stable and re-usable discriminators, as we detail in Section 8.3.2. At inference, we tune a small convolutional block $\Omega(\cdot)$ on each test sample x , independently, until the predicted mask \tilde{y} satisfies the adversarial shape prior. We only need a single sample to do the fine-tuning.

on the annotated data and an adversarial cost on a set of unpaired images. Meanwhile, we train the discriminator to distinguish real from predicted masks. At inference, for each test sample, we only tune the adaptor $\Omega(\cdot)$ using the (unsupervised) adversarial loss and improve performance. We highlight that developing novel segmentors and adaptors is not our scope. Thus, we use previously developed architectures that showed success in segmentation tasks.

Obtaining discriminators re-usable at inference is not trivial and requires specific solutions to overcome crucial challenges. These solutions, with our optimisation strategy and model design, are one major contribution of this work.

8.3.2 Re-usable Discriminators: Challenges and Proposed Solutions

Challenge 1. To obtain a re-usable discriminator $\Delta(\cdot)$, we must prevent it from *overfitting* and *catastrophically forget*, or its predictions on the masks generated during inference will not be reliable. Generally speaking, this is a challenging task because: GANs can easily memorise data

if trained for too long (Nagarajan, Raffel, and Goodfellow, 2018).¹ Moreover, the discriminator may forget how unrealistic segmentation masks look like after the segmentor training has converged (Shrivastava, Pfister, et al., 2017). Although $\Delta(\cdot)$ may work well at training in these cases, it would not generalise to the test data, as we explain below.

If properly trained, a segmentor $\Sigma(\cdot)$ predicts *realistic* segmentation masks in the latest stages of training. Thus, in standard GANs, we stop training while optimising $\Delta(\cdot)$ to tell apart *real* from more and more *real-looking* masks. At convergence, this becomes similar to training the discriminator using only *real* images and labelling them as *real* half the times, as *fake* the other half. At this point, gradients become uninformative, and the discriminator collapses to one of the following cases: **i)** it always predicts its equilibrium point (which in vanilla GANs is the number 0.5, equidistant from the labels *real*: 1, *fake*: 0) but it can still detect unrealistic images; **ii)** it predicts the equilibrium point independently of the input image, forgetting what *fake* samples look like (Shrivastava, Pfister, et al., 2017; Kim, Kim, and Kim, 2018); or **iii)** it memorises the real masks (which, differently from the generated ones, appear unchanged since the beginning of training) and it always classifies them as *real*, while classifying *any other input* as *fake*. It is crucial to prevent the behaviours **ii)** and **iii)** to have a re-usable discriminator. For this reason, we use:

- *Fake anchors*: we ensure to expose the discriminator to unrealistic masks (labelled as *fake*) until the end of training. In particular, we train $\Delta(\cdot)$ 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 image² 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

¹Memorisation can also happen just in the discriminator. In fact, contrarily to the segmentors, we do not use any additional supervised cost to regularise the discriminator training. We show how to detect memorisation from the losses in Appendix C.1

²We use patches having size equal to 10% of the image size.

forgetting how unrealistic masks look like, providing informative gradients until we stop training³

Challenge 2. An additional challenge is to train *stable* discriminators, which do not change much in the latest training epochs. In other words, we want small oscillations in the discriminator loss. 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. To this end, we suggest using:

- *Spectral normalisation* (Miyato et al., 2018), *tanh* activations, and *Gradient Penalty* (Gulrajani et al., 2017): to increase the smoothness of the function learned by the discriminator (Chu, Minami, and Fukumizu, 2020).
- *Discriminator data augmentation*: consisting of random roto-translations, and Instance Noise (Sønderby et al., 2017; Müller, Kornblith, and Hinton, 2019), to map similar inputs to the same prediction label. We translate images up to 10% of image pixels on both vertical and horizontal axes, and we rotate them between $0 \div \pi/2$. We generate noise using a Normal distribution with zero mean and 0.1 standard deviation.

8.3.3 Architectures and Training Objectives for $\Sigma(\cdot)$ and $\Delta(\cdot)$

We use a UNet (Ronneberger, Fischer, and Brox, 2015) segmentor with batch normalisation (Ioffe and Szegedy, 2015). Given an image \mathbf{x} , let us consider its adapted version obtained as the output of the adaptor $\mathbf{x}' = \Omega(\mathbf{x})$. Received \mathbf{x}' as input, the segmentor $\Sigma(\cdot)$ predicts a multi-channel label map $\tilde{\mathbf{y}} = \Sigma(\mathbf{x}') = \Sigma \circ \Omega(\mathbf{x})$. For the annotated images, we

³Concurrent to our work, Sinha, Ayush, et al., 2021 recently introduced a similar idea named Negative Data Augmentation, which improved the training of GAN generators. However, differently from Sinha, Ayush, et al., 2021 our scope is to build a stable discriminator, which can be re-used at inference.

minimise the supervised weighted cross-entropy loss:

$$\mathcal{L}(\Omega, \Sigma) = - \sum_{i=1}^c w_i \cdot \mathbf{y}_i \log(\tilde{\mathbf{y}}_i), \quad (8.1)$$

where i is a class index, c the number of classes, and w_i a class scaling factor used to address the class imbalance problem. The value $w_i = 1 - n_i/n_{tot}$ considers both the total number of pixels n_{tot} and the number of pixels n_i having label i .

As discriminator $\Delta(\cdot)$, we use a convolutional encoder, processing the predicted masks with a series of 5 convolutional layers. Layers use a number of 4×4 filters following the series: 32, 64, 128, 256, 512. After the first two layers, we downsample the features maps using a stride of 2. As discussed in Section 8.3.2, we increase discriminator smoothness using spectral normalisation and *tanh* activations. Finally, a fully-connected layer integrates the extracted features and predicts a scalar linear output, which we use to compute the adversarial objectives (Mao et al., 2018):

$$\begin{aligned} \min_{\Delta} \left\{ \mathcal{V}_{LS}(\Delta) = \frac{1}{2} E_{\mathbf{y} \sim p(\mathbf{y})} [(\Delta(\mathbf{y}) - 1)^2] + \frac{1}{2} E_{\mathbf{x} \sim p(\mathbf{x})} [(\Delta(\Sigma \circ \Omega(\mathbf{x})) + 1)^2] \right\} \\ \min_{\Omega, \Sigma} \left\{ \mathcal{V}_{LS}(\Omega, \Sigma) = \frac{1}{2} E_{\mathbf{x} \sim p(\mathbf{x})} [(\Delta(\Sigma \circ \Omega(\mathbf{x})))^2] \right\}, \end{aligned} \quad (8.2)$$

where -1 and $+1$ are the labels for *fake* and *real* images, respectively, and 0 is the GAN equilibrium point. During training, we alternately minimise Equation 8.1 on a batch of annotated images and Equation 8.2 on a batch of unpaired images and unpaired masks. To avoid the adversarial loss from prevailing over the supervised cost on the segmentor, we rescale $\mathcal{V}_{LS}(\Omega, \Sigma)$ by multiplying it by a dynamic weighting value $a = 0.1 \cdot \frac{\|\mathcal{L}(\Omega, \Sigma)\|}{\|\mathcal{V}_{LS}(\Omega, \Sigma)\|}$ (Valvano, Leo, and Tsaftaris, 2021c). As a result, we ensure that the supervised cost on the segmentor is always one order of magnitude larger than the adversarial cost, which can judge predictions only qualitatively. We use Adam optimiser (Kingma and Ba, 2015), with a learning rate of 0.0001 and a batch size of 12. Training proceeds until the segmentation loss stops decreasing on a validation set.

8.3.4 Adversarial Test-Time Training: Adapting $\Omega(\cdot)$

At inference, we do not fine-tune the whole segmentor but only adapt a few convolutional layers at its input. Our choice is motivated by Asano, Rupprecht, and Vedaldi, [2020], who argued that early layers are the most suited for one-shot learning, and is similar to that of Karani et al., [2021]. By keeping the deeper layers of $\Sigma(\cdot)$ unchanged, we also limit the model flexibility and let it adapt only to changes at lower abstraction levels, ultimately preventing trivial solutions. Thus, given a test sample \mathbf{x} , we tune a shallow convolutional residual block (i.e. the adaptor $\Omega(\cdot)$) 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 $\Omega(\cdot)$, the input to the segmentor becomes an augmented version of \mathbf{x} , which can be more easily classified.

The adaptor is taken from Karani et al., [2021] and has 3 convolutional layers with $16\ 3 \times 3$ kernels and activation $\Phi(\mathbf{T}) = e^{-\mathbf{T}^2/s^2}$, where \mathbf{T} is an input tensor and s is a trainable scaling parameter, randomly initialised and optimised at test-time.

Test-time Iterations and Computational Aspects At inference, our method needs n_{iter} forward and backward passes to correct a segmentation. Despite this is slower than standard inference, where each image requires only one forward pass, we highlight that obtaining fast inference is not the purpose of this work. We leave the development of strategies for faster inference to future work.

For our experiments, we defined a different optimal n_{iter} for each test sample. We first define a maximum number of iterations $n_{iter}^{max} = 1000$ to define an upper bound on the inference time. Then, we stop TTT when the adversarial loss (or the sum of adversarial and reconstruction cost, in Section [8.5.3]) on the predicted mask has not decreased for the last 200 steps, or the number of iterations is equal to n_{iter}^{max} . Finally, we pick n_{iter} as the number of iterations that led to the prediction associated with the minimum adversarial loss, which we consider to be the best one.

8.4 Experimental Setup

Below, we first describe the datasets used to test the proposed approach. Then, we detail the evaluation protocol used for our experiments.

8.4.1 Data

We consider four medical datasets: ACDC (described in Section 3.8.1), LVSC (Section 3.8.2), CHAOS (we consider the T1 in-phase images, described in Section 3.8.4), and M&Ms (Section 3.8.3). We employ specific datasets based on two different learning scenarios, where we assume:

- **Identifiable Distribution Shift:** in this case, we use ACDC and M&Ms data to model test-time distribution shifts that we can easily identify as changes in the 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 as $ACDC_{1.5 \rightarrow 3T}$. For M&Ms, we consider training and validation set built using data from 3 out of the four available MRI vendors and construct the test set using data from the held-out vendor. As a result, both in $ACDC_{1.5 \rightarrow 3T}$ and M&Ms, we can be sure there is a distribution shift between training and test data. In both cases, we maintain a 2:1 ratio between the number of samples in the training and validation set.
- **Non-identifiable Distribution Shift:** in this second case, 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. To prevent information leakage, we divide datasets by patients and use groups of 40% for training, 20% for validation, and 40% for the test set, respectively. Out of the 40% training patients, we consider annotations for one fourth of the training subjects in ACDC and LVSC (10 patients) and one half for CHAOS (4 patients). We treat

the remaining data as unpaired and use them for adversarial training (Equation 8.2). 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. In this case, although we cannot identify distribution shifts a priori, they may still be present in the test set and lead to a performance drop (Recht et al., 2018).

8.4.2 Evaluation Protocol

For all the experiments, we report results of 3-fold cross-validation. We measure performance in terms of segmentation quality, using the Dice score, the IoU score, and the 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, which, differently from rank-based tests like the non-parametric Wilcoxon test, has the advantage that it can distinguish between large and small metric variations. Moreover, it allows us to compare performance mean/median directly. We use significance at $p \leq 0.05$ or $p \leq 0.01$ denoted by one (*) or two (**) asterisks, respectively.

8.5 Experiments and Discussion

We present and discuss the performance of our method in various experimental scenarios. At first, we present the advantage of the proposed approach during inference: either under identifiable or non-identifiable distribution shifts (Section 8.5.1). In Section 8.5.2 we discuss a limitation of the model and define a possible solution that we analyse in Section 8.5.3. In the latter, we study the effect of additional reconstruction losses to aid the adversarial discriminator and do better Test-time Training. After that, Section 8.5.4 shows that the adversarial TTT is different from and compatible with post-processing operations, leading to complementary performance gains. Lastly, Section 8.5.5 shows that the method can be potentially used for Online Continual Learning.

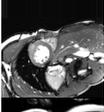
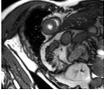
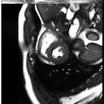
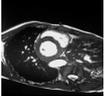
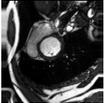
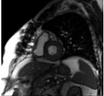
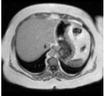
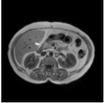
Dataset	Input	Prediction		True
		Before TTT	After TTT	
ACDC _{1.5T→3T}				
				
M & Ms				
				
ACDC				
				
LVSC				
				
CHAOS				
				

Figure 43: We show examples of mistaken predictions and their corrections obtained 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.

8.5.1 Adversarial Test-time Training Under Distribution Shifts

We report a qualitative example of test-time adaptation in Figure 43, 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 holes.

In Figure 44, we represent segmentation performance on the test set with violin plots, before and after Test-time Training. These plots show the whole distribution of performance values for patients in the test set. We observe performance improvements across metrics and datasets both in terms of 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. However, we observe empirical improvements in terms of Dice and IoU scores on CHAOS, too. From these results, we can argue that adversarial TTT could lead to substantial benefits for medical applications, where systems must be robust and prevent trivial mistakes.

In Figure 45, we compare the performance of our method vs one using a shape prior separately learned by a DAE (TTANN, Karani et al., 2021). To the best of our knowledge, the TTANN method is the only prior work on Test-time Training in semantic segmentation. Thus, we compare to improvements obtained with TTANN and discuss the pros and cons of driving the adaptation using a DAE vs a mask discriminator.

Our experiments show advantages in using 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 $\Delta(\cdot)$ 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 observed in Larrazabal et al., 2020.

