This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3069634, IEEE Transactions on Medical Imaging

Learning to Segment from Scribbles using Multi-scale Adversarial Attention Gates

Gabriele Valvano, Andrea Leo, Sotirios A. Tsaftaris

***Abstract*—Large, fine-grained image segmentation datasets, annotated at pixel-level, are difficult to obtain, particularly in medical imaging, where annotations also require expert knowledge. Weakly-supervised learning can train models by relying on weaker forms of annotation, such as scribbles. Here, we learn to segment using scribble annotations in an adversarial game. With unpaired segmentation masks, we train a multi-scale GAN to generate realistic segmentation masks at multiple resolutions, while we use scribbles to learn their correct position in the image. Central to the model's success is a novel attention gating mechanism, which we condition with adversarial signals to act as a shape prior, resulting in better object localization 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 layers of the segmentor. We evaluated our model on several medical (ACDC, LVSC, CHAOS) and non-medical (PPSS) datasets, and 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). We release expert-made scribble annotations for the ACDC dataset, and the code used for the experiments, at https://vios-s.github.io/multiscale-adversarial-attention-gates.**

***Index Terms*—Weak Supervision, Scribbles, Segmentation, GAN, Attention, Shape Priors.**

I. 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 [1], [2]. However, the information inside unlabeled data can improve CNNs only under specific assumptions [1], and SSL requires representative image-segmentation pairs being available.

Alternatively, weakly-supervised approaches [3], [4], [5], [6] attempt to train models relying only on weak annotations (e.g., image-level labels, sparse pixel annotations, or noisy annotations [7]), 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 [8]. 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 [5]. 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.

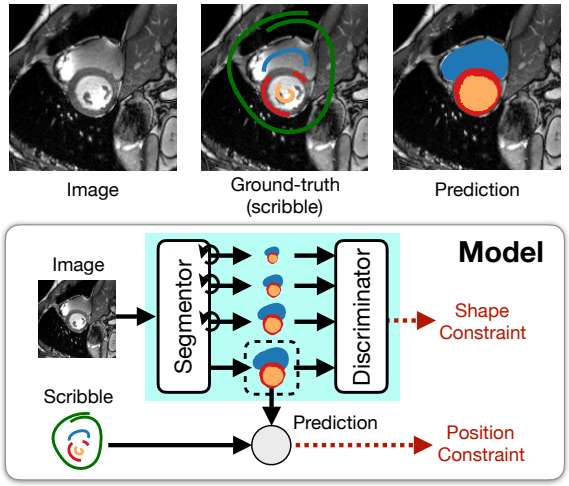

Fig. 1: In an adversarial game, our model learns to generate segmentation masks that look realistic at multiple scales and overlap with the available scribble annotations. Loopy arrows in the figure, on the segmentor, represent the proposed attention gates, which under adversarial conditioning suppress irrelevant information in the extracted features maps.

A. 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 [9] 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.

G. Valvano (email: gabriele.valvano@imtlucca.it) and A. Leo are with IMT School for Advanced Studies Lucca, Lucca 55100 LU, Italy. G. Valvano is also with School of Engineering, University of Edinburgh, Edinburgh EH9 3FB, UK. S. A. Tsaftaris is with School of Engineering, University of Edinburgh, Edinburgh EH9 3FB, UK.

This work was supported by the Erasmus+ programme of the European Union. S.A. Tsaftaris acknowledges the support of the Royal Academy of Engineering and the Research Chairs and Senior Research Fellowships scheme. We thank NVIDIA for donating the GPU used for this research.

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 [10], [11], [12], 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 [13] ours are conditioned by the adversarial signals, which enforce a stronger object localization in the image. Moreover, differently from other multi-scale GANs [10], [11], [12] 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 source[1] 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 summarize the contributions of the paper 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 [16], LVSC [17] and CHAOS [18]) and computer vision data (PPSS [19]).
- 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 few-shot learning method trained with densely annotated segmentation masks, and show the advantage of collecting large-scale weakly annotated datasets.

II. Related Work

A large body of research aimed at developing learning algorithms that rely less on high-quality annotations [2], [7]. 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.

A. 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 [20]. For this reason, many approaches have tried to expand scribble annotations by assigning the same class to pixels with similar intensity and nearby position [20], [21]. 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.* [5], who suggest estimating the class of unlabelled pixels via a learned two-step procedure. At first, they train a CNN directly with scribbles; then, they relabel the training set by refining the CNN predictions with Conditional Random Fields (CRF); finally, 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 [5], [22] have investigated alternatives that avoid this step, post-processing the model predictions with CRF [23] or introducing CRF as a trainable layer [24]. Tang *et al.* [22] 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 *et al.* [22], does not rely on CRF, and can handle both long-range and short-range inconsistencies in the predicted masks.

Concurrent to our work, Zhang *et al.* [25] recently introduced a method that learns to segment images from scribbles using an adversarial shape prior. However, they suggest using a PatchGAN [26] discriminator, which only focuses on *local* properties of the generated segmentations, while we introduce a method that focuses on both *local* and *global* aspects.

B. 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 [27]. Below, we summarise recent work on shape priors in Deep Learning.