We also emphasise that mask discriminators have the advantage of

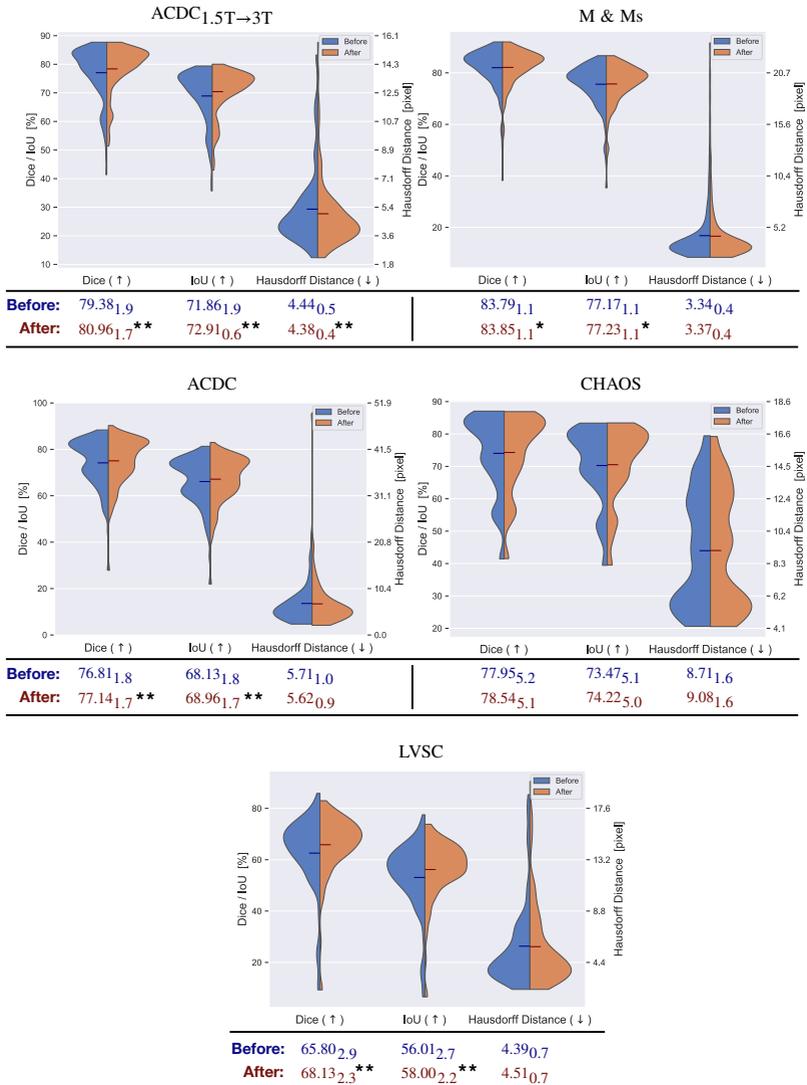


Figure 44: Dice (↑), IoU (↑) and Hausdorff distance (↓) obtained **before** and **after** tuning the segmentor on the individual test instances. Arrows show metric improvement directions. Under each violin plot, we also report the median performance, with 95% confidence interval as subscript. Observe how adversarial Test-time Training improves performance under different metrics and datasets (bootstrapped t-test, * $p < 0.05$, ** $p < 0.01$).

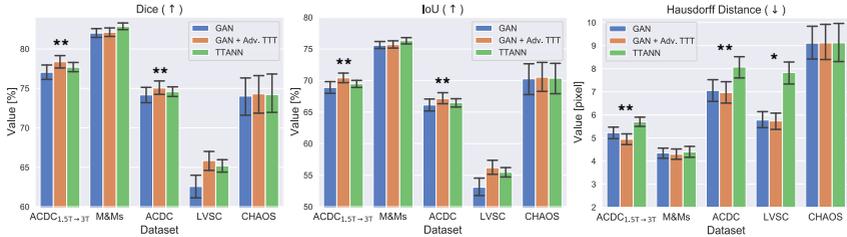


Figure 45: Adversarial TTT has competitive performance with TTANN, and it has the advantage of re-using an already available GAN component. Bar plots report average performance and standard errors. Stars on top of the bar plots show if differences between adversarial TTT and TTANN are significant (bootstrapped t-test, * $p < 0.05$, ** $p < 0.01$).

being already available after training GANs, and, thus, they are always ready to use without any additional training effort. Moreover, thanks to their encoder-like architecture (rather than auto-encoder-like, as in DAEs), the discriminators need a reduced computation, making inference faster.

Lastly, we perform an ablation study to analyse the effect of the adaptor, the smoothness constraints and the proposed *fake anchors* regularisation on the model. As illustrated in Table 8, the techniques stabilise training and make the adversarial shape prior stronger. As a result: i) the adversarial training leads to a better segmentor, and ii) the re-usable discriminator further increases model performance. For comparison, training a simple UNet on the same data leads to an average Dice score of 70.1 (standard deviation of 13).

8.5.2 Limitations and a Possible Solution

From our experiments, we observe that, in some rare cases, adversarial TTT makes segmentation performance worse. This happens because we drive adaptation based on a predicted mask without considering any additional information about the image. In other terms, a limitation of the method is that the discriminator learns to approximate the shape prior characterised by the probability distribution $p(y)$ rather than the joint

	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 8: 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 8.3.2 (ablation #3), and the adaptor (standard GAN, #4). Performance is in terms of average Dice score on ACDC data, with standard deviation as subscript.

distribution $p(\mathbf{x}, \mathbf{y})$. Thus, the discriminator does not penalise realistic masks even when they are wrong segmentations for a given image (as shown in Figure 46). Hence, it becomes natural to wonder if considering both the image and the segmentation mask to drive the adaptation process may provide additional context and help during test-time adaptation. We highlight that this problem is also present in TTANN (Karani et al., 2021) and in all the methods that learn the marginal $p(\mathbf{y})$ rather than the joint distribution of images and masks. We explore a possible solution below.

8.5.3 Toward Causal Test-time Training

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

In our model, we optimise the segmentation modules (Ω and Σ) 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. However, tuning the adaptor based only

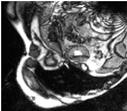
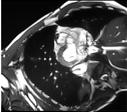
Input	Prediction		True
	Before TTT	After TTT	
			
			

Figure 46: Failure cases. Since the information contained inside the predicted mask is limited, the discriminator will not penalise realistic but wrong segmentation masks (**top row**). In some cases, it might even encourage the segmentor to make bigger mistakes (**bottom row**).

on the adversarial loss may be considered 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, Walker, and Glocker, 2020). Instead, GANs whose discriminator only receives segmentation masks as input would penalise the segmentor without considering the image causing that mask.

From a causal perspective, our approach is non-optimal because we should also consider the inverse conditional probability $p(\mathbf{x}|\mathbf{y})$ to capture the causal structure better and update the model parameters, improving the approximation $p(\mathbf{y}|\mathbf{x})$ according to Bayes theorem:

$$p(\mathbf{y}|\mathbf{x}) = p(\mathbf{x}|\mathbf{y}) \frac{p(\mathbf{y})}{p(\mathbf{x})}. \quad (8.3)$$

Hence, to obtain a more coherent description, we should learn an inverse function which maps 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. 8.3), 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-masks 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 that 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 8.4.1).⁴

Another alternative to learning $p(\mathbf{x}|\mathbf{y})$ is to substitute it with a proxy distribution. For example, we could learn $p(\mathbf{x}|\mathbf{y}, \mathbf{R})$, where \mathbf{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} and thus break the one-to-many, many-to-one problem described above. In this case, we would establish the relationship:

$$p(\mathbf{y}|\mathbf{x}) \leftrightarrow p(\mathbf{x}|\mathbf{y}, \mathbf{R}) \frac{p(\mathbf{y})}{p(\mathbf{x})}. \quad (8.4)$$

An example of such a model is SDNet (previously discussed in Chapter 4), which uses the extracted mask and its residuals to reconstruct the image⁵ while also having an adversarial discriminator learning $p(\mathbf{y})$.

We experimented with this framework to explore if reconstructing the test samples during inference adds benefits during adaptation in terms of *performance* and *adaptation speed*. Thus, we first included the adaptor Ω in

⁴For completeness, we also conducted an experiment in fully-supervised learning, where all the training images are associated to a segmentation mask and the discriminator can learn the joint distribution. In this case, we observed that the discriminator was more prone to overfit the training data, and its generalisation under distribution shifts got worse.

⁵To 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 this experiment, we assume it is a reasonable approximation.

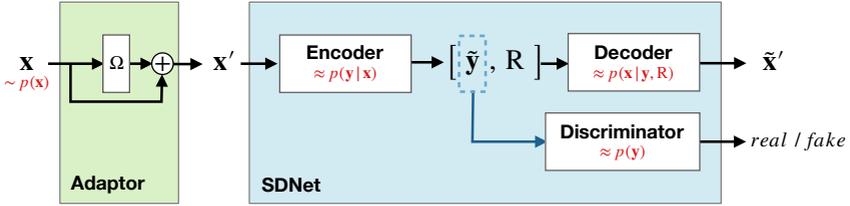


Figure 47: The proposed approach in a causal setting. We add the adaptor Ω in front of SDNet to transform an image $\mathbf{x} \sim p(\mathbf{x})$ into its adapted version \mathbf{x}' . **During training**, the SDNet encoder extracts the segmentation mask $\hat{\mathbf{y}}$ and a residual representation \mathbf{R} . A decoder uses both of them to reconstruct the adapted image, predicting $\tilde{\mathbf{x}}' \approx \mathbf{x}'$. A mask discriminator learns to say apart real segmentation masks from the predicted ones. **At inference**, we perform Test-time Training by minimising the sum of the reconstruction cost (computed comparing \mathbf{x}' and $\tilde{\mathbf{x}}'$) and the adversarial loss (computed on the predicted $\hat{\mathbf{y}}$ according to Equation 8.2).

front of SDNet (as shown in Figure 47). Then, we trained the full model according to the SDNet training objectives, which include an adversarial loss \mathcal{L}_A and a reconstruction loss \mathcal{L}_R .

The adversarial loss is the same we defined in Equation 8.2, and we also followed the precautions discussed in Section 8.3.2

We left the reconstruction term as in the original SDNet framework, minimising the mean absolute error between an image and its reconstruction. However, we trained 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 when there is a distribution shift between training and inference data, the SDNet decoder may not be able to reconstruct the test image correctly. On the contrary, after tuning Ω to the test image, the SDNet can effectively reconstruct the adapted image \mathbf{x}' ⁶ We left 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 patient. We set the number of TTT steps n_{iter} as described in Section 8.3.4, and performed TTT in three different

⁶An 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.

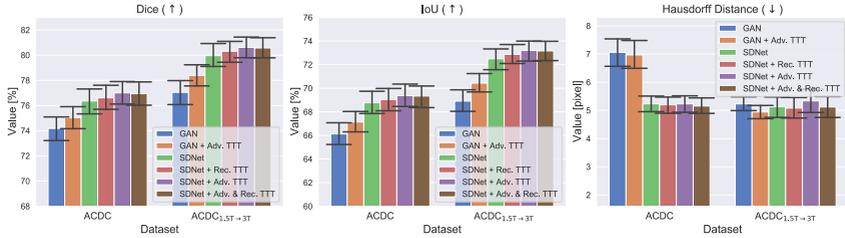


Figure 48: Toward Causal Test-time Training. We compare the performance of: a GAN before and after adversarial Test-time Training; the SDNet model (discussed in Section 8.5.3); the SDNet after Test-time Training performed minimising only a reconstruction cost (“+ Rec. TTT”), only an adversarial cost (“+ Adv. TTT”) and their sum (“+ Adv. & Rec. TTT”). Bar plots report average performance and standard errors.

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 (ACDC_{1.5→3T} data) and non-identifiable shifts (ACDC data). We report per-dataset results in terms of segmentation quality in Figure 48.

From the 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 to the target image. There is only one experimental exception to this result: the Hausdorff Distance of “SDNet + Adv. TTT” on ACDC_{1.5→3T} data. In this case, despite Dice and IoU scores increase, the Hausdorff

Distance gets worse. We argue that this happens because the model makes more errors in the most apical and basal slices of the heart⁷ a behaviour that we also observe for “GAN” and “GAN” + Adv. TTT”, where the Hausdorff distance is high.

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 Figure 48), we can see that there is 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”).

On the other hand, including \mathcal{L}_R during TTT has a bigger impact on the test-time adaptation 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 ACDC_{1.5→3T} data. By including also the reconstruction cost in TTT, the average number of TTT steps becomes 66 on ACDC and 119 on ACDC_{1.5→3T}. Our results are also in line with recent findings arguing that correct causal structures adapts faster (Bengio, Deleu, et al., 2020).

From the experiments we conclude 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.

8.5.4 Combining Adversarial TTT with Post-processing Operations

Adversarial TTT should not be confused with post-processing operations because it does not directly modify the predicted segmentation masks. On the contrary, our approach lets the model adapt to the input image and, as such, is compatible with post-processing techniques. Moreover,

⁷By 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, making results worse.

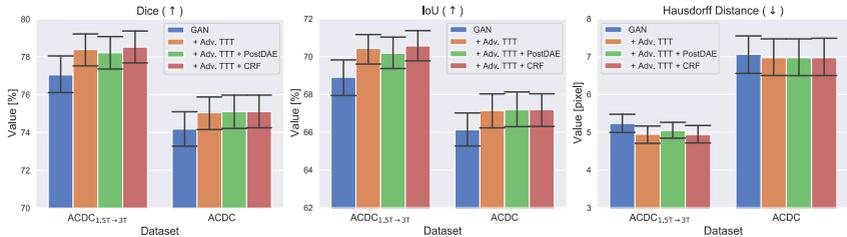


Figure 49: Compatibility with post-processing techniques (PostDAE and CRF). Bar plots report average performance and standard errors.

contrary to standard post-processing operations, our method has the advantage that it can also learn from a continuous stream of data, as we will show in the next section.

As examples, we consider two popular post-processing techniques. First, we examine post-processing the predicted masks with Conditional Random Fields (CRF), as in the DeepLab framework (Chen, Papandreou, et al., 2017). We highlight that this approach adapts the predicted mask to the image, while our method 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 a corrupted mask on a previously learned manifold of realistic masks without considering the input image.

Figure 49 shows that our method can be combined with both techniques and, sometimes, even improves performance. In particular, we find that PostDAE does not always help: probably because adversarial TTT already adapts the model using a data-driven shape prior, so additional DAEs may be useless or even harmful. On the contrary, CRF increases performance because it introduces a different type of prior knowledge in the model (Zheng et al., 2015), from which the segmentor can benefit (similar to what happens in model ensembling).

Dataset	Adv. TTT	Continual	Dice (\uparrow)	IoU (\uparrow)	Hausdorff Distance (\downarrow)
ACDC	\times	\times	74.2 ₁₀	66.1 ₁₀	7.1 ₀₅
	\checkmark	\times	75.0 ₀₉	67.1 ₁₀	6.9 ₀₅
	\checkmark	\checkmark	75.1₀₉	67.2₁₀	6.9₀₅
ACDC _{1.5\rightarrow3T}	\times	\times	77.0 ₀₉	68.9 ₀₉	5.2 ₀₂
	\checkmark	\times	78.4 ₀₈	70.4 ₀₈	5.0 ₀₂
	\checkmark	\checkmark	78.6₀₈	70.6₀₈	4.9₀₂

Table 9: Online Continual Learning. We show that our model can continuously learn from a stream of test data, leading to gradually higher segmentation scores on the test samples. Numbers are average performance, with standard deviation as subscript. Best results in **bold**.

8.5.5 Online Continual Learning

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). When learning from new data, the model performance should increase in time. Moreover, as the model starts getting better on the test distribution, the need for TTT should decrease, making test-time adaptation gradually faster.

We conducted experiments for both ACDC and ACDC_{1.5 \rightarrow 3T}, and report results in Table 9. In this continual learning scenario, we do not restart TTT from zero when testing on a new image; on the contrary, we simply continue the learning process from one test patient to the other.

Overall, we find that the segmentor benefits from learning on new data, increasing test-time performance. More interestingly, the average number of TTT steps needed for adaptation decreased from 322 to 315 on ACDC data and from 120 to 114 on ACDC_{1.5 \rightarrow 3T}. This reduced number of steps suggests that gradually introducing new knowledge into the model lessens the need for adaptation, and the segmentor might be able to do without TTT after a while.

We believe that learning without supervision on new test data is a promising research avenue. However, there are still several challenges to

solve, such as alleviating the risk of forgetting previous experience when we continually learn with new data. To this end, it would be interesting to combine our method with other continual learning approaches in the future. For example, Elastic Weight Consolidation (Kirkpatrick et al., 2017) and Variational Continual Learning (Nguyen et al., 2018) penalise big updates of the segmentor parameters after their initial optimisation, and may help in the continual learning scenario (Mai et al., 2021).

8.6 Conclusion

In this chapter, we demonstrated that by satisfying simple design assumptions, it is possible to re-use adversarial discriminators during inference. In particular, we re-used a mask discriminator to detect and then correct segmentation mistakes made by a segmentor. The proposed method is simple and can be applied to any GAN, potentially increasing its test-time performance on the most challenging images. We showed that reconstruction costs could complement the adversarial discriminator and improve inference speed. Moreover, the proposed framework benefits from continual learning, making test-time inference more accurate and faster.

More broadly, the possibility of re-using adversarial discriminators to correct generator errors may open opportunities even outside image segmentation. Given 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, Walia, and Babyn, 2019). With improved architectures and regularisation techniques (Kurach et al., 2019; Chu, Minami, and Fukumizu, 2020), we believe adversarial networks will be even more popular in the future. In this context, training stable and re-usable discriminators opens opportunities for using flexible data-driven losses at test time, making inference better. We will discuss these and other promising research directions more in detail in the next chapter, which concludes this manuscript.

Chapter 9

Summary and Future Directions

In this thesis, we presented deep learning methods for the segmentation of medical images. We discussed that collecting large-scale fully-annotated datasets is a challenging problem in medical imaging. Thus, it is necessary to introduce regularisation: *directly* at the features representation level or *indirectly*, imposing prior-driven constraints on the model predictions. In the following sections, we briefly summarise the approaches developed to achieve this goal in the previous chapters. We also discuss some limitations and future research avenues.

9.1 Summary

In this manuscript, we presented methods for learning with limited levels of supervision. In particular, we discussed how it is possible to use prior knowledge about the patient anatomy to learn in semi-supervised, weakly-supervised, and unsupervised (test-time) settings.

Motivated by reducing the dependency on annotations, we presented a disentanglement-based approach to regularise semi-supervised segmentation learning in Chapter 4. We first illustrated how to decompose an image into anatomical and modality-specific components.

Then, we presented a method for regularising these representations to be temporally coherent in cardiac cine MRI, ultimately improving semi-supervised segmentation quality. Our experiments showed that disentangled methods are well suited for semi-supervised learning and increase model robustness in a lack of labels. However, we also experimented that these frameworks can be challenging to train, in practice, because they require carefully balancing several different losses. Moreover, when the number of available annotations increases, their training does not necessarily improve model performance. On the other hand, more straightforward methods, such as Generative Adversarial Networks, can offer simpler and effective alternatives for semi-supervised learning.

In Chapter 5, we explored the use of GANs to learn data-driven shape priors. We compared several GAN variants and found that adversarial shape priors are useful regularisers for semi-supervised learning. In general, we observed that GANs are harder to train when the training images are too scarce but, provided a sufficient amount of data, they become more stable and constantly improve semi-supervised performance. The chapter also explored several regularisation techniques and presented a novel approach adding textures on top of the unpaired segmentation masks. In general, we found that adding continuous values on top of the segmentation masks helps training better discriminators. We believe that one of the key reasons why this helps is that continuous values contribute to making the distribution observed by the discriminator broader, thus limiting its overfitting risk.

Aiming to overcome some of the limitations of GANs, such as the lack of multi-scale consistency priors, Chapter 6 presented a novel multi-scale GAN formulation that forces the segmentor to consider short-range and long-range object dependencies and scale. As a result, the data-driven shape prior introduced in the segmentor became much more effective, increasing segmentation quality. The key to the model success was to tightly couple a segmentor and a mask discriminator through adversarial conditioning of attention gates, helping to suppress scattered false positives in the predicted masks. We showed that such a powerful shape

prior helps recover missing information in semi-supervised and weakly-supervised learning. Finally, we also discussed that it would be interesting to explore the adversarial conditioning of attention gates in a broader context, from image registration to synthesis tasks.

When it is not possible to include unpaired segmentation masks in the training process, we can still introduce multi-scale consistency through a self-supervised loss. We showed this in Chapter 7, removing the need for adversarial discriminators and unpaired masks while accepting only slight or no performance compromises.

Lastly, we observed that it can be useful to introduce shape priors at inference, too. In fact, when the training data is not representative of the test distribution, previously optimised segmentors may underperform and produce unrealistic outputs. Detecting trivial mistakes is crucial for medical applications, and it should be made possible. For this reason, in Chapter 8 we showed that we can re-use shape priors learned by adversarial mask discriminators to detect and correct segmentation mistakes at inference. This chapter also opened new research directions for comprehensive test-time usage of the already developed GAN components.

9.2 Future Directions

We believe that disentangled models, such as those presented in Chapter 4, have a great potential to increase model robustness when labels are scarce. We also find that disentangled representations offer an intuitive interpretation that is well suited for image generation and for learning from multiple imaging modalities. However, disentanglement usually needs to balance many supervised and unsupervised objectives, making models hard to train. We argue that solving this limitation is crucial to make this type of methods easier to develop and more broadly applicable. In this context, relying on architectural and data biases may reduce the number of training objectives and lead to simpler models. Moreover, we believe that *full* disentanglement is not necessarily the best option as it can limit too much model flexibility, as also observed in Liu*, Thermos*, et al., 2021. On the contrary, architectural biases may be less

stringent and allow more expressiveness in the features space. However, finding the right balance of biases remains an open research problem.

In Chapter 6 and 7 we highlighted that learning hierarchical dependencies in the object shapes increases model robustness on unseen data. It would be interesting to devise disentanglement methods to obtain modality-specific and anatomy-specific *hierarchical factors*, which can better capture the semantic information of the image (Vahdat and Kautz, 2020; Chai, Wulff, and Isola, 2021). For example, it would be exciting to develop compositional decoders that, given an input representation, process it in a disentangled and hierarchical manner, changing specific portions of the reconstructed image while maintaining global coherence. This would also be useful for data augmentation, or to fight dataset imbalance (Chai, Wulff, and Isola, 2021).

Finally, it would be nice to consider the temporal transformation predicted by the disentanglement method in Chapter 4 as acting at multiple scales, similar to the registration method proposed by Krebs et al., 2019.

In the thesis, we widely adopted GANs to learn shape priors. As we discussed, GAN discriminators learn flexible data-driven losses, but optimising GANs can still be challenging with limited data. Regularisation techniques and architectural biases are promising research directions to render GANs more stable and easy to use. Thanks to their ability to learn data-driven losses, we believe GANs will not see their popularity decrease. For this reason, methods re-using the previously learned loss functions have great potential to help to validate the model predictions in the real world, where distribution shifts may hamper segmentation performance and where monitoring the correct functioning of the model could be hard. Hence, we believe that exploring a broader application context for methods such as those presented in Chapter 8 is a promising research direction, potentially helpful even outside image segmentation tasks. For example, it would be exciting to explore the re-use of discriminators trained for image generation and style transfer tasks. Similarly, we find attractive the idea to ensemble discriminators obtained from multiple GANs or that have learned different types of losses.

Having a better understanding of the properties of the learned loss

functions is also of crucial importance. Experiments in Chapter 6 showed that using just a few unpaired masks was sufficient to learn an adequate shape prior and regularise training in weakly supervised settings. A natural question arises whether the discriminator learns an expressive shape prior-driven loss or it would only learn proxies, such as object compactness and approximated object size, should it be sufficient for training. In this context, it would be fascinating to explore the use of invertible mask discriminators. This type of discriminators would encourage the features extracted before its last fully connected layer to maintain all the available information of the input, thus ensuring semantically rich representations. In this case, the training signals for the segmentor may become of a higher quality and maybe lead to better models. Invertible discriminators may also benefit more from discriminator data augmentation, thanks to equivariant (rather than invariant) learning.

Deepening our understanding of the type of loss function that adversarial discriminators can learn is especially relevant for medical applications, where interpretability is important. Thus we must ask: *does the adversarial framework introduce unexpected biases? And: how do model architecture, optimisation strategy, and data affect the training of the GAN generator?* Being aware of these biases is extremely important to understand the framework's limitations, especially for adversarial Test-time Training, where the discriminator could potentially make the model perform worse. For example, how would mask discriminators behave when there is only one labelled pixel in the predicted mask? Would they push the generator to suppress it or enlarge it? Addressing these questions is not trivial, and we believe that introducing more context (e.g., spatial coordinates, adjacent slices, image intensity) could be beneficial. Thus, it would be interesting to tune the discriminator behaviour to correctly address ambiguous cases, both in training and inference. It would also be exciting to develop advanced techniques to weigh the adversarial contribution based on specific environmental factors, such as patient age, pathology or slice position. In this context, integrating vision and text data with multi-modal frameworks could be a successful strategy, too. Hence, we believe Visual Reasoning and Causal Learning could play an

important role in learning data-driven losses and fully exploit contextual information for model predictions.

On the segmentor side, another exciting opportunity resides in mixing strategies developed in multiple learning paradigms and for different machine learning problems. For example, methods developed for Natural Language Processing (such as Transformers, Vaswani et al., 2017, Caron et al., 2021) have recently been applied with success in vision tasks, too. We think that a closer collaboration between domain experts will foster computer vision progress even further.

Lastly, while we presented methods for the segmentation of two-dimensional images, we envision their extension to segment 3D volumes. In fact, several medical imaging modalities, such as MRI and CT, provide three-dimensional information about the patient. Compared to the two-dimensional segmentation we focused on, 3D views benefit from additional contextual information to segment challenging images. However, 3D models still have several challenges to solve. First, the effective dataset size decreases because, in contrast to two-dimensional models, each subject 3D volume constitutes a single training sample rather than multiple 2D images. Moreover, the use of 3D convolutional layers significantly increases the number of parameters. As a result, preventing the risk of model overfitting becomes more challenging. In this context, self-supervised and transfer learning could play an important role to improve data efficiency. Similarly, research in Few-shot Learning (Wang, Yao, et al., 2020) is going through lots of progress, and it could provide effective training strategies also for 3D object segmentation.

Appendix A

Experimental Details of Chapter 5

A.1 Experimental Details of Section 5.2 and 5.3

In the following, we detail the architectures and the optimisation strategy used for the experiments in Section 5.2 and 5.3

A.1.1 Model Architectures

The GAN consists in a Segmentor and a Discriminator neural networks. We detail each of them below.

Segmentor We use a UNet segmentor (Ronneberger, Fischer, and Brox, 2015). The UNet has an auto-encoding architecture, where the encoder extracts feature maps at multiple depth levels and propagates them to the decoder using skip connections and a concatenation operation. The convolutional layers have $3 \times 3 \times k$ filters, with k equal to the number of input channels. After each convolutional layer, we apply batch normalisation (Ioffe and Szegedy, 2015) and use a *ReLU* activation function. The number of filters for each layer follows the series 32, 64, 128, 256, 512 for each depth level of the encoder, respectively. The decoder has a

symmetrical structure.

The segmentor output consists in a convolutional layer with c kernels having size $1 \times 1 \times k$, where k is the number of input channels, and c is the number of possible classes to segment, including the background. The layer output is then processed using a *softmax* function, which maps the values to a probabilistic range for each object class.

Discriminator For each depth level d , a convolutional layer processes the input using $4 \times 4 \times k$ filters and stride of 2, where k are the input channels. The number of filters is the same as in the segmentor encoder (i.e. 32, 64, 128, 256, 512, for depth d , respectively). A second convolutional layer compresses the features maps using 12 kernels with size $1 \times 1 \times k$. We use *tanh* activation function for both layers. Finally, a fully connected layer integrates the high-level features extracted from the input and produces an output scalar, which we use to compute the adversarial loss.

In Section 5.2 we stabilise the training process using: i) spectral normalisation (Miyato et al., 2018); ii) instance noise (Sønderby et al., 2017) with zero mean and 0.2 standard deviation; and iii) label noise with 10% flipping probability (Salimans et al., 2016). In Section 5.3 we only use spectral normalisation and the regularisation technique explicitly described in the chapter.

A.1.2 Optimisation

We use Adam optimiser (Kingma and Ba, 2015), learning rate of 0.0001, and batch size of 12 to minimise the training cost. Training proceeds until convergence according to an early stopping criterion on a validation set. The criterion stops training when the supervised cost between ground-truth and predicted masks stops decreasing.

Appendix B

Additional Experiments and Results of Chapter 6

B.1 Dice score and Hausdorff Distance for Single Anatomical Regions

We report Dice score and Hausdorff Distance (HD, in pixels) for each organ of the medical datasets in Table 10, 11, 12, 13. Results consider training the segmentors with half of the weakly annotated train set (see Section 6.4.1). Notice that the average of the Dice score obtained for a method across classes is different from the multi-class Dice score (Crum, Camara, and Hill, 2006) reported in Table 6. In fact, given a multi-class segmentation mask \mathbf{y} and the prediction $\tilde{\mathbf{y}}$:

$$\frac{1}{c} \sum_{i=1}^c \frac{2|\tilde{\mathbf{y}}_i \cdot \mathbf{y}_i|}{|\tilde{\mathbf{y}}_i| + |\mathbf{y}_i|} \neq \frac{2|\tilde{\mathbf{y}} \cdot \mathbf{y}|}{|\tilde{\mathbf{y}}| + |\mathbf{y}|},$$

where i refers to each class and c is the number of classes.