Recently, Clough *et al.* used Persistent Homology to enforce shape priors in medical image segmentation [28]. Oktay *et al.* [29] demonstrated that we can learn a data-driven shape prior with a convolutional autoencoder trained on unpaired

[1]We simulate a realistic clinical setting, where the unpaired masks can be obtained from a different modality or acquisition protocol [14], [15].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3069634, IEEE Transactions on Medical Imaging

segmentation masks, and it can be used as regulariser to train a segmentor. Dalca *et al.* [30] suggested learning the shape prior with a variational autoencoder (VAE) [31], and then share part of the VAE weights with a segmentor. Other approaches included shape priors as post-processing, regularising the training [32], or adjusting predictions at inference, using VAEs [33] or denoising autoencoders [14]. Kervadec *et al.* [34] suggested introducing size information as a differentiable penalty, during training. Alternatively, Dalca *et al.* [35] proposed to learn to warp a segmentation atlas. Other methods [36], [37] 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 [38], [39], as well as using temporal consistency constraints on the object shapes dynamics [40].

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.

C. 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.* [41] 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 *et al.* [12] 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 [12], 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.

D. Attention Gates

Due to the ability to suppress irrelevant and ambiguous information, attention gates have become an integral part of many sequence modeling [42] and image classification [43] frameworks. Recently, they have also been successfully employed for segmentation [13], [44], [45], [46], [47], 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.

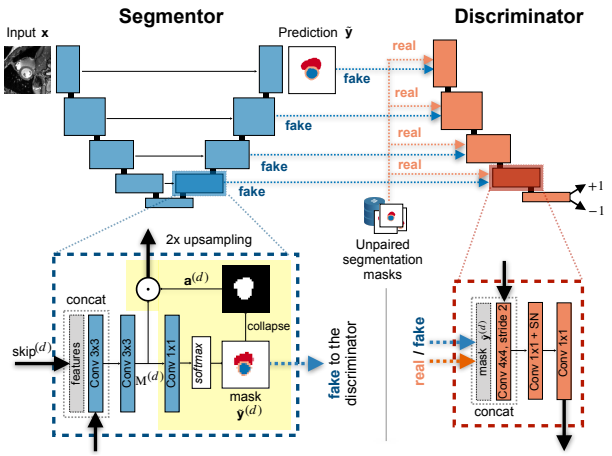

Fig. 2: Model architectures. Top: segmentor and discriminator interact at multiple scales. Bottom: convolutional blocks detail. In yellow background, the Adversarial Attention Gate (AAG).

III. Proposed Approach

In this section, we first describe the adopted notation, and then we present a general overview of the proposed method. Finally, we detail model architectures and training objectives.

Notation: For the remainder, we will use italic lowercase letters to denote scalars s. Two-dimensional images (matrices) will be denoted with bold lowercase letters, as $\mathbf{x} \in \mathbb{R}^{n \times m}$, where $n, m \in \mathbb{N}$ are scalars denoting dimensions. Tensors $\mathrm{T} \in \mathbb{R}^{r \times s \times t}$ are denoted as uppercase letters, where $r, s, t \in \mathbb{N}$. Finally, capital Greek letters will denote functions $\Phi(\cdot)$.

We will assume a weakly supervised setting, where we have access to: i) image-scribble pairs $(\mathbf{x}, \mathbf{y_s})$, being $\mathbf{x}$ the image and $\mathbf{y_s}$ the associated scribble; ii) unlabelled images; and iii) a set of segmentation masks $\mathbf{y}$ unrelated to any of the images.[2]

A. 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 Fig. 2).

[2]In Section V-F, we will also investigate a mixed setting, where we additionally have: iv) pairs of image-segmentation masks $(\mathbf{x}, \mathbf{y})$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3069634, IEEE Transactions on Medical Imaging

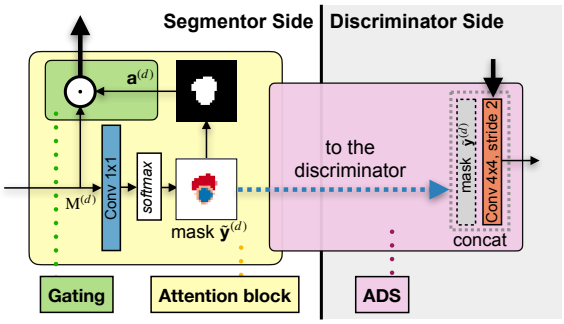

Fig. 3: Adversarial Attention Gates consist of an attention block (yellow background) pairing Adversarial Deep Supervision (ADS, obtained via the connection in pink background) and a multiplicative gating operation (green background).

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 $\tilde{\mathbf{y}} = \Sigma(\mathbf{x})$ 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., $\Delta(\mathbf{y})$ vs $\Delta(\tilde{\mathbf{y}})$) [9], 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 [37], [48]. 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.

B. Architectures

Segmentor $\Sigma(\cdot)$*:* We modify a UNet [49] to include AAG modules in the decoder and to allow collaborative training between segmentor and discriminator at multiple scales (Fig. 2). 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 [49]. 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 $\mathrm{M}^{(d)}$. Then, the AAG classifier uses $\mathrm{M}^{(d)}$ to predict a segmentation $\tilde{\mathbf{y}}^{(d)}$ at the given resolution level d. As a classifier, we use a convolutional layer with c $1 \times 1 \times k$ 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 $\mathbf{a}^{(d)}$, localizing object positions in the image (Fig. 2). To force the segmentor to use $\mathbf{a}^{(d)}$, we multiply the extracted features $\mathrm{M}^{(d)}$ with $\mathbf{a}^{(d)}$ 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 [50] 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. 3).