Model	Dice			HD		
	RV	MYO	LV	RV	MYO	LV
UNet _{PCE}	69.3 ₁₁	76.4 ₀₆	84.2 ₀₇	84.7 ₂₉	79.5 ₂₃	74.4 ₂₈
UNet _{WPCE}	56.3 ₁₃	67.5 ₀₆	78.4 ₀₉	120.5 ₁₆	99.6 ₁₃	97.4 ₁₄
UNet _{CRF}	59.0 ₁₄	66.1 ₀₆	76.6 ₀₉	117.8 ₂₀	103.2 ₁₁	99.6 ₁₃
TS-UNet _{CRF}	27.2 ₁₀	40.8 ₀₈	47.9 ₁₂	133.9 ₁₂	111.9 ₀₉	115.6 ₁₀
PostDAE	55.6 ₁₂	66.7 ₀₇	80.6 ₀₇	103.4 ₁₈	88.7 ₁₂	80.6 ₁₅
UNet _D	40.4 ₁₅	59.7 ₀₈	75.3 ₀₉	33.5 ₁₀	25.7 ₁₂	25.2 ₁₄
ACCL	73.5 ₁₀	79.7 ₀₅	87.8 ₀₆	26.1 ₂₄	28.8 ₂₅	16.6 ₂₀
Ours	75.2 ₁₂	81.7 ₀₅	87.9 ₀₅	22.7 ₂₇	26.8 ₃₀	25.2 ₂₇

Table 10: Dice score and Hausdorff distance (HD) for single organs in ACDC. Abbreviations are as follows: RV: right ventricle, MYO: myocardium, LV: left ventricle.

Model	Dice	HD
	MYO	MYO
UNet _{PCE}	62.3 ₀₉	55.7 ₂₈
UNet _{WPCE}	59.1 ₀₇	52.4 ₂₃
UNet _{CRF}	60.4 ₀₈	53.0 ₂₇
TS-UNet _{CRF}	50.5 ₀₇	93.4 ₂₇
PostDAE	58.6 ₀₇	47.5 ₂₂
UNet _D	31.7 ₀₉	44.7 ₂₀
ACCL	65.9 ₀₈	24.0 ₁₉
Ours	65.5 ₀₈	27.5 ₂₅

Table 11: Dice score and Hausdorff distance (HD) for single organs in LVSC. MYO stands for myocardium.

Model	Dice				HD			
	L	RK	LK	S	L	RK	LK	S
UNet _{PCE}	43.5 ₀₇	21.3 ₀₄	9.1 ₀₃	25.9 ₀₇	133.5 ₀₁	157.1 ₀₄	151.7 ₀₁	133.8 ₀₇
UNet _{WPCE}	42.5 ₀₉	29.2 ₀₂	16.6 ₀₂	25.7 ₀₅	121.3 ₀₁	114.5 ₀₃	154.6 ₀₁	128.7 ₀₁
UNet _{CRF}	37.3 ₀₉	20.0 ₀₆	16.3 ₀₄	27.9 ₁₃	119.2 ₁₀	148.9 ₀₄	148.4 ₀₆	101.8 ₀₈
TS-UNet _{CRF}	41.1 ₁₂	13.2 ₀₄	6.2 ₀₂	16.5 ₀₆	110.6 ₁₈	157.7 ₀₄	153.8 ₀₅	163.8 ₀₈
PostDAE	32.8 ₀₇	57.9 ₀₇	57.1 ₀₆	58.4 ₁₁	100.1 ₁₃	192.0 ₀₀	184.8 ₀₆	192.0 ₀₀
UNet _D	60.2 ₀₅	46.4 ₁₀	46.9 ₀₆	41.3 ₁₂	59.9 ₀₂	93.5 ₃₇	151.0 ₀₃	123.1 ₀₃
ACCL	65.0 ₁₂	57.3 ₀₆	49.4 ₀₉	51.2 ₁₄	35.3 ₁₂	178.3 ₁₉	85.0 ₀₄	100.9 ₁₈
Ours	64.0 ₀₇	68.5 ₀₆	59.6 ₀₉	39.7 ₀₈	62.0 ₀₅	27.4 ₁₃	34.7 ₀₃	60.8 ₂₇

Table 12: Dice score and Hausdorff distance (HD) for single organs in CHAOS-T1. Abbreviations are as follows: L: liver, RK: right kidney, LK: left kidney, S: spleen.

Model	Dice				HD			
	L	RK	LK	S	L	RK	LK	S
UNet _{PCE}	48.4 ₀₈	23.9 ₀₅	9.7 ₀₂	27.7 ₀₇	133.1 ₀₁	155.9 ₀₄	151.3 ₀₁	114.6 ₀₉
UNet _{WPCE}	55.6 ₀₉	31.5 ₀₄	28.4 ₀₃	32.2 ₁₀	106.0 ₀₉	129.9 ₀₄	135.9 ₀₂	101.0 ₀₃
UNet _{CRF}	48.0 ₀₉	26.4 ₁₅	19.9 ₀₃	32.9 ₁₂	117.2 ₁₁	151.1 ₀₄	141.1 ₀₉	91.0 ₁₁
TS-UNet _{CRF}	44.5 ₁₀	7.0 ₀₃	6.4 ₀₃	18.4 ₀₅	90.8 ₁₄	157.6 ₀₄	154.5 ₀₅	157.3 ₀₈
PostDAE	43.4 ₀₇	57.9 ₀₇	57.5 ₀₆	58.4 ₁₁	76.4 ₁₂	192.0 ₀₀	190.2 ₀₃	192.0 ₀₀
UNet _D	63.6 ₀₄	53.0 ₁₀	45.0 ₀₈	34.1 ₁₀	52.8 ₀₆	127.7 ₂₃	108.5 ₀₁	113.0 ₀₆
ACCL	63.2 ₁₀	42.8 ₁₀	46.5 ₀₉	56.5 ₁₂	47.3 ₁₈	77.4 ₃₆	94.7 ₂₉	98.3 ₄₄
Ours	56.3 ₀₆	68.6 ₀₇	61.4 ₀₉	44.2 ₀₈	65.7 ₀₃	44.6 ₁₂	40.0 ₁₈	63.6 ₂₇

Table 13: Dice score and Hausdorff distance (HD) for single organs in CHAOS-T2. Abbreviations are as follows: L: liver, RK: right kidney, LK: left kidney, S: spleen.

B.2 Results on ACDC Evaluation Platform

In Table 14 we report metrics obtained after training on all the available ACDC data and testing on 50 extra patients using the online evaluation platform¹

Cardiac Phase	Dice			HD		
	RV	MYO	LV	RV	MYO	LV
End-diastole	89	81	93	16.6	45.3	20.9
End-systole	84	84	88	20.3	44.2	27.1

Table 14: Dice score and Hausdorff distance (HD) of the proposed approach trained on all the available ACDC data, and tested on 50 extra patients using the challenge server. Note that the server does not provide information about the standard deviation, nor a higher precision for the Dice score. Abbreviations are as follows: RV: right ventricle, MYO: myocardium, LV: left ventricle.

B.3 The Effect of the Dynamic Loss Weighting

As we discussed in Section 6.3.3, we optimise the loss function:

$$\mathcal{L} = a_0 \mathcal{L}_{SUP} + a_1 \mathcal{V}_{LS}(\Sigma),$$

We argue that, for a proper model convergence, it is important to prevent that during training one contribution prevails over the other. Thus, we suggest maintaining a fixed ratio between the amplitude of supervised and adversarial contributions using a *dynamic* value for a_0 :

$$a_0 = \frac{\|\mathcal{V}_{LS}(\Sigma)\|}{\|\mathcal{L}_{SUP}\|},$$

Empirically, if we remove the dynamic scaling and leave $a_0 = 1$, we observe a performance decrease, and obtain a Dice score of 71.6% (6.5%)

¹<https://acdc.creatis.insa-lyon.fr/#challenges>

on the test set. In particular, this happens because in the initial stages of training the supervised loss is the largest, while the adversarial loss becomes the main loss contribution during the final learning stages. As a result, the model ends its training relying more on the adversarial cost than the supervised one, and performance decreases.

B.4 Fully Supervised Learning

We conducted experiments to analyse model performance when it is trained with mask supervision rather than with scribbles. We report results in Table 15. As can be seen from the table, the same model works well with full supervision, and it improves performance when training with masks, rather than when using only scribble annotations.

We highlight that we conducted these experiments while keeping exactly the *same framework and hyperparameters*. It is possible that the choice of better hyperparameters could further improve the reported numbers (for example, changing the learning rate). However, since our scope is not related to training with full supervision, we don't investigate this further.

Supervision	Dataset				
	ACDC	LVSC	CHAOS-T1	CHAOS-T2	PPSS
Scribbles	84.3 ₀₄	65.5 ₀₈	56.8 ₀₅	57.8 ₀₄	74.6 ₀₄
Masks	84.3 ₀₂	68.8 ₀₇	65.7 ₀₃	65.9 ₀₂	76.9 ₀₄

Table 15: Training our method with scribbles and with mask supervision. We report the Dice average (standard deviation as subscript) obtained on the test data for each dataset.

B.5 Additional Figures

We report examples of segmentation failures for the proposed approach and for the benchmark models in Figure 50.

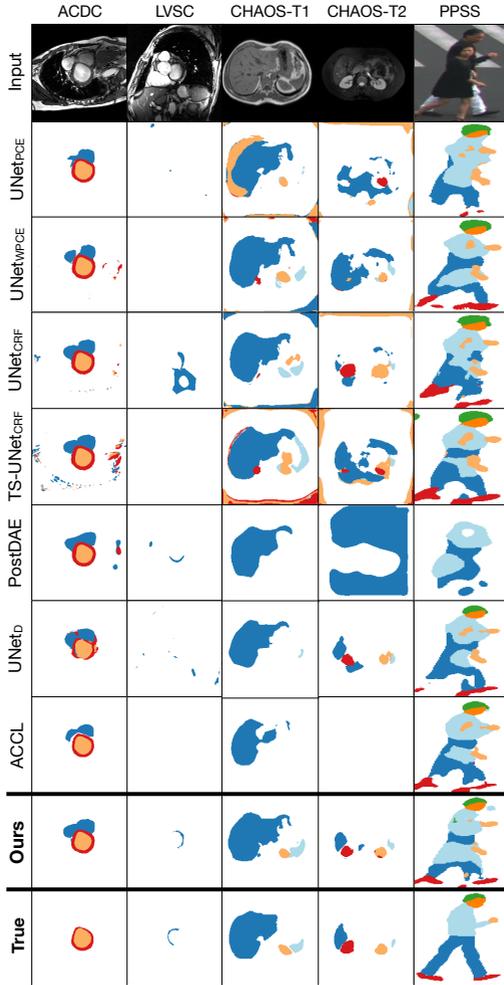


Figure 50: Example of model failures. In both ACDC and LVSC, the apical and the basal slices of the heart are the hardest to segment, due to intrinsic uncertainty of the cardiac boundaries, resulting in over/under-segmentations in all the models. For CHAOS, we show that all models make mistakes when the organ boundaries have low contrast, though our model preserves realistic outputs. In PPSS, we show that occlusions make the segmentation task harder; for example, if two people overlap, all model will try to segment both people, rather than only one.

Appendix C

Additional Experiments and Results of Chapter 8

C.1 Discriminator: Convergence and Memorisation

We report examples of the training and validation losses for the GAN discriminator $\Delta(\cdot)$. We use a Least-square GAN, whose discriminator loss to minimise is:

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

where $+1$ and -1 are the labels for *real* and *fake* (generated) images, respectively, and 0 is the equilibrium value.

We report examples of convergence modes in Figure 51 and Figure 52. We show losses on the training set on the left, losses on the validation set on the right. 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.

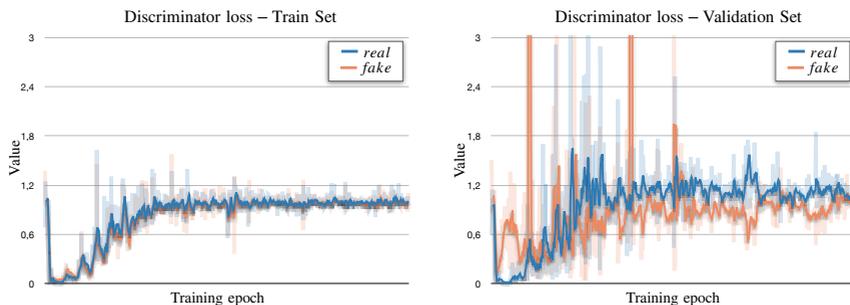


Figure 51: At convergence, the discriminator reaches an equilibrium stage where it always predicts the value 0, equidistant from the *true* and the *fake* labels. As a result, losses converge to the equilibrium value 1.0 both for train and validation.

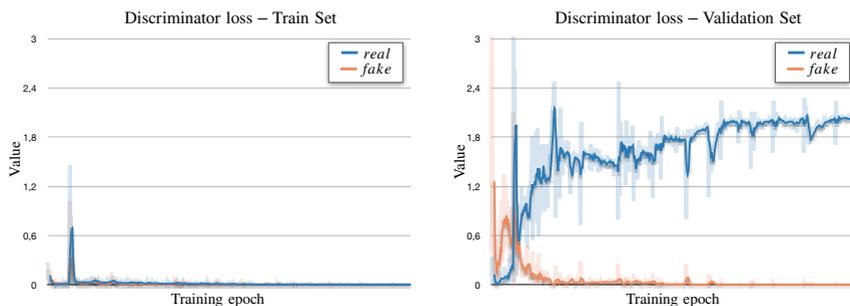


Figure 52: At convergence, the discriminator shows signals of memorisation. The discriminator memorises the *real* training images, and it predicts the label *fake* (i.e. the value -1) for any other case. During validation, the *fake* images are still classified correctly, while the *real* ones are classified as *fake* and the associated loss converges to the value of 2.0.