The *real samples* $\{\mathbf{y}^{(d)}\}_{d=1}^{4}$ consist of expert-made segmentations, that we supply at full or downsampled resolution at multiple discriminator depths, while *fake samples* $\{\tilde{\mathbf{y}}^{(d)}\}_{d=1}^{4}$ 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. 2, right).

The discriminator is a convolutional encoder adapted from [38]. At every depth d, at first, we process and downsample 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 [51] to improve training. Obtained feature maps are then compressed with a second convolutional layer using 12 $1 \times 1 \times k$ 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* [52] and *instance noise* [53]. 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 ($\Delta(\mathbf{y})$, $\Delta(\tilde{\mathbf{y}})$).

C. 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, only the latter when we are using 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 trough 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:

$$\mathcal{L}_{SUP} = \mathbb{1}(\mathbf{y_s}) * \Big[- \sum\nolimits_{i=1}^{c} w_i \cdot \mathbf{y_{s}}_i \log(\tilde{\mathbf{y}}_i) \Big], \qquad (1)$$

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 [54] 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 [9] 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 [55], we adopt the Least Square GAN objective [55] 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:

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

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.

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 [56] 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 [57] 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 [40]). Similarly to previous work with weak annotations [20], [58], 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.

IV. Experimental Setup

A. Data

Below, we first describe the adopted datasets; then, we detail the procedure used to generate scribble annotations; finally, we define how we construct train, validation, and test set. We consider medical and vision datasets, for the segmentation of heart, abdominal organs, and human pose parts:

1) **ACDC** [16]. This dataset contains 2-dimensional cine-MR images obtained by 100 patients using various 1.5T and 3T MR scanners and different temporal resolutions. Manual segmentations are provided for the end-diastolic (ED) and end-systolic (ES) cardiac phases for right ventricle (RV), left ventricle (LV) and myocardium (MYO). We resample the data to $1.51mm^2$ and cropped or padded them to match a size of 224×224. We normalize the images of each patient by removing the median and dividing by the interquartile range computed per volume.
2) **LVSC** [17]. It contains gated SSFP cine images of 100 patients, obtained from a mix of 1.5T scanner types and imaging parameters. Manual segmentations are provided for the left ventricular myocardium (MYO) in all the cardiac phases. To compare with ACDC, we only consider segmentations for ES and ED phase instants. We resample images to the average resolution of $1.45mm^2$ and crop or pad them to a size of 224×224. We normalize the images of each patient by removing the median and dividing by the interquartile range computed on his MRI scan.
3) **CHAOS** [18]. It has abdominal MR images of 20 subjects, with segmentation masks of liver, kidneys, and spleen. We test our model on the T1 in-phase and T2 images. We resample images to a resolution of $1.89mm^2$ and crop to 192×192 pixels, after normalising in $[-1, 1]$.
4) **PPSS** [19]. To demonstrate the broad utility of our method, we use the (non-medical) Pedestrian Parsing in Surveillance Scenes data. PPSS contains RGB images of pedestrians with occlusions, derived from 171 surveillance videos, using different cameras and resolutions. Besides images, ground truth segmentations are given for seven parts of the pedestrians: hair, face, upper clothes, arms, legs, shoes, and background. Since provided segmentations have size 80×160, we resample all the images to the same spatial resolution. We also normalize images between 0 and 1, dividing them by their maximum value.

Scribble Generation: To obtain scribbles with these datasets we follow different processes. Examples of those scribbles are shown in Fig. 4. 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 simulated data from automatic procedures. Thus, in ACDC, we use ITK-SNAP [59] to manually draw scribbles for ES and ED phases within the available segmentation masks. We

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3069634, IEEE Transactions on Medical Imaging

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 [60]. 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 [20]).

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 $\mathrm{P} \equiv (x_\mathrm{P}, y_\mathrm{P})$ 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 [38], [52], 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 [38]); 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 [19], 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.

B. 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 I. In particular, we consider:

- **UNet$_{\mathbf{PCE}}$** and **UNet$_{\mathbf{WPCE}}$** [54]: The UNet [49] 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 [54], or the WPCE loss introduced in (1).
- **UNet$_{\mathbf{CRF}}$**: We also consider the previous UNet$_{\mathrm{WPCE}}$ whose prediction is further processed by CRF as RNN layer [23], [24], [61]. 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 [5]: $\sigma_\alpha = 160, \sigma_\beta = 3$ and $\sigma_\gamma = 10$. These parameters model the pairwise potentials of CRF as weighted Gaussians [24]. As in [5], we use 5 iterations for the RNN. For the other datasets, we set $\sigma_\gamma = 3$, as suggested in [24].
- **TS-UNet$_{\mathbf{CRF}}$**: We compare our model to the two-steps procedure in [5], 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$_{\mathrm{CRF}}$.

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** [14]: 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$_{\mathbf{D}}$**: 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** [25]: similar to UNet$_{\mathrm{D}}$, ACCL trains with scribbles using a PatchGAN discriminator [26].

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

- **UNet$^{\mathbf{UB}}$**: 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 (1)).