Bibliography

- Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. (2016). “Tensorflow: A System for Large-scale Machine Learning”. In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–283.
- Achille, Alessandro and Stefano Soatto (2018). “Emergence of Invariance and Disentanglement in Deep Representations”. In: *The Journal of Machine Learning Research* 19.1, pp. 1947–1980.
- Agatston, Arthur S., Warren R. Janowitz, Frank J. Hildner, Noel R. Zusmer, Manuel Viamonte, and Robert Detrano (1990). “Quantification of Coronary Artery Calcium Using Ultrafast Computed Tomography”. In: *Journal of the American College of Cardiology* 15.4, pp. 827–832.
- Alemi, Alexander A., Ian Fischer, Joshua V. Dillon, and Kevin Murphy (2016). “Deep Variational Information Bottleneck”. In: *arXiv preprint arXiv:1612.00410*.
- Arjovsky, Martin and Léon Bottou (2017). “Towards Principled Methods for Training Generative Adversarial Networks”. In: *arXiv preprint arXiv:1701.04862*.
- Arjovsky, Martin, Soumith Chintala, and Léon Bottou (2017). “Wasserstein Generative Adversarial Networks”. In: *International Conference on Machine Learning (ICML)*. PMLR, pp. 214–223.
- Asano, Yuki M., Christian Rupprecht, and Andrea Vedaldi (2020). “A Critical Analysis of Self-Supervision, or What We Can Learn From a Single Image”. In: *International Conference on Learning Representations (ICLR)*.
- Azad, Reza, Abdur R Fayjie, Claude Kauffman, Ismail Ben Ayed, Marco Pedersoli, and Jose Dolz (2021). “On the Texture Bias for Few-Shot

- CNN Segmentation". In: *Winter Conference on Applications of Computer Vision (WACV)*.
- Azadi, Samaneh, Michael Tschannen, Eric Tzeng, Sylvain Gelly, Trevor Darrell, and Mario Lucic (2019). "Semantic Bottleneck Scene Generation". In: *arXiv:1911.11357*.
- Bai, Wenjia, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E. Petersen, Yike Guo, Paul M. Matthews, and Daniel Rueckert (2019). "Self-Supervised Learning For Cardiac MR Image Segmentation by Anatomical Position Prediction". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, pp. 541–549.
- Bai, Wenjia, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew King, Paul M. Matthews, and Daniel Rueckert (2017). "Semi-supervised Learning For Network-Based Cardiac MR Image Segmentation". In: *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, pp. 253–260.
- Bai, Wenjia, Hideaki Suzuki, Chen Qin, Giacomo Tarroni, Ozan Oktay, Paul M. Matthews, and Daniel Rueckert (2018). "Recurrent Neural Networks for Aortic Image Sequence Segmentation With Sparse Annotations". In: *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, pp. 586–594.
- Baumgartner, Christian F., Kerem C. Tezcan, Krishna Chaitanya, Andreas M. Hötter, Urs J. Muehlemaier, Khoschy Schawkat, Anton S. Becker, Olivio Donati, and Ender Konukoglu (2019). "Phiseg: Capturing uncertainty in Medical Image Segmentation". In: *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, pp. 119–127.
- Bearman, Amy, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei (2016). "What's the point: Semantic segmentation with point supervision". In: *European Conference on Computer Vision (ECCV)*. Springer, pp. 549–565.
- Belharbi, Soufiane, Jérôme Rony, Jose Dolz, Ismail Ben Ayed, Luke McCaffrey, and Eric Granger (2020). "Deep Interpretable Classification and Weakly-Supervised Segmentation of Histology Images via Max-Min Uncertainty". In: *arXiv preprint arXiv:2011.07221*.
- Bengio, Yoshua (2009). "Learning Deep Architectures for AI". In: *Foundations and trends® in Machine Learning 2.1*, pp. 1–127.

- Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013). “Representation Learning: A Review and New Perspectives”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8, pp. 1798–1828.
- Bengio, Yoshua, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal (2020). “A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms”. In: *International Conference on Learning Representations (ICLR)*.
- Bengio, Yoshua, Nicholas Léonard, and Aaron Courville (2013). “Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation”. In: *arXiv preprint arXiv:1308.3432*.
- Benjamens, Stan, Pranavsinh Dhunoo, and Bertalan Meskó (2020). “The State of Artificial Intelligence-Based FDA-Approved Medical Devices and Algorithms: An Online Database”. In: *NPJ Digital Medicine* 3.1, pp. 1–8.
- Bernard, Olivier, 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 (2018). “Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?” In: *IEEE Transactions on Medical Imaging* 37.11, pp. 2514–2525.
- Berthelot, David, Thomas Schumm, and Luke Metz (2017). “BEGAN: Boundary Equilibrium Generative Adversarial Networks”. In: *International Conference on Learning Representations (ICLR)*.
- Blum, Avrim and Tom Mitchell (1998). “Combining Labeled and Unlabeled Data With Co-Training”. In: *Annual Conference on Computational Learning Theory*, pp. 92–100.
- Blumberg, Henry (1920). “Hausdorff’s Grundzüge der Mengenlehre”. In: *Bulletin of the American Mathematical Society* 27.3, pp. 116–129.
- Bond-Taylor, Sam, Adam Leach, Yang Long, and Chris G. Willcocks (2021). “Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models”. In: *arXiv preprint arXiv:2103.04922*.

- Bourlard, Hervé and Yves Kamp (1988). “Auto-Association by Multilayer Perceptrons and Singular Value Decomposition”. In: *Biological cybernetics* 59.4-5, pp. 291–294.
- Brendel, Wieland and Matthias Bethge (2019). “Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet”. In: *International Conference on Learning Representations (ICLR)*.
- Burgess, Christopher P., Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner (2018). “Understanding Disentangling in β -VAE”. In: *arXiv preprint arXiv:1804.03599*.
- Campello, Victor M., Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M. Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, et al. (2021). “Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge”. In: *IEEE Transactions on Medical Imaging*.
- Can, Yigit B., Krishna Chaitanya, Basil Mustafa, Lisa M Koch, Ender Konukoglu, and Christian F. Baumgartner (2018). “Learning to Segment Medical Images With Scribble-Supervision Alone”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, pp. 236–244.
- Caron, Mathilde, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin (2021). “Emerging Properties in Self-Supervised Vision Transformers”. In: *arXiv preprint arXiv:2104.14294*.
- Caruana, Rich (1997). “Multitask Learning”. In: *Machine Learning* 28.1, pp. 41–75.
- Caselles-Dupré, Hugo, Michael Garcia-Ortiz, and David Filliat (2019). “Symmetry-Based Disentangled Representation Learning requires Interaction with Environments”. In: *Advances in Neural Information Processing Systems (NeurIPS)*.
- Castro, Daniel C., Ian Walker, and Ben Glocker (2020). “Causality Matters in Medical Imaging”. In: *Nature Communications* 11.1, pp. 1–10.
- Chai, Lucy, Jonas Wulff, and Phillip Isola (2021). “Using Latent Space Regression to Analyze and Leverage Compositionality in GANs”. In: *International Conference on Learning Representations (ICLR)*.
- Chai, Seoin, Daniel Rueckert, and Ahmed E. Fetit (2020). “Reducing Textural Bias Improves Robustness of Deep Segmentation CNNs”. In: *arXiv preprint arXiv:2011.15093*.
- Chaitanya, Krishna, Ertunc Erdil, Neerav Karani, and Ender Konukoglu (2020). “Contrastive Learning of Global and Local Features for Med-

- ical Image Segmentation with Limited Annotations”. In: *Advances in Neural Information Processing Systems (NeurIPS)* 33.
- Chaitanya, Krishna, Neerav Karani, Christian F. Baumgartner, Anton Becker, Olivio Donati, and Ender Konukoglu (2019). “Semi-Supervised and Task-Driven Data Augmentation”. In: *International Conference on Information Processing in Medical Imaging (IPMI)*. Springer, pp. 29–41.
- Chapelle, Olivier, Bernhard Scholkopf, and Alexander Zien (2009). “Semi-supervised Learning”. In: *IEEE Transactions on Neural Networks* 20.3, pp. 542–542.
- Charles, J., T. Pfister, D. R. Magee, D. C. Hogg, and A. Zisserman. (2013). “Domain Adaptation for Upper Body Pose Tracking In Signed TV Broadcasts”. In: *British Machine Vision Conference (BMVC)*.
- Chartsias, Agisilaos, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David E. Newby, Rohan Dharmakumar, and Sotirios A. Tsaftaris (2019). “Disentangled Representation Learning In Cardiac Image Analysis”. In: *Medical Image Analysis* 58, p. 101535.
- Chartsias, Agisilaos, Giorgos Papanastasiou, Chengjia Wang, Scott Semple, David Newby, Rohan Dharmakumar, and Sotirios Tsaftaris (2020). “Disentangle, Align and Fuse for Multimodal and Semi-Supervised Image Segmentation”. In: *IEEE Transactions on Medical Imaging*.
- Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer (2002). “SMOTE: synthetic minority over-sampling technique”. In: *Journal of artificial intelligence research* 16, pp. 321–357.
- Chen, Chen, Cheng Ouyang, Giacomo Tarroni, Jo Schlemper, Huaqi Qiu, Wenjia Bai, and Daniel Rueckert (2019). “Unsupervised Multi-Modal Style Transfer for Cardiac MR Segmentation”. In: *International Workshop on Statistical Atlases and Computational Models of the Heart (STACOM)*. Springer, pp. 209–219.
- Chen, Liang, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert (2019). “Self-supervised learning for Medical Image Analysis using image context restoration”. In: *Medical Image Analysis* 58, p. 101539.
- Chen, Liang-Chieh, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille (2017). “DeepLab: Semantic Image Segmentation With Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.4, pp. 834–848.

- Chen, Xi, Yan Duan, Rein Houthoofd, John Schulman, Ilya Sutskever, and Pieter Abbeel (2016). "InfoGAN: Interpretable Representation Learning By Information Maximizing Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2172–2180.
- Cheplygina, Veronika, Marleen de Bruijne, and Josien P. W. Pluim (2019). "Not-so-Supervised: A Survey of Semi-Supervised, Multi-Instance, and Transfer Learning In Medical Image Analysis". In: *Medical Image Analysis* 54, pp. 280–296.
- Chu, Casey, Kentaro Minami, and Kenji Fukumizu (2020). "Smoothness and Stability in GANs". In: *International Conference on Learning Representations (ICLR)*.
- Clough, James R., Ilkay Oksuz, Nicholas Byrne, Veronika A. Zimmer, Julia A. Schnabel, and Andrew P. King (2020). "A Topological Loss Function for Deep-Learning Based Image Segmentation Using Persistent Homology". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Crum, William R., Oscar Camara, and Derek L.G. Hill (2006). "Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis". In: *IEEE Transactions on Medical Imaging* 25.11, pp. 1451–1461.
- Dai, Jifeng, Kaiming He, and Jian Sun (2015). "BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation". In: *International Conference on Computer Vision (ICCV)*, pp. 1635–1643.
- Dalca, Adrian V., John Guttag, and Mert R. Sabuncu (2018). "Anatomical Priors in Convolutional Networks for Unsupervised Biomedical Segmentation". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9290–9299.
- Dalca, Adrian V., Evan Yu, Polina Golland, Bruce Fischl, Mert R. Sabuncu, and Juan Eugenio Iglesias (2019). "Unsupervised Deep Learning For Bayesian Brain MRI Segmentation". In: *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, pp. 356–365.
- Delange, Matthias, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars (2021). "A Continual Learning Survey: Defying Forgetting in Classification Tasks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Della Latta, Daniele, Gianmarco Santini, Gabriele Valvano, Nicola Martini, Andrea Ripoli, Francesco Avogliero, Alberto Clemente, Carla

- Luisa Susini, Dante Chiappino, et al. (2018). “Contrast-free Estimation of Cardiac Volumes from CT Scans Using Deep Learning”. In: *European Congress of Radiology (ECR)*. DOI: [10.1594/ecr2018/C-1413](https://doi.org/10.1594/ecr2018/C-1413)
- Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei (2009). “ImageNet: A Large-Scale Hierarchical Image Database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 248–255.
- Denton, Emily L., Soumith Chintala, Rob Fergus, et al. (2015). “Deep Generative Image Models Using a Laplacian Pyramid of Adversarial Networks”. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1486–1494.
- Dice, Lee R. (1945). “Measures of the amount of ecologic association between species”. In: *Ecology* 26.3, pp. 297–302.
- Donahue, Jeff, Philipp Krähenbühl, and Trevor Darrell (2017). “Adversarial Feature Learning”. In: *International Conference on Learning Representations (ICLR)*.
- Dong, Qi, Shaogang Gong, and Xiatian Zhu (2017). “Class rectification hard mining for imbalanced deep learning”. In: *International Conference on Computer Vision (ICCV)*, pp. 1851–1860.
- Dorent, Reuben, Samuel Joutard, Jonathan Shapey, Neil Bisdas Sotirios A.nd Kitchen, Robert Bradford, Shakeel Saeed, Marc Modat, Sébastien Ourselin, and Tom Vercauteren (2020). “Scribble-based Domain Adaptation via Co-segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, pp. 479–489.
- Dou, Qi, Daniel C. Castro, Konstantinos Kamnitsas, and Ben Glocker (2019). “Domain Generalization via Model-Agnostic Learning Of Semantic Features”. In: *arXiv preprint arXiv: 1910.13580*.
- Dou, Qi, Lequan Yu, Hao Chen, Yueming Jin, Xin Yang, Jing Qin, and Pheng-Ann Heng (2017). “3D Deeply Supervised Network for Automated Segmentation of Volumetric Medical Images”. In: *Medical Image Analysis* 41, pp. 40–54.
- Dumoulin, Vincent, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville (2017). “Adversarially Learned Inference”. In: *International Conference on Learning Representations (ICLR)*.
- Dumoulin, Vincent, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio (2018). “Feature-wise

- Transformations". In: *Distill*. <https://distill.pub/2018/feature-wise-transformations>. DOI: [10.23915/distill.00011](https://doi.org/10.23915/distill.00011)
- Fabio Cozman, Ira Cohen (2006). "Risks of Semi-supervised Learning". In: *Semi-supervised Learning*, pp. 56–72.
- Fedus, William, Mihaela Rosca, Balaji Lakshminarayanan, Andrew M. Dai, Shakir Mohamed, and Ian Goodfellow (2017). "Many Paths to Equilibrium: GANs Do Not Need to Decrease a Divergence at Every Step". In: *arXiv preprint arXiv:1710.08446*.
- Feyjje, Abdur R., Reza Azad, Marco Pedersoli, Claude Kauffman, Ismail Ben Ayed, and Jose Dolz (2020). "Semi-Supervised Few-Shot Learning For Medical Image Segmentation". In: *arXiv preprint arXiv:2003.08462*.
- Fonseca, Carissa G., Michael Backhaus, David A. Bluemke, Randall D. Britten, Jae Do Chung, Brett R. Cowan, Ivo D. Dinov, J. Paul Finn, Peter J. Hunter, Alan H. Kadish, et al. (2011). "The Cardiac Atlas Project - An Imaging Database for Computational Modeling and Statistical Atlases of the Heart". In: *Bioinformatics* 27.16, pp. 2288–2295.
- Freeman, Jeremy, Corey M. Ziemba, David J. Heeger, Eero P. Simoncelli, and J. Anthony Movshon (2013). "A Functional and Perceptual Signature of the Second Visual Area in Primates". In: *Nature Neuroscience* 16.7, p. 974.
- Fu, Jun, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu (2019). "Dual Attention Network for Scene Segmentation". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3146–3154.
- Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel (2018). "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness". In: *arXiv preprint arXiv:1811.12231*.
- Gong, Yunye, Srikrishna Karanam, Ziyang Wu, Kuan-Chuan Peng, Jan Ernst, and Peter C. Doerschuk (2018). "Learning Compositional Visual Concepts With Mutual Consistency". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8659–8668.
- Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2014). "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2672–2680.
- Grady, Leo (2006). "Random Walks for Image Segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.11, pp. 1768–1783.