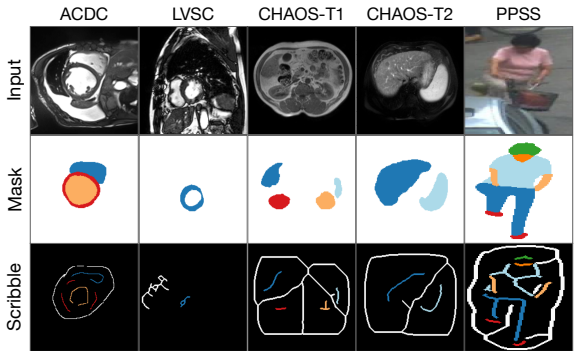

Fig. 4: Example of scribbles for each dataset (images resized to the same resolution to easy visualisation). ACDC: manual annotations; LVSC: random walks; CHAOS and PPSS: skeletonisation. Please, refer to Section IV-A for additional details.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3069634, IEEE Transactions on Medical Imaging

Model	Uses Prior	Type of Prior
UNet_{PCE}	✗	–
$\text{UNet}_{\text{WPCE}}$	✗	–
UNet_{CRF}	✓	Mean Field Assumption [24]
$\text{TS-UNet}_{\text{CRF}}$	✓	Mean Field Assumption [24]
PostDAE	✓	Shape, via DAE
UNet_{D}	✓	Shape, via Discriminator
ACCL	✓	Shape, via Patch Discriminator [26]
Ours	✓	Multi-scale Shape, via AAGs

TABLE I: Type of prior used by each model.

- $\textbf{UNet}_{\textbf{D}}^{\textbf{UB}}$: 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.* [5], we train the CRF as RNN layer of $\text{TS-UNet}_{\text{CRF}}$ 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.

Evaluation: We measure performance with the multi-class Dice score: $Dice = \frac{2|\tilde{\mathbf{y}} \cdot \mathbf{y}|}{|\tilde{\mathbf{y}}|+|\mathbf{y}|}$, where $\tilde{\mathbf{y}}$ and $\mathbf{y}$ 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.

V. 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 V-A, V-B); *and what happens when we have fewer scribble annotations, or less unpaired data* (Section V-C, V-D)? Then, we consider two natural questions that extend the applicability of our approach: *Can we learn from multiple scribbles per training image* (Section V-E)? *Can we mix per-pixel annotations with scribbles during training* (Section V-F)? Finally, we ask: *Why does Adversarial Attention Gating work* (Section V-G)?

A. 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 II, and qualitatively, in Fig. 5.

In particular, Table II reports average and standard deviation of the Dice score on test data for each dataset.[3] We clarify that, as discussed in Section IV-A, 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 Section I of the Supplementary Material.

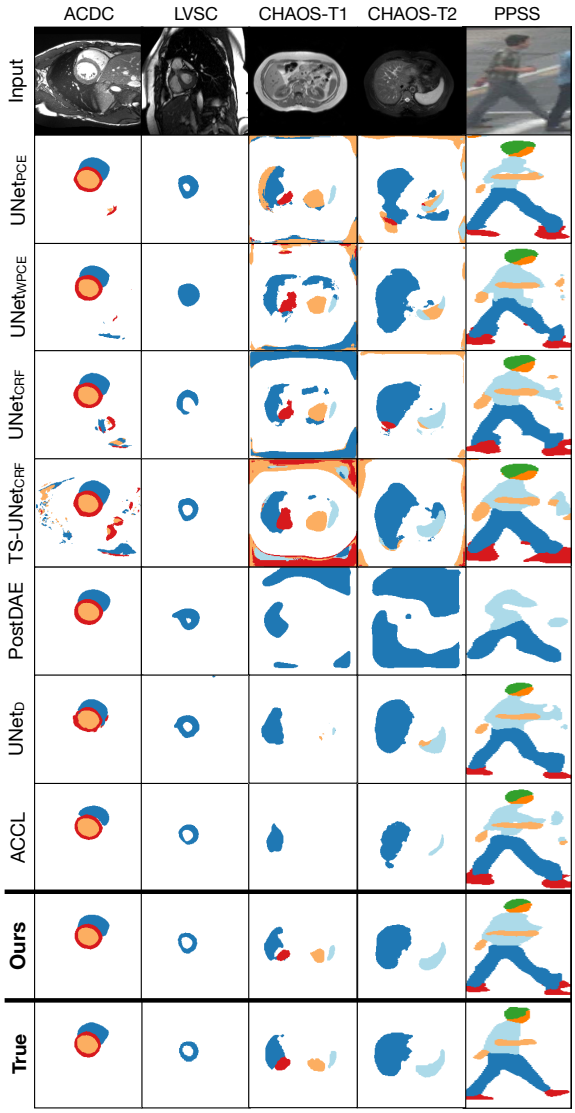

Fig. 5: Example of predicted segmentation masks for the considered methods on each task. Observe that our approach (bottom row) learns spatial relationships in the image, thus preventing the prediction of isolated pixels in the mask, as well as unrealistic spatial relationship among the object parts.

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 results.[4] When compared to the 2^{nd}

[3]For 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 Section I of the Supplementary Material.

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3069634, IEEE Transactions on Medical Imaging

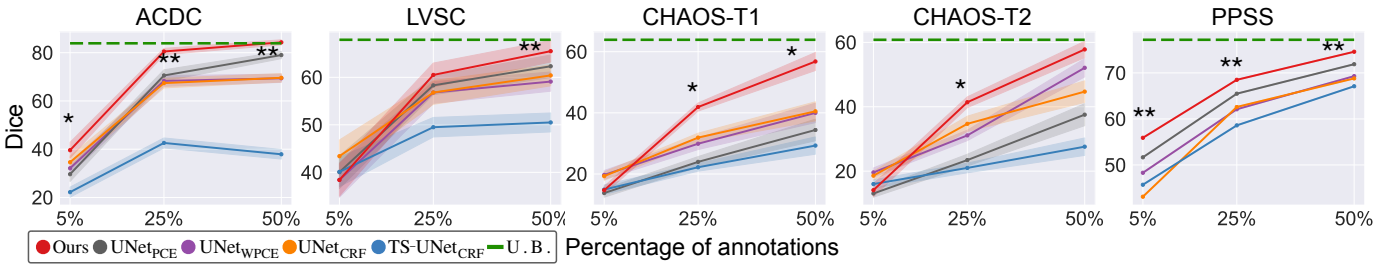

Fig. 6: Dice score obtained on the test data by our and methods that don't use shape priors when changing the percentage of available labels in the training set (shaded bands show standard errors instead of deviation for clarity). As upper bound (U.B.) we consider $\text{UNet}_\text{D}{}^{\text{UB}}$, trained using all the densely annotated masks. Asterisks (*,**) have the same role as in Table II.

Supervision Type	Model	Dataset: ACDC	LVSC	CHAOS-T1	CHAOS-T2	PPSS
Scribble	UNet_{PCE}	79.0_{06}	62.3_{09}	34.4_{06}	37.5_{06}	71.9_{04}
	$\text{UNet}_{\text{WPCE}}$	69.4_{07}	59.1_{07}	40.0_{05}	$\underline{52.1}_{05}$	69.3_{04}
	UNet_{CRF}	69.6_{07}	60.4_{08}	40.5_{05}	44.7_{06}	68.8_{04}
	$\text{TS-UNet}_{\text{CRF}}$	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	$\underline{82.6}_{05}$	$\mathbf{65.9_{08}}$	$\underline{48.3}_{07}$	49.7_{05}	$\underline{73.2}_{04}$
	Ours	$^{**}\mathbf{84.3_{04}}$	$\underline{65.5}_{08}$	$^{*}\mathbf{56.8_{05}}$	$\mathbf{57.8_{04}}$	$^{**}\mathbf{74.6_{04}}$
Mask	UNet^{UB}	82.0_{05}	67.2_{07}	60.8_{06}	58.6_{01}	72.8_{04}
	$\text{UNet}_{\text{D}}{}^{\text{UB}}$	83.9_{05}	67.9_{09}	63.9_{05}	60.8_{01}	77.2_{04}

TABLE II: 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 (* $p \leq 0.05$, ** $p \leq 0.01$).

best model, we obtain up to ∼8.5% of improvement on CHAOS-T1. As our ablation study shows in Section V-G, 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. 5).

Interestingly, we observe that weighting the loss contribution of each class based on their numerosity (UNet_{PCE} vs $\text{UNet}_{\text{WPCE}}$) is not always beneficial to the model, probably because, being sparse, scribble supervision 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 ($\text{UNet}_{\text{WPCE}}$ vs UNet_{CRF}).

The two-step paradigm of $\text{TS-UNet}_{\text{CRF}}$ is one of the worst. We observed that errors reinforce themselves in self-learning schemes [1], 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 II shows, post-processing the segmentor output with a DAE does not improve performance (PostDAE). As discussed by the PostDAE authors [14], 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 Section III of the Supplementary Material, for visual examples of these and other models' failures.

Instead, mask discriminators are an effective choice (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.

B. Segmentation Masks vs Scribbles

To understand the trade-off between time-to-segment and 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 [8], 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 [7], [63], [64], [65], [66], [67]. 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 [65].[6] TSDA uses a GAN

[5]In 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.* [5] [62]. With the enlarged scribbles, $\text{TS-UNet}_{\text{CRF}}$ 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 $\text{TS-UNet}_{\text{CRF}}$ 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.

[6]We used the code provided by the authors at https://github.com/krishnabits001/task_driven_data_augmentation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3069634, IEEE Transactions on Medical Imaging

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 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 II).

Our results confirm recent findings [68] 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.

C. Model Robustness to Limited Annotations

We analyze the robustness of the models with a scarcity of annotations in Fig. 6. 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 of $\text{UNet}_\text{D}{}^{\text{UB}}$, trained with all the available image-segmentation pairs.

As shown in Fig. 6, 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.

D. 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 V-C training our model without the additional unpaired images, and by varying the number of annotated data from 5% to 50%. At

[7]We 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 [69].