- Guan, Hao and Mingxia Liu (2021). "Domain Adaptation for Medical Image Analysis: A Survey". In: *arXiv preprint arXiv:2102.09508*.
- Gulrajani, Ishaan, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C. Courville (2017). "Improved Training of Wasserstein GANs". In: *NeurIPS* 30. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pp. 5767–5777. URL: <http://papers.nips.cc/paper/7159-improved-training-of-wasserstein-GANs.pdf>
- Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger (2017). "On Calibration of Modern Neural Networks". In: *International Conference on Machine Learning (ICML)*. PMLR, pp. 1321–1330.
- Guo, Lan-Zhe, Yu-Feng Li, Ming Li, Jin-Feng Yi, Bo-Wen Zhou, and Zhi-Hua Zhou (2019). "Reliable Weakly Supervised Learning: Maximize Gain and Maintain Safeness". In: *arXiv preprint arXiv:1904.09743*.
- Hadsell, Raia, Sumit Chopra, and Yann LeCun (2006). "Dimensionality Reduction by Learning an Invariant Mapping". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2. IEEE, pp. 1735–1742.
- Havaei, Mohammad, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle (2017). "Brain Tumor Segmentation With Deep Neural Networks". In: *Medical Image Analysis* 35, pp. 18–31.
- Hermann, Katherine L. and Simon Kornblith (2020). "Exploring the Origins and Prevalence of Texture Bias in Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems (NeurIPS)*.
- Higgins, Irina, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner (2018). "Towards a Definition of Disentangled Representations". In: *arXiv preprint arXiv:1812.02230*.
- Higgins, Irina, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner (2017). " β -VAE: Learning Basic Visual Concepts With a Constrained Variational Framework". In: *International Conference on Learning Representations (ICLR)*. Vol. 3.
- Hinton, Geoffrey, Alex Krizhevsky, Navdeep Jaitly, Tijmen Tieleman, and Yichuan Tang (2012). "Does the Brain Do Inverse Graphics". In: *Brain and Cognitive Sciences Fall Colloquium*. Vol. 2.
- Hsieh, Jun-Ting, Bingbin Liu, De-An Huang, Li F. Fei-Fei, and Juan Carlos Niebles (2018). "Learning to Decompose and Disentangle Repre-

- sentations for Video Prediction”. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 517–526.
- Huang, Xun, Ming-Yu Liu, Serge Belongie, and Jan Kautz (2018). “Multi-modal Unsupervised Image-to-image Translation”. In: *European Conference on Computer Vision (ECCV)*, pp. 179–196.
- Ioffe, Sergey and Christian Szegedy (2015). “Batch Normalization: Accelerating Deep Network Training By Reducing Internal Covariate Shift”. In: *International Conference on Machine Learning (ICML)*. PMLR, pp. 448–456.
- Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros (2017). “Image-to-Image Translation With Conditional Adversarial Networks”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1125–1134.
- Jaccard, Paul (1912). “The distribution of the flora in the alpine zone. 1”. In: *New phytologist* 11.2, pp. 37–50.
- Japkowicz, Nathalie, Stephen Jose Hanson, and Mark A Gluck (2000). “Nonlinear Autoassociation Is Not Equivalent to PCA”. In: *Neural Computation* 12.3, pp. 531–545.
- Jetley, Saumya, Nicholas A. Lord, Namhoon Lee, and Philip H. S. Torr (2018). “Learn To Pay Attention”. In: *International Conference on Learning Representations (ICLR)*. eprint: [1804.02391](https://arxiv.org/abs/1804.02391).
- Ji, Zhanghexuan, Yan Shen, Chunwei Ma, and Mingchen Gao (2019). “Scribble-based Hierarchical Weakly Supervised Learning for Brain Tumor Segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, pp. 175–183.
- Jiang, Haochuan, Agisilaos Chartsias, Xinheng Zhang, Giorgos Papanastasiou, Scott Sempale, Mark Dweck, David Sempale, Rohan Dharmakumar, and Sotirios A. Tsaftaris (2020). “Semi-supervised Pathology Segmentation with Disentangled Representations”. In: *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*. Springer, pp. 62–72.
- Jolicoeur-Martineau, Alexia (2019). “The Relativistic Discriminator: A Key Element Missing From Standard GAN”. In: *International Conference on Learning Representations (ICLR)*.
- (2020). “On relativistic f-divergences”. In: *International Conference on Machine Learning (ICML)*. PMLR, pp. 4931–4939.
- Joyce, Thomas, Agisilaos Chartsias, and Sotirios A. Tsaftaris (2017). “Robust Multi-Modal MR Image Synthesis”. In: *International Conference*

- on *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, pp. 347–355.
- (2018). “Deep Multi-Class Segmentation Without Ground-Truth Labels”. In: *Medical Imaging with Deep Learning (MIDL)*.
- Jurdi, Rosana El, Caroline Petitjean, Paul Honeine, Veronika Cheplygina, and Fahed Abdallah (2020). “High-level Prior-based Loss Functions for Medical Image Segmentation: A Survey”. In: *arXiv preprint arXiv:2011.08018*.
- (2021). “A Surprisingly Effective Perimeter-based Loss for Medical Image Segmentation”. In: *Medical Imaging with Deep Learning (MIDL)*.
- Karani, Neerav, Ertunc Erdil, Krishna Chaitanya, and Ender Konukoglu (2021). “Test-Time Adaptable Neural Networks for Robust Medical Image Segmentation”. In: *Medical Image Analysis* 68, p. 101907.
- Karpathy, Andrej, Pieter Abbeel, Greg Brockman, Peter Chen, Vicki Cheung, Rocky Duan, Ian Goodfellow, Durk Kingma, Jonathan Ho, Rein Houthoofd, Tim Salimans, John Schulman, Ilya Sutskever, and Wojciech Zaremba (2016). *Generative Models*. <https://openai.com/blog/generative-models>. Accessed: 2020-07-22.
- Karras, Tero, Timo Aila, Samuli Laine, and Jaakko Lehtinen (2017). “Progressive Growing of GANs for Improved Quality, Stability, and Variation”. In: *International Conference on Learning Representations (ICLR)*.
- Karras, Tero, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila (2020). “Training Generative Adversarial Networks With Limited Data”. In: *Advances in Neural Information Processing Systems (NeurIPS)* 33.
- Kavur, A. Emre, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. (2021). “CHAOS Challenge-combined (CT-MR) Healthy Abdominal Organ Segmentation”. In: *Medical Image Analysis* 69, p. 101950.
- Kavur, A. Emre, Naciye Sinem Gezer, Mustafa Barış, Yusuf Şahin, Savaş Özkan, Bora Baydar, Ulaş Yüksel, Çağlar Kılıkçier, Şahin Olut, Gözde Bozdağı Akar, et al. (2020). “Comparison of Semi-Automatic and Deep Learning-Based Automatic Methods for Liver Segmentation in Living Liver Transplant Donors”. In: *Diagnostic and Interventional Radiology* 26.1, p. 11.
- Kavur, Ali Emre, M. Alper Selver, Oğuz Dicle, Mustafa Barış, and N. Sinem Gezer (Apr. 2019). *CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data*. Version v1.03. Zenodo.

DOI: [10.5281/zenodo.3362844](https://doi.org/10.5281/zenodo.3362844), URL: <https://doi.org/10.5281/zenodo.3362844>

- Kayhan, Osman Semih and Jan C. van Gemert (2020). "On Translation Invariance in CNNs: Convolutional Layers Can Exploit Absolute Spatial Location". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14274–14285.
- Kervadec, Hoel, Jose Dolz, Meng Tang, Eric Granger, Yuri Boykov, and Ismail Ben Ayed (2019). "Constrained-CNN Losses for Weakly Supervised Segmentation". In: *Medical Image Analysis* 54, pp. 88–99.
- Kervadec, Hoel, Jose Dolz, Shanshan Wang, Eric Granger, and Ismail Ben Ayed (2020). "Bounding Boxes for Weakly Supervised Segmentation: Global Constraints Get Close to Full Supervision". In: *Medical Imaging with Deep Learning (MIDL)*.
- Khoreva, Anna, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele (2017). "Simple Does It: Weakly Supervised Instance and Semantic Segmentation". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 876–885.
- Khrulkov, Valentin, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Osledeets, and Victor Lempitsky (2020). "Hyperbolic Image Embeddings". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6418–6428.
- Kim, Hyunjik and Andriy Mnih (2018). "Disentangling by Factorising". In: *International Conference on Machine Learning (ICML)*. PMLR, pp. 2649–2658.
- Kim, Myeongjin and Hyeran Byun (2020). "Learning Texture Invariant Representation for Domain Adaptation of Semantic Segmentation". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Kim, Youngjin, Minjung Kim, and Gunhee Kim (2018). "Memorization Precedes Generation: Learning Unsupervised GANs With Memory Networks". In: *International Conference on Machine Learning (ICML)*.
- Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations (ICLR)*.
- Kingma, Diederik P. and Max Welling (2014). "Auto-Encoding Variational Bayes". In: *International Conference on Learning Representations (ICLR)*.
- Kingma, Durk P, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling (2016). "Improved Variational Inference With Inverse Autoregressive Flow". In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4743–4751.

- Kirkpatrick, James, 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 (2017). "Overcoming Catastrophic Forgetting in Neural Networks". In: *Proceedings of the National Academy of Sciences* 114.13, pp. 3521–3526.
- Kodali, Naveen, Jacob Abernethy, James Hays, and Zsolt Kira (2017). "On Convergence and Stability of GANs". In: *arXiv preprint arXiv:1705.07215*.
- Kohl, Simon, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger (2018). "A Probabilistic U-net for Segmentation of Ambiguous Images". In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6965–6975.
- Kong, Shu and Charless C. Fowlkes (2018). "Recurrent Pixel Embedding For Instance Grouping". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9018–9028.
- Krähenbühl, Philipp and Vladlen Koltun (2011). "Efficient Inference in Fully Connected Crfs With Gaussian Edge Potentials". In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 109–117.
- Krebs, Julian, Hervé Delingette, Boris Maillhé, Nicholas Ayache, and Tommaso Mansi (2019). "Learning a Probabilistic Model for Diffeomorphic Registration". In: *IEEE Transactions on Medical Imaging* 38.9, pp. 2165–2176.
- Kubat, Miroslav, Stan Matwin, et al. (1997). "Addressing the curse of imbalanced training sets: one-sided selection". In: *International Conference on Machine Learning (ICML)*. Vol. 97. Citeseer, pp. 179–186.
- Kurach, Karol, Mario Lučić, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly (2019). "A Large-Scale Study on Regularization and Normalization in GANs". In: *International Conference on Machine Learning (ICML)*. PMLR, pp. 3581–3590.
- Kurzman, Lironne, David Vazquez, and Issam Laradji (2019). "Class-Based Styling: Real-time Localized Style Transfer with Semantic Segmentation". In: *ICCV Workshops*.
- Landau, Barbara, Linda B. Smith, and Susan S. Jones (1988). "The Importance of Shape in Early Lexical Learning". In: *Cognitive development* 3.3, pp. 299–321.
- Larrazabal, Agostina J., César Martínez, Ben Glocker, and Enzo Ferrante (2020). "Post-DAE: Anatomically plausible segmentation via post-

- processing with Denoising Autoencoders". In: *IEEE Transactions on Medical Imaging*.
- Lee, Chen-Yu, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu (2015). "Deeply-Supervised Nets". In: *Artificial intelligence and statistics*, pp. 562–570.
- Lee, Hsin-Ying, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang (2018). "Diverse Image-to-Image Translation via Disentangled Representations". In: *European Conference on Computer Vision (ICCV)*, pp. 36–52.
- Leshno, Moshe, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken (1993). "Multilayer Feedforward Networks With a Nonpolynomial Activation Function Can Approximate Any Function". In: *Neural Networks* 6.6, pp. 861–867.
- Li, Hanchao, Pengfei Xiong, Jie An, and Lingxue Wang (2018). "Pyramid Attention Network for Semantic Segmentation". In: *British Machine Vision Conference (BMVC)*.
- Li, Haoliang, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot (2018). "Domain Generalization With Adversarial Feature Learning". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5400–5409.
- Li, Yuheng, Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee (2020). "MixNMatch: Multifactor Disentanglement and Encoding for Conditional Image Generation". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8039–8048.
- Liao, Haofu, Wei-An Lin, S. Kevin Zhou, and Jiebo Luo (2019). "ADN: Artifact Disentanglement Network for Unsupervised Metal Artifact Reduction". In: *IEEE Transactions on Medical Imaging* 39.3, pp. 634–643.
- Lin, Di, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun (2016). "Scribble-Sup: Scribble-supervised Convolutional Networks for Semantic Segmentation". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3159–3167.
- Lin, Long-Ji (1992). "Self-Improving Reactive Agents Based On Reinforcement Learning, Planning and Teaching". In: *Machine learning* 8.3-4, pp. 293–321.
- Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár (2017). "Focal Loss for Dense Object Detection". In: *International Conference on Computer Vision (ICCV)*, pp. 2980–2988.
- Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick (2014). "Mi-

- crosoft COCO: Common Objects in Context". In: *European Conference on Computer Vision (ICCV)*. Springer, pp. 740–755.
- Liu, Ming-Yu, Thomas Breuel, and Jan Kautz (2017). "Unsupervised Image-to-Image Translation Networks". In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 700–708.
- Liu, Quande, Qi Dou, and Pheng-Ann Heng (2020). "Shape-Aware Meta-Learning For Generalizing Prostate MRI Segmentation to Unseen Domains". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, pp. 475–485.
- Liu, Rosanne, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski (2018). "An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution". In: *Advances in Neural Information Processing Systems (NeurIPS)* 31, pp. 9605–9616.
- Liu*, Xiao, Spyridon Thermos*, Gabriele Valvano*, Agisilaos Chatsias, Alison O’Neil, and Sotirios A. Tsaftaris (2021). "Measuring the Biases and Effectiveness of Content-Style Disentanglement". In: *Proceedings of the British Machine Vision Conference 2021*. British Machine Vision Association.
- Liu, Yanbin, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang (2019). "Learning to propagate labels: Transductive propagation network for few-shot learning". In: *International Conference on Learning Representations (ICLR)*.
- Locatello, Francesco, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem (2019). "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations". In: *International Conference on Learning Representations Workshops (ICLRW)*, pp. 4114–4124.
- (2020). "A Commentary on the Unsupervised Learning Of Disentangled Representations". In: *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 13681–13684.
- Lorenz, Dominik, Leonard Bereska, Timo Milbich, and Bjorn Ommer (2019). "Unsupervised Part-Based Disentangling Of Object Shape and Appearance". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10964.
- Lowe, David G. (2004). "Distinctive Image Features From Scale-Invariant Keypoints". In: *International journal of computer vision* 60.2, pp. 91–110.
- Lucic, Mario, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet (2017). "Are GANs Created Equal? A Large-scale Study". In: *arXiv preprint arXiv:1711.10337*.