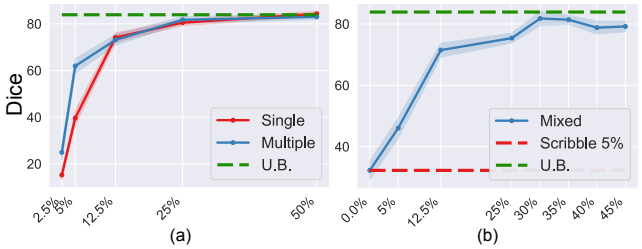

Fig. 7: (a) Effect of training with labels from multiple annotators; and (b) performance in presence of mixed supervision (mask and scribbles) on ACDC. The upper bound (U.B.) is the $\text{UNet}_\text{D}{}^{\text{UB}}$, trained with all the dense segmentation masks.

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.

E. 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 different levels of expertise in annotators [8], [70]. 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 IV-A for a description of ii) and iii)).

For every training image, we combine multiple scribbles summing up the supervised loss (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 [70]), but they are out of the scope of this manuscript.

In Fig. 7a, 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.

F. 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: (1) for scribble-annotated data, (2)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3069634, IEEE Transactions on Medical Imaging

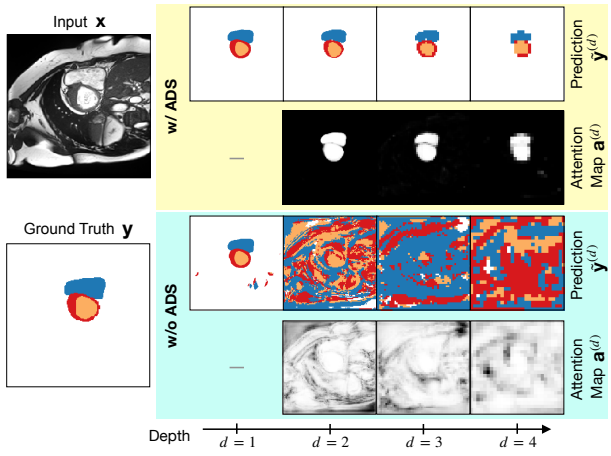

Fig. 8: UNet-like segmentor with (top) vs without (bottom) adversarial conditioning of the attention gates in its decoder. Conditioned by an adversarial shape prior (w/ ADS), the model learns semantic attention maps able to localize the object to segment at multiple scales. Also, the shape prior encourages the segmentor to learn multi-scale relationships in the objects.

for unlabeled data, and the weighted cross-entropy for fully annotated images. We report results on ACDC in Fig. 7b, 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 manuscript, 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 Section II of the Supplementary Material).

G. Why does Adversarial Attention Gating work?

Prior-conditioned Attention Maps are Object Localizers: Here we 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 III-B, 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 Fig. 3). In Fig. 8 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. 8, 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 UNet's prediction for $d = 1$ in Fig. 8) are prevented, and the model performance improves (see also Fig. 5).

Adversarial Attention Gating Trains Deep Layers Better: We qualitatively show that AAGs increase the training of the segmentor deepest layers. In Fig. 9, 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

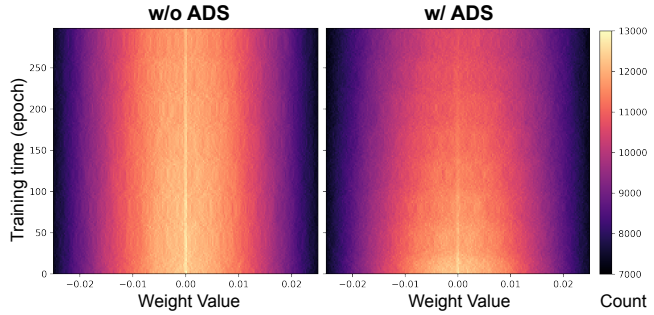

Fig. 9: Weight distribution for the convolutional layers at depth d=4 of the segmentor. We compare how the weight distribution changes during training, with and without the use of ADS on the segmentor. Notice that ADS helps the layer training, and the initially narrow distribution becomes broader in time.

	Attention	Discriminator single	Discriminator multi	5%	25%	50%
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 III: 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.

segmentor to update its weights also in deeper layers, which would otherwise suffer from vanishing gradients [71], [72].

Ablation Study: We show ablations on ACDC in Table III. 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 Fig. 3). When we completely remove the discriminator, the segmentor is trained just with scribble supervision and no adversarial signals. As Table III shows, each model component contributes to the final performance.

In particular, Table III 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 III, we further observe that both ADS and the multiplicative gating are important aspects of the model, and they increase the segmentation 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.

VI. 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 [73], conditional image generation [74] and localised style transfer [75].

Hoping to inspire new studies in weakly-supervised learning, we release manual scribble annotations for ACDC data, and the code used for the experiments.

References

[1] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning," *IEEE Trans Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[2] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *MIA*, vol. 54, pp. 280–296, 2019.

[3] Y. Zhou, Z. Li, S. Bai, C. Wang, X. Chen, M. Han, E. Fishman, and A. L. Yuille, "Prior-aware neural network for partially-supervised multi-organ segmentation," in *ICCV*, 2019, pp. 10 672–10 681.

[4] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *CVPR*, 2017, pp. 876–885.

[5] Y. B. Can, K. Chaitanya, B. Mustafa, L. M. Koch, E. Konukoglu, and C. F. Baumgartner, "Learning to segment medical images with scribble-supervision alone," in *DLMIA/ML-CDS*. Springer, 2018, pp. 236–244.

[6] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *ICCV*, 2017, pp. 5688–5696.

[7] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *MIA*, p. 101693, 2020.