- Luo, Ping, Xiaogang Wang, and Xiaoou Tang (2013). "Pedestrian Parsing Via Deep Compositional Network". In: *International Conference on Computer Vision (ICCV)*, pp. 2648–2655.
- Luo, Yawei, Zhedong Zheng, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang (2018). "Macro-Micro Adversarial Network for Human Parsing". In: *European Conference on Computer Vision (ICCV)*, pp. 418–434.
- Mao, Xudong, Qing Li, Haoran Xie, Raymond Yiu Keung Lau, Zhen Wang, and Stephen Paul Smolley (2018). "On the Effectiveness of Least Squares Generative Adversarial Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Mai, Zheda, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner (2021). "Online Continual Learning in Image Classification: An Empirical Survey". In: *arXiv preprint arXiv:2101.10423*.
- Malhotra, Gaurav and Jeffrey Bowers (2019). "The contrasting roles of shape in human vision and convolutional neural networks". In: *Annual Conference of the Cognitive Science Society*, pp. 2261–2267.
- Mao, Xin, Zhaoyu Su, Pin Siang Tan, Jun Kang Chow, and Yu-Hsing Wang (2019). "Is Discriminator a Good Feature Extractor?" In: *arXiv preprint arXiv:1912.00789*.
- Mao, Xudong, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley (2017). "Least Squares Generative Adversarial Networks". In: *International Conference on Computer Vision (ICCV)*, pp. 2794–2802.
- Martini, Nicola, Alessio Vatti, Andrea Ripoli, Sara Salaris, Gianmarco Santini, Gabriele Valvano, Maria Filomena Santarelli, Dante Chiappino, and Daniele Della Latta (2019). "Robust Reconstruction of Cardiac T1 Maps Using RNNs". In: *Medical Imaging with Deep Learning (MIDL)*.
- Meng, Qingjie, Jacqueline Matthew, Veronika A. Zimmer, Alberto Gomez, David F. A. Lloyd, Daniel Rueckert, and Bernhard Kainz (2020). "Mutual Information-based Disentangled Neural Networks for Classifying Unseen Categories in Different Domains: Application to Fetal Ultrasound Imaging". In: *IEEE Transactions on Medical Imaging*.
- Milletari, Fausto, Nassir Navab, and Seyed-Ahmad Ahmadi (2016). "V-net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation". In: *Fourth international conference on 3D vision (3DV)*. IEEE, pp. 565–571.
- Mirza, Mehdi and Simon Osindero (2014). "Conditional generative adversarial nets". In: *arXiv preprint arXiv:1411.1784*.

- Miyato, Takeru, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida (2018). "Spectral Normalization for Generative Adversarial Networks". In: *International Conference on Learning Representations (ICLR)*.
- Monteiro, Miguel, Mário A.T. Figueiredo, and Arlindo L. Oliveira (2018). "Conditional Random Fields as Recurrent Neural Networks for 3D Medical Imaging Segmentation". In: *arXiv:1807.07464*.
- Müller, Rafael, Simon Kornblith, and Geoffrey E Hinton (2019). "When Does Label Smoothing Help?" In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4694–4703.
- Nagarajan, Vaishnavh, Colin Raffel, and I. Goodfellow (2018). "Theoretical Insights Into Memorization in GANs". In: *NeurIPS Workshop*.
- Neto, José A. Gonçalves, Mohamed Elazzazzi, Ersan Altun, and Richard C Semelka (2008). "When Should Abdominal Magnetic Resonance Imaging Be Used?" In: *Clinical Gastroenterology and Hepatology* 6.6, pp. 610–615.
- Ngo, Phuc Cuong, Amadeus Aristo Winarto, Connie Khor Li Kou, Sojeong Park, Farhan Akram, and Hwee Kuan Lee (2019). "Fence GAN: Towards Better Anomaly Detection". In: *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, pp. 141–148.
- Nguyen, Cuong V, Yingzhen Li, Thang D Bui, and Richard E Turner (2018). "Variational Continual Learning". In: *International Conference on Learning Representations (ICLR)*.
- Nickel, Maximillian and Douwe Kiela (2017). "Poincaré Embeddings for Learning Hierarchical Representations". In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6338–6347.
- Nie, Weili, Tero Karras, Animesh Garg, Shoubhik Debnath, Anjul Patney, Ankit Patel, and Animashree Anandkumar (2020). "Semi-Supervised StyleGAN for Disentanglement Learning". In: *International Conference on Machine Learning (ICML)*. PMLR, pp. 7360–7369.
- Noroozi, Vahid, Sara Bahaadini, Lei Zheng, Sihong Xie, Weixiang Shao, and S Yu Philip (2018). "Semi-Supervised Deep Representation Learning For Multi-View Problems". In: *2018 IEEE International Conference on Big Data*. IEEE, pp. 56–64.
- Nosrati, Masoud S. and Ghassan Hamarneh (2016). "Incorporating Prior Knowledge in Medical Image Segmentation: A Survey". In: *arXiv pre-print arXiv:1607.01092*.
- Nosrati, Masoud S. and Ghassan Hamarneh (2016). "Incorporating Prior Knowledge in Medical Image Segmentation: A Survey". In: *arXiv:1607.01092*.

- O'Regan, J. Kevin and Alva Noë (2001). "A Sensorimotor Account of Vision and Visual Consciousness". In: *Behavioral and brain sciences* 24.5, p. 939.
- Oktay, Ozan, 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 (2017). "Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation". In: *IEEE Transactions on Medical Imaging* 37.2, pp. 384–395.
- Oktay, Ozan, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, et al. (2018). "Attention U-net: Learning Where to Look For the Pancreas". In: *Medical Imaging with Deep Learning (MIDL)*.
- Ørting, Silas, Andrew Doyle, Matthias Hirth Arno van Hilten, Oana Inel, Christopher R. Madan, Panagiotis Mavridis, Helen Spiers, and Veronika Cheplygina (2019). "A Survey of Crowdsourcing in Medical Image Analysis". In: *arXiv:1902.09159*.
- Quali, Yassine, Céline Hudelot, and Myriam Tami (2020). "An Overview of Deep Semi-supervised Learning". In: *arXiv preprint arXiv:2006.05278*.
- Ouyang, Cheng, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert (2020). "Self-Supervision With Superpixels: Training Few-Shot Medical Image Segmentation Without Annotation". In: *European Conference on Computer Vision (ECCV)*. Springer, pp. 762–780.
- Painchaud, Nathan, Youssef Skandarani, Thierry Judge, Olivier Bernard, Alain Lalonde, and Pierre-Marc Jodoin (2020). "Cardiac Segmentation With Strong Anatomical Guarantees". In: *IEEE Transactions on Medical Imaging* 39.11, pp. 3703–3713.
- Painchaud, Nathan, Youssef Skandarani, Thierry Judge, Olivier Bernard, Alain Lalonde, and Pierre-Marc Jodoin (2019). "Cardiac MRI Segmentation With Strong Anatomical Guarantees". In: *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, pp. 632–640.
- Park, Seong Ho and Kyunghwa Han (2018). "Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction". In: *Radiology* 286.3, pp. 800–809.
- Park, Taesung, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu (2019). "Semantic Image Synthesis With Spatially-Adaptive Normalization".

- In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2337–2346.
- Patel, Gaurav and Jose Dolz (2021). “Weakly Supervised Segmentation With Cross-Modality Equivariant Constraints”. In: *arXiv preprint arXiv: 2104.02488*.
- Perez, Ethan, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville (2018). “FiLM: Visual Reasoning With a General Conditioning Layer”. In: *AAAI Conference on Artificial Intelligence (AAAI)*. Vol. 32.
- Petersen, Steffen E., Musa Abdulkareem, and Tim Leiner (2019a). “Artificial Intelligence Will Transform Cardiac Imaging—Opportunities and Challenges”. In: *Frontiers in Cardiovascular Medicine* 6, p. 133. ISSN: 2297-055X. DOI: [10.3389/fcvm.2019.00133](https://doi.org/10.3389/fcvm.2019.00133), URL: <https://www.frontiersin.org/article/10.3389/fcvm.2019.00133>.
- Petersen, Steffen Erhard, Musa Abdulkareem, and Tim Leiner (2019b). “Artificial Intelligence Will Transform Cardiac Imaging - Opportunities and Challenges”. In: *Frontiers in cardiovascular medicine* 6, p. 133.
- Qiao, Siyuan, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille (2018). “Deep Co-Training for Semi-Supervised Image Recognition”. In: *European Conference on Computer Vision (ICCV)*, pp. 135–152.
- Qin, Chen, Bibo Shi, Rui Liao, Tommaso Mansi, Daniel Rueckert, and Ali Kamen (2019). “Unsupervised Deformable Registration for Multi-Modal Images via Disentangled Representations”. In: *International Conference on Information Processing in Medical Imaging (IPMI)*. Springer, pp. 249–261.
- Qu, Hui, Pengxiang Wu, Qiaoying Huang, Jingru Yi, Zhennan Yan, Kang Li, Gregory M. Riedlinger, Subhrajyoti De, Shaoting Zhang, and Dimitris N. Metaxas (2020). “Weakly Supervised Deep Nuclei Segmentation Using Partial Points Annotation in Histopathology Images”. In: *IEEE Transactions on Medical Imaging* 39.11, pp. 3655–3666.
- Radford, Alec, Luke Metz, and Soumith Chintala (2015). “Unsupervised Representation Learning With Deep Convolutional Generative Adversarial Networks”. In: *arXiv preprint arXiv:1511.06434*.
- Rajchl, Martin, Lisa M. Koch, Christian Ledig, Jonathan Passerat-Palmbach, Kazunari Misawa, Kensaku Mori, and Daniel Rueckert (2017). “Employing Weak Annotations for Medical Image Analysis Problems”. In: *arXiv:1708.06297*.
- Recht, Benjamin, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar (2018). “Do Cifar-10 Classifiers Generalize to Cifar-10?” In: *arXiv preprint arXiv:1806.00451*.

- Reinke, Annika, Matthias Eisenmann, Minu D Tizabi, Carole H Sudre, Tim Rädtsch, Michela Antonelli, Tal Arbel, Spyridon Bakas, M Jorge Cardoso, Veronika Cheplygina, Keyvan others Farahani, Ben Glocker, Doreen Heckmann-Nötzel, Fabian Isensee, Pierre Jannin, Charles E. Kahn, Jens Kleesiek, Tahsin Kurc, Michal Kozubek, Bennett A. Landman, Geert Litjens, Klaus Maier-Hein, Bjoern Menze, Henning Müller, Jens Petersen, Mauricio Reyes, Nicola Rieke, Bram Stieltjes, Ronald M. Summers, Sotirios A. Tsaftaris, Bram van Ginneken, Annette Kopp-Schneider, Paul Jäger, and Lena Maier-Hein (2021). “Common Limitations of Image Processing Metrics: A Picture Story”. In: *arXiv preprint arXiv:2104.05642*.
- Rezende, Danilo Jimenez, Shakir Mohamed, and Daan Wierstra (2014). “Stochastic Backpropagation and Approximate Inference in Deep Generative Models”. In: *arXiv preprint arXiv:1401.4082*.
- Ringer, Sam, Will Williams, Tom Ash, Remi Francis, and David MacLeod (2019). “Texture Bias Of CNNs Limits Few-Shot Classification Performance”. In: *NeurIPS Workshop on Meta-Learning*.
- Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). “U-net: Convolutional Networks for Biomedical Image Segmentation”. In: *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, pp. 234–241.
- Roth, Holger R., Dong Yang, Ziyue Xu, Xiaosong Wang, and Daguang Xu (2020). “Going to Extremes: Weakly Supervised Medical Image Segmentation”. In: *arXiv preprint arXiv:2009.11988*.
- Salimans, Tim, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen (2016). “Improved Techniques for Training GANs”. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2234–2242.
- Santini, Gianmarco, Daniele Della Latta, Nicola Martini, Gabriele Valvano, Andrea Gori, Andrea Ripoli, Carla L. Susini, Luigi Landini, and Dante Chiappino (2017). “An Automatic Deep Learning Approach for Coronary Artery Calcium Segmentation”. In: *European Medical and Biological Engineering Conference & Nordic-Baltic Conference on Biomedical Engineering and Medical Physics (EMBECE & NBC)*. Springer, pp. 374–377. DOI: [10.1007/978-981-10-5122-7_94](https://doi.org/10.1007/978-981-10-5122-7_94).
- Saxena, Divya and Jiannong Cao (2020). “Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions”. In: *arXiv preprint arXiv:2005.00065*.

- Schaul, Tom, John Quan, Ioannis Antonoglou, and David Silver (2016). "Prioritized experience replay". In: *International Conference on Learning Representations (ICLR)*.
- Schlemper, Jo, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert (2019). "Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images". In: *Medical Image Analysis* 53, pp. 197–207.
- Schölkopf, Bernhard, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij (2013). "Semi-supervised Learning in Causal and Anticausal Settings". In: *Empirical Inference*. Springer, pp. 129–141.
- Schonfeld, Edgar, Bernt Schiele, and Anna Khoreva (2020). "A U-net Based Discriminator for Generative Adversarial Networks". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8207–8216.
- Schwartz, Odelia and Eero P. Simoncelli (2001). "Natural Signal Statistics and Sensory Gain Control". In: *Nature Neuroscience* 4.8, pp. 819–825.
- Shaban, Amirreza, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots (2017). "One-Shot Learning for Semantic Segmentation". In: *British Machine Vision Conference (BMVC)*.
- Shrivastava, Abhinav, Abhinav Gupta, and Ross Girshick (2016). "Training Region-Based Object Detectors With Online Hard Example Mining". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 761–769.
- Shrivastava, Ashish, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb (2017). "Learning From Simulated and Un-supervised Images Through Adversarial Training". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2107–2116.
- Sinha, Abhishek, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, and Stefano Ermon (2021). "Negative Data Augmentation". In: *International Conference on Learning Representations (ICLR)*.
- Sinha, Ashish and Jose Dolz (2020). "Multi-Scale Self-Guided Attention for Medical Image Segmentation". In: *IEEE Journal of Biomedical and Health Informatics*.
- Sinha, Samarth, Animesh Garg, and Hugo Larochelle (2020). "Curriculum By Smoothing". In: *Advances in Neural Information Processing Systems (NeurIPS)* 33.
- Smith, Leslie N. (2017). "Cyclical Learning Rates for Training Neural Networks". In: *Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp. 464–472.