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *ECCV*. Springer, 2014, pp. 740–755.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014, pp. 2672–2680.

[10] E. L. Denton, S. Chintala, R. Fergus *et al.*, "Deep generative image models using a laplacian pyramid of adversarial networks," in *NeurIPS*, 2015, pp. 1486–1494.

[11] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *ICLR*, 2017.

[12] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Macro-micro adversarial network for human parsing," in *ECCV*, 2018, pp. 418–434.

[13] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *MIA*, vol. 53, pp. 197–207, 2019.

[14] A. J. Larrazabal, C. Martínez, B. Glocker, and E. Ferrante, "Post-DAE: Anatomically plausible segmentation via post-processing with Denoising Autoencoders," *IEEE TMI*, 2020.

[15] N. Painchaud, Y. Skandarani, T. Judge, O. Bernard, A. Lalande, and P.-M. Jodoin, "Cardiac segmentation with strong anatomical guarantees," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3703–3713, 2020.

[16] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, Camara *et al.*, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE TMI*, vol. 37, no. 11, pp. 2514–2525, 2018.

[17] A. Suinesiaputra, B. R. Cowan, A. O. Al-Agamy, M. A. Elattar, N. Ayache, A. S. Fahmy, A. M. Khalifa, P. Medrano-Gracia, M.-P. Jolly, A. H. Kadish *et al.*, "A collaborative resource to build consensus for automated left ventricular segmentation of cardiac mr images," *MIA*, vol. 18, no. 1, pp. 50–62, 2014.

[18] A. Emre Kavur, N. Sinem Gezer, M. Barış, P.-H. Conze, V. Groza, D. Duy Pham, S. Chatterjee, P. Ernst, S. Özkan, B. Baydar *et al.*, "CHAOS Challenge–Combined (CT-MR) Healthy Abdominal Organ Segmentation," *arXiv*, pp. arXiv–2001, 2020.

[19] P. Luo, X. Wang, and X. Tang, "Pedestrian parsing via deep decompositional network," in *ICCV*, 2013, pp. 2648–2655.

[20] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *CVPR*, 2016, pp. 3159–3167.

[21] Z. Ji, Y. Shen, C. Ma, and M. Gao, "Scribble-based hierarchical weakly supervised learning for brain tumor segmentation," in *MICCAI*. Springer, 2019, pp. 175–183.

[22] M. Tang, F. Perazzi, A. Djelouah, I. Ben Ayed, C. Schroers, and Y. Boykov, "On regularized losses for weakly-supervised CNN segmentation," in *ECCV*, 2018, pp. 507–522.

[23] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE PAMI*, vol. 40, no. 4, pp. 834–848, 2017.

[24] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *ICCV*, 2015, pp. 1529–1537.

[25] P. Zhang, Y. Zhong, and X. Li, "ACCL: Adversarial constrained-CNN loss for weakly supervised medical image segmentation," *arXiv:2005.00328*, 2020.

[26] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.

[27] M. S. Nosrati and G. Hamarneh, "Incorporating prior knowledge in medical image segmentation: a survey," *arXiv:1607.01092*, 2016.

[28] J. R. Clough, I. Oksuz, N. Byrne, V. A. Zimmer, J. A. Schnabel, and A. P. King, "A Topological Loss Function for Deep-Learning based Image Segmentation using Persistent Homology," *arXiv:1910.01877*, 2019.

[29] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan *et al.*, "Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation," *IEEE TMI*, vol. 37, no. 2, pp. 384–395, 2017.

[30] A. V. Dalca, J. Guttag, and M. R. Sabuncu, "Anatomical priors in convolutional networks for unsupervised biomedical segmentation," in *CVPR*, 2018, pp. 9290–9299.

[31] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *ICLR*, 2014.

[32] Q. Yue, X. Luo, Q. Ye, L. Xu, and X. Zhuang, "Cardiac segmentation from LGE MRI using deep neural network incorporating shape and spatial priors," in *MICCAI*. Springer, 2019, pp. 559–567.

[33] N. Painchaud, Y. Skandarani, T. Judge, O. Bernard, A. Lalande, and P.-M. Jodoin, "Cardiac MRI segmentation with strong anatomical guarantees," in *MICCAI*. Springer, 2019, pp. 632–640.

[34] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, and I. B. Ayed, "Constrained-CNN losses for weakly supervised segmentation," *MIA*, vol. 54, pp. 88–99, 2019.

[35] A. V. Dalca, E. Yu, P. Golland, B. Fischl, M. R. Sabuncu, and J. E. Iglesias, "Unsupervised deep learning for Bayesian brain MRI segmentation," in *MICCAI*. Springer, 2019, pp. 356–365.

[36] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. A. Eslami, D. J. Rezende, and O. Ronneberger, "A probabilistic U-Net for segmentation of ambiguous images," in *NeurIPS*, 2018, pp. 6965–6975.

[37] C. F. Baumgartner, K. C. Tezcan, K. Chaitanya, A. M. Hötker, U. J. Muehlematter, K. Schawkat, A. S. Becker, O. Donati, and E. Konukoglu, "Phiseg: Capturing uncertainty in medical image segmentation," in *MICCAI*. Springer, 2019, pp. 119–127.