- Sønderby, Casper Kaae, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár (2017). “Amortised map inference for image super-resolution”. In: *International Conference on Learning Representations (ICLR)*.
- Sørensen, Thorvald (1948). “A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons”. In: *Royal Danish Academy of Sciences and Letters* 5.4, pp. 1–34.
- Souly, Nasim, Concetto Spampinato, and Mubarak Shah (2017). “Semi Supervised Semantic Segmentation Using Generative Adversarial Network”. In: *International Conference on Computer Vision (ICCV)*, pp. 5688–5696.
- Suinesiaputra, Avan, 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 (2014). “A Collaborative Resource to Build Consensus for Automated Left Ventricular Segmentation of Cardiac MR Images”. In: *Medical Image Analysis* 18.1, pp. 50–62.
- Sun, Yu, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt (2020). “Test-Time Training With Self-Supervision for Generalization Under Distribution Shifts”. In: *International Conference on Machine Learning (ICML)*. PMLR, pp. 9229–9248.
- Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich (2015). “Going Deeper With Convolutions”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9.
- Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna (2016). “Rethinking the Inception Architecture for Computer Vision”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826.
- Taigman, Yaniv, Ming Yang, Marc’Aurelio Ranzato, and Lior Wolf (2014). “DeepFace: Closing the Gap to Human-Level Performance in Face Verification”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1701–1708.
- Tajbakhsh, Nima, Laura Jeyaseelan, Qian Li, Jeffrey N. Chiang, Zhihao Wu, and Xiaowei Ding (2020). “Embracing Imperfect Datasets: A Review of Deep Learning Solutions for Medical Image Segmentation”. In: *Medical Image Analysis*, p. 101693.

- Tang, Meng, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers (2018). “Normalized Cut Loss for Weakly-Supervised CNN Segmentation”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1818–1827.
- Tang, Meng, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov (2018). “On Regularized Losses for Weakly-supervised CNN Segmentation”. In: *European Conference on Computer Vision (ICCV)*, pp. 507–522.
- Tanimoto, Taffee T (1958). *Elementary mathematical theory of classification and prediction*. International Business Machines Corporation (IBM).
- Toldo, Marco, Andrea Maracani, Umberto Michieli, and Pietro Zanuttigh (2020). “Unsupervised Domain Adaptation in Semantic Segmentation: a Review”. In: *arXiv preprint arXiv:2005.10876*.
- Vahdat, Arash and Jan Kautz (2020). “NVAE: A Deep Hierarchical Variational Autoencoder”. In: *arXiv:2007.03898*.
- Valvano, Gabriele, Agisilaos Chartsias, Andrea Leo, and Sotirios A. Tsaftaris (2019). “Temporal Consistency Objectives Regularize the Learning Of Disentangled Representations”. In: *Domain Adaptation and Representation Transfer (DART)*. Springer, pp. 11–19. ISBN: 978-3-030-33391-1.
- Valvano, Gabriele, Daniele Della Latta, Nicola Martini, Gianmarco Santini, Andrea Gori, Chiara Iacconi, Andrea Ripoli, Luigi Landini, and Dante Chiappino (2017). “Evaluation of a Deep Convolutional Neural Network Method for the Segmentation of Breast Microcalcifications in Mammography Imaging”. In: *European Medical and Biological Engineering Conference & Nordic-Baltic Conference on Biomedical Engineering and Medical Physics (EMBECE & NBC)*. Springer, pp. 438–441. DOI: [10.1007/978-981-10-5122-7_110](https://doi.org/10.1007/978-981-10-5122-7_110).
- Valvano, Gabriele, Andrea Leo, and Sotirios A Tsaftaris (2021a). “Self-supervised Multi-scale Consistency for Weakly Supervised Segmentation Learning”. In: *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Springer, pp. 14–24.
- (2021b). “Stop Throwing Away Discriminators! Re-using Adversaries for Test-Time Training”. In: *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Springer, pp. 68–78.
- (2021c). “Learning to Segment From Scribbles Using Multi-Scale Adversarial Attention Gates”. In: *IEEE Transactions on Medical Imaging*. DOI: [10.1109/TMI.2021.3069634](https://doi.org/10.1109/TMI.2021.3069634).

- Valvano, Gabriele, Andrea Leo, and Sotirios A. Tsaftaris (2021d). “Re-using Adversarial Mask Discriminators for Test-time Training under Distribution Shifts”. In: *arXiv preprint arXiv:2108.11926*.
- (2021e). “Regularising Disentangled Representations With Anatomical Temporal Consistency”. In: *Under Review at: Biomedical Image Synthesis and Simulations, Elsevier*.
- Valvano, Gabriele, Nicola Martini, Andrea Leo, Gianmarco Santini, Daniele Della Latta, Emiliano Ricciardi, Dante Chiappino, and Pietro Pietrini (2019). “Evaluation of Planar and Volumetric Convolutional Neural Networks for Brain Segmentation”. In: *Organization for Human Brain Mapping (OHBM)*.
- Valvano, Gabriele, Gianmarco Santini, Nicola Martini, Andrea Ripoli, Chiara Iacconi, Dante Chiappino, and Daniele Della Latta (2019). “Convolutional Neural Networks for the Segmentation of Microcalcification in Mammography Imaging”. In: *Journal of Healthcare Engineering* 2019. DOI: [10.1155/2019/9360941](https://doi.org/10.1155/2019/9360941)
- Van Steenkiste, Sjoerd, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem (2019). “Are Disentangled Representations Helpful for Abstract Visual Reasoning?” In: *arXiv preprint arXiv:1905.12506*.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). “Attention Is All You Need”. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008.
- Vincent, Pascal, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol (2008). “Extracting and Composing Robust Features With Denoising Autoencoders”. In: *International Conference on Machine Learning (ICML)*, pp. 1096–1103.
- Vondrick, Carl, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy (2018). “Tracking Emerges By Colorizing Videos”. In: *European Conference on Computer Vision (ECCV)*, pp. 391–408.
- Wang, Dequan, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, Trevor Darrell, UC Berkeley, and Adobe Research (2021). “Tent: Fully Test-Time Adaptation by Entropy Minimization”. In: *International Conference on Learning Representations (ICLR)*. Vol. 4, p. 6.
- Wang, Haohan, Xindi Wu, Pengcheng Yin, and Eric P. Xing (2020). “High Frequency Component Helps Explain the Generalization of Convolutional Neural Networks”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*.

- Wang, Xiaolong, Allan Jabri, and Alexei A. Efros (2019). “Learning Correspondence From the Cycle-Consistency of Time”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2566–2576.
- Wang, Yaqing, Quanming Yao, James T. Kwok, and Lionel M. Ni (2020). “Generalizing From a Few Examples: A Survey on Few-Shot Learning”. In: *ACM Computing Surveys (CSUR)* 53.3, pp. 1–34.
- Wang, Yi, Zijun Deng, Xiaowei Hu, Lei Zhu, Xin Yang, Xuemiao Xu, Pheng-Ann Heng, and Dong Ni (2018). “Deep Attentional Features for Prostate Segmentation in Ultrasound”. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, pp. 523–530.
- Watanabe, Satoshi (1960). “Information Theoretical Analysis of Multivariate Correlation”. In: *IBM Journal of research and development* 4.1, pp. 66–82.
- Wikipedia (2020a). *Heart* — *Wikipedia, The Free Encyclopedia*. [Online; accessed 11-November-2020]. URL: <https://en.wikipedia.org/wiki/Heart>
- (2020b). *Wiggers diagram* — *Wikipedia, The Free Encyclopedia*. [Online; accessed 11-November-2020]. URL: https://en.wikipedia.org/wiki/Wiggers_diagram
- Wood, Justin N. (2016). “A Smoothness Constraint on the Development of Object Recognition”. In: *Cognition* 153, pp. 140–145.
- Wu, Chenshen, Luis Herranz, Xialei Liu, Joost van de Weijer, Bogdan Raducanu, et al. (2018). “Memory replay GANs: Learning to generate new categories without forgetting”. In: *Advances in Neural Information Processing Systems (NeurIPS)* 31, pp. 5962–5972.
- Xia, Tian, Agisilaos Chatsias, and Sotirios A. Tsaftaris (2020). “Pseudo-Healthy Synthesis With Pathology Disentanglement and Adversarial Learning”. In: *Medical Image Analysis* 64, p. 101719.
- Xiangli, Yuanbo, Yubin Deng, Bo Dai, Chen Change Loy, and Dahua Lin (2020). “Real or Not Real, That Is the Question”. In: *International Conference on Learning Representations (ICLR)*.
- Xie, Yutong, Jianpeng Zhang, Zehui Liao, Yong Xia, and Chunhua Shen (2020). “PGL: Prior-Guided Local Self-Supervised Learning for 3D Medical Image Segmentation”. In: *arXiv preprint arXiv: 2011.12640*.
- Xu, Yan (2019). “Deep Learning in Multimodal Medical Image Analysis”. In: *International Conference on Health Information Science (HIS)*. Springer, pp. 193–200.
- Xue, Yuan, Tao Xu, Han Zhang, L. Rodney Long, and Xiaolei Huang (2018). “SeGAN: Adversarial Network With Multi-Scale l1 Loss for

- Medical Image Segmentation". In: *Neuroinformatics* 16.3-4, pp. 383–392.
- Yang, Junlin, Nicha C. Dvornek, Fan Zhang, Julius Chapiro, MingDe Lin, and James S. Duncan (2019). "Unsupervised Domain Adaptation via Disentangled Representations: Application to Cross-Modality Liver Segmentation". In: *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, pp. 255–263.
- Yi, Xin, Ekta Walia, and Paul Babyn (2019). "Generative Adversarial Network in Medical Imaging: A Review". In: *Medical Image Analysis* 58, p. 101552.
- Yue, Qian, Xinzhe Luo, Qing Ye, Lingchao Xu, and Xiahai Zhuang (2019). "Cardiac Segmentation From LGE MRI Using Deep Neural Network Incorporating Shape and Spatial Priors". In: *arXiv preprint arXiv:1906.07347*.
- Yushkevich, Paul A., Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C. Gee, and Guido Gerig (2006). "User-Guided 3D Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability". In: *NeuroImage* 31.3, pp. 1116–1128.
- Zaech, Jan-Nico, Dengxin Dai, Martin Hahner, and Luc Van Gool (2019). "Texture Underfitting For Domain Adaptation". In: *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, pp. 547–552.
- Zamir, Amir R, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese (2018). "Taskonomy: Disentangling Task Transfer Learning". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3712–3722.
- Zenati, Houssam, Manon Romain, Chuan-Sheng Foo, Bruno Lecouat, and Vijay Chandrasekhar (2018). "Adversarially Learned Anomaly Detection". In: *IEEE International Conference on Data Mining (ICDM)*. IEEE, pp. 727–736.
- Zhang, Han, Zizhao Zhang, Augustus Odena, and Honglak Lee (2020). "Consistency Regularization for Generative Adversarial Networks". In: *International Conference on Learning Representations (ICLR)*.
- Zhang, Pengyi, Yunxin Zhong, and Xiaoqiong Li (2020). "ACCL: Adversarial Constrained-CNN Loss for Weakly Supervised Medical Image Segmentation". In: *arXiv:2005.00328*.
- Zhang, Yexun, Ya Zhang, Qinwei Xu, and Ruipeng Zhang (2020). "Learning Robust Shape-Based Features for Domain Generalization". In: *IEEE Access* 8, pp. 63748–63756.

- Zhao, Amy, Guha Balakrishnan, Fredo Durand, John V. Guttag, and Adrian V. Dalca (2019). "Data Augmentation Using Learned Transformations for One-Shot Medical Image Segmentation". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8543–8553.
- Zhao, Jing, Xijiong Xie, Xin Xu, and Shiliang Sun (2017). "Multi-View Learning Overview: Recent Progress and New Challenges". In: *Information Fusion* 38, pp. 43–54.
- Zhao, Junbo, Michael Mathieu, and Yann LeCun (2017). "Energy-based Generative Adversarial Network". In: *International Conference on Learning Representations (ICLR)*.
- Zhao, Shengyu, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han (2020). "Differentiable Augmentation for Data-efficient GAN Training". In: *Advances in Neural Information Processing Systems (NeurIPS)* 33.
- Zheng, Shuai, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr (2015). "Conditional Random Fields as Recurrent Neural Networks". In: *International Conference on Computer Vision (ICCV)*, pp. 1529–1537.
- Zhou, Kaiyang, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy (2021). "Domain Generalization: A Survey". In: *arXiv preprint arXiv:2103.02503*.
- Zhou, S. Kevin, Hayit Greenspan, Christos Davatzikos, James S. Duncan, Bram van Ginneken, Anant Madabhushi, Jerry L. Prince, Daniel Rueckert, and Ronald M. Summers (2021). "A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises". In: *Proceedings of the IEEE*.
- Zhou, Tongxue, Su Ruan, and Stéphane Canu (2019). "A Review: Deep Learning For Medical Image Segmentation Using Multi-Modality Fusion". In: *Array* 3, p. 100004.
- Zhou, Yuyin, Zhe Li, Song Bai, Chong Wang, Xinlei Chen, Mei Han, Elliot Fishman, and Alan L Yuille (2019). "Prior-Aware Neural Network for Partially-Supervised Multi-Organ Segmentation". In: *International Conference on Computer Vision (ICCV)*, pp. 10672–10681.
- Zhou, Zhi-Hua (2018). "A Brief Introduction to Weakly Supervised Learning". In: *National Science Review* 5.1, pp. 44–53.
- Zhou, Zongwei, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B. Gotway, and Jianming Liang (2019). "Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis". In: *International Conference on Medical Image*

Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 384–393.

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A. Efros (2017). “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks”. In: *International Conference on Computer Vision (ICCV)*, pp. 2223–2232.



Unless otherwise expressly stated, all original material of whatever nature created by Gabriele Valvano and included in this thesis, is licensed under a [Creative Commons Attribution Noncommercial Share Alike 3.0 Italy License](https://creativecommons.org/licenses/by-nc-sa/3.0/it/legalcode/).

Check on Creative Commons site:

<https://creativecommons.org/licenses/by-nc-sa/3.0/it/legalcode/>

<https://creativecommons.org/licenses/by-nc-sa/3.0/it/deed.en>

[Ask the author](#) about other uses.