[38] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. E. Newby, R. Dharmakumar, and S. A. Tsaftaris, "Disentangled representation learning in cardiac image analysis," *Medical image analysis*, vol. 58, p. 101535, 2019.

[39] J. Yang, N. C. Dvornek, F. Zhang, J. Chapiro, M. Lin, and J. S. Duncan, "Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation," in *MICCAI*. Springer, 2019, pp. 255–263.

[40] G. Valvano, A. Chartsias, A. Leo, and S. A. Tsaftaris, "Temporal consistency objectives regularize the learning of disentangled representations," in *DART*. Springer, 2019, pp. 11–19.

[41] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang, "SegAN: Adversarial network with multi-scale l1 loss for medical image segmentation," *Neuroinformatics*, vol. 16, no. 3-4, pp. 383–392, 2018.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[43] S. Jetley, N. A. Lord, N. Lee, and P. H. S. Torr, "Learn To Pay Attention," *ICLR*, 2018.

[44] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention U-net: Learning where to look for the pancreas," *MIDL*, 2018.

[45] Y. Wang, Z. Deng, X. Hu, L. Zhu, X. Yang, X. Xu, P.-A. Heng, and D. Ni, "Deep attentional features for prostate segmentation in ultrasound," in *MICCAI*. Springer, 2018, pp. 523–530.

[46] A. Sinha and J. Dolz, "Multi-scale self-guided attention for medical image segmentation," *IEEE J Biomed Health Inform*, 2020.

[47] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *CVPR*, 2019, pp. 3146–3154.

[48] A. Vahdat and J. Kautz, "NVAE: A Deep Hierarchical Variational Autoencoder," *arXiv:2007.03898*, 2020.

[49] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[50] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *ICML*, 2015.

[51] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *ICLR*, 2018.

[52] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *NeurIPS*, 2016, pp. 2234–2242.

[53] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár, "Amortised map inference for image super-resolution," *ICLR*, 2017.

[54] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, and C. Schroers, "Normalized cut loss for weakly-supervised CNN segmentation," in *CVPR*, 2018, pp. 1818–1827.

[55] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "On the effectiveness of least squares generative adversarial networks," *IEEE PAMI*, 2018.

[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ICLR*, 2015.

[57] L. N. Smith, "Cyclical learning rates for training neural networks," in *WACV*. IEEE, 2017, pp. 464–472.

[58] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *ICCV*, 2015, pp. 1635–1643.

[59] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig, "User-Guided 3D Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[60] M. Rajchl, L. M. Koch, C. Ledig, J. Passerat-Palmbach, K. Misawa, K. Mori, and D. Rueckert, "Employing weak annotations for medical image analysis problems," *arXiv:1708.06297*, 2017.

[61] M. Monteiro, M. A. Figueiredo, and A. L. Oliveira, "Conditional random fields as recurrent neural networks for 3d medical imaging segmentation," *arXiv:1807.07464*, 2018.

[62] L. Grady, "Random walks for image segmentation," *IEEE PAMI*, vol. 28, no. 11, pp. 1768–1783, 2006.

[63] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," 2017.

[64] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transformations for one-shot medical image segmentation," in *CVPR*, 2019, pp. 8543–8553.

[65] K. Chaitanya, N. Karani, C. F. Baumgartner, A. Becker, O. Donati, and E. Konukoglu, "Semi-supervised and task-driven data augmentation," in *IPMI*. Springer, 2019, pp. 29–41.

[66] Y. Liu, J. Lee, M. Park, S. Kim, E. Yang, S. J. Hwang, and Y. Yang, "Learning to propagate labels: Transductive propagation network for few-shot learning," *ICLR*, 2019.

[67] A. R. Feyjie, R. Azad, M. Pedersoli, C. Kauffman, I. B. Ayed, and J. Dolz, "Semi-supervised few-shot learning for medical image segmentation," *arXiv preprint arXiv:2003.08462*, 2020.

[68] Y. M. Asano, C. Rupprecht, and A. Vedaldi, "A critical analysis of self-supervision, or what we can learn from a single image," *ICLR*, 2020.

[69] L.-Z. Guo, Y.-F. Li, M. Li, J.-F. Yi, B.-W. Zhou, and Z.-H. Zhou, "Reliable weakly supervised learning: Maximize gain and maintain safeness," *arXiv preprint arXiv:1904.09743*, 2019.

[70] S. Ørting, A. Doyle, M. H. A. van Hilten, O. Inel, C. R. Madan, P. Mavridis, H. Spiers, and V. Cheplygina, "A survey of crowdsourcing in medical image analysis," *arXiv:1902.09159*, 2019.

[71] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[72] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial Intellig and Stat*, 2015, pp. 562–570.

[73] J. Krebs, H. Delingette, B. Mailhé, N. Ayache, and T. Mansi, "Learning a probabilistic model for diffeomorphic registration," *IEEE TMI*, vol. 38, no. 9, pp. 2165–2176, 2019.

[74] S. Azadi, M. Tschannen, E. Tzeng, S. Gelly, T. Darrell, and M. Lucic, "Semantic Bottleneck Scene Generation," *arXiv:1911.11357*, 2019.

[75] L. Kurzman, D. Vazquez, and I. Laradji, "Class-based styling: Real-time localized style transfer with semantic segmentation," in *ICCV Workshops*, 2019, pp. 0–0.