CycleMix: A Holistic Strategy for Medical Image Segmentation from Scribble Supervision

Ke Zhang and Xiahai Zhuang*
Fudan University

Abstract

Curating a large set of fully annotated training data can be costly, especially for the tasks of medical image segmentation. Scribble, a weaker form of annotation, is more obtainable in practice, but training segmentation models from limited supervision of scribbles is still challenging. To address the difficulties, we propose a new framework for scribble learning-based medical image segmentation, which is composed of mix augmentation and cycle consistency and thus is referred to as CycleMix. *For augmentation of supervision, CycleMix adopts the mixup strategy with a dedicated design of random occlusion, to perform increments and decrements of scribbles. For regularization of supervision, CycleMix intensifies the training objective with consistency losses to penalize inconsistent segmentation, which results in significant improvement of segmentation performance. Results on two open datasets, i.e., ACDC and MSCMRseg, showed that the proposed method achieved exhilarating performance, demonstrating comparable or even better accuracy than the fully-supervised methods. The code and expert-made scribble annotations for MSCMRseg are publicly available at* https://github.com/BWGZK/CycleMix.

1. Introduction

Large fully-annotated datasets are crucial to the generalization ability of deep neural networks. However, the manual labeling of medical images requires great efforts from experienced clinical experts, which is both expensive and time-consuming. To alleviate it, existing works have exploited weakly labeled and unlabeled training data to assist model training, such as semi-supervised learning (SSL) [20, 23, 27] and weakly-supervised learning (WSL) [14, 22, 31]. However, SSL generally requires part of the images in the dataset to be accurately and precisely annotated. As an alternative, we propose to investigate a specific form of WSL approaches, which only utilize scribble annotations for model training.

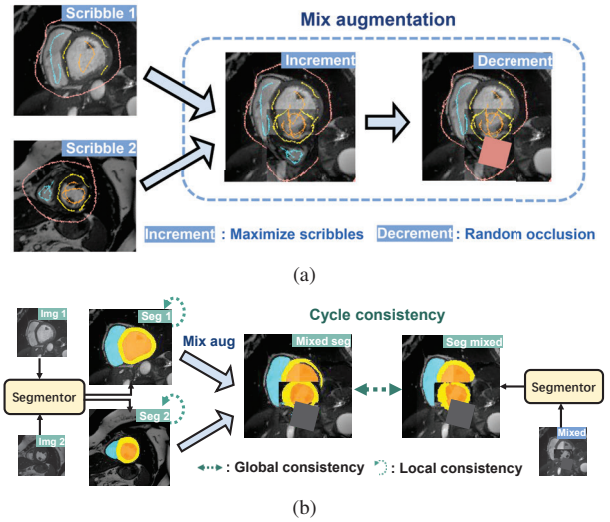

Figure 1. Illustration of CycleMix based on mix augmentation and cycle consistency on scribble training images: (a) shows the operations and results of mixing images and corresponding labels; (b) illustrates the segmentation results for consistency regularization.

WSL is proposed to exploit weak annotations, such as image-level labels, sparse annotations, and noisy annotations [24]. Among them, scribble, as images in Figure. 1 (a) illustrate, is one of the most convenient forms of weak label and has great potential in medical image segmentation [4]. However, due to the lack of supervision, it is still ardours to learn the shape priors of objects, which makes the segmentation of the boundaries particularly difficult.

The existing scribble learning mainly includes two groups. The first line of researches leverage *a priori* assumption to expand scribble annotation [24], such as labeling pixels with similar gray values and similar positions in the same category [13, 19]. However, the process of scribble expansion may generate noisy labels, which deteriorates the segmentation performance of trained models. The sec-

*Xiahai Zhuang is corresponding author. This work was funded by the National Natural Science Foundation of China (grant no. 61971142, 62111530195 and 62011540404) and the development fund for S talents (no. 2020015).

ond one learns adversarial shape priors, but requires extra fully-annotated masks [18, 28, 35].

There is a line of augmentation strategies, well known as *mixup*, have been proposed, which focus on generating previously-unseen virtual examples [8, 15, 16, 33, 34]. However, these strategies are proposed for image classification, and they may change the shape priors of target objects, leading to unrealistic segmentation results for a segmentation task. When only scribble supervision is available, the segmentation performance using mixup augmentation could become even worse and unstable, due to the lack of precise annotations.

To address above mentioned challenges, we propose *CycleMix* to learn segmentation from scribbles. As illustrated in Figure. 1, CycleMix maximizes supervision of scribbles based on mix augmentation and random occlusion, and regularizes training of models using consistency losses. Firstly, we surmise that a segmentation model should benefit from finer gradient flow via larger portion of annotated pixels. Therefore, we propose the two-step *mix augmentation* strategy to augment supervision, including image combination to increase scribbles and random occlusion to reduce scribbles. In addition, we develop two-level *consistency* regularization, at both of the global and local levels. The global consistency loss penalizes the inconsistent segmentation of the same image patch in two scenarios, *i.e.*, in the original image and mixed image; while the local consistency loss minimizes the distance between prediction and its largest connected component, exploiting the prior knowledge of anatomy that the target structures are interconnected.

The contributions of this paper are summarized as follows:

- We propose a novel weakly-supervised segmentation framework for scribble supervision, *i.e.*, CycleMix, by integrating mix augmentation of supervision and regularization of supervision from consistency, and introduce a new scribble annotated cardiac segmentation dataset of MSCMRseg.
- To the best of our knowledge, the proposed CycleMix is the first framework to incorporate mixup strategies for augmentation of weakly-supervised segmentation, where one can achieve both increments and decrements of scribbles from the mixed training images.
- We propose the consistency losses to regularize the limited supervision from scribbles by penalizing inconsistent segmentation results, at both the global and local levels, which can lead to profound improvement of model performance.
- CycleMix has been evaluated on two open datasets, *i.e.*, ACDC and MSCMR, and demonstrated promising performance by generating comparable or even better segmentation accuracy than the fully-supervised approaches.

2. Related works

2.1. Learning from scribble supervision

Scribble refers to sparse annotations where masks are provided for a small fraction of pixels in images [24]. Existing methods mostly used selective pixel loss for annotated pixels. There are works [1, 13, 19] attempting to expand scribbles or reconstruct the complete mask for model training. However, the pixel-relabeling process required iterative training, which is slow and prone to noisy labels. To avoid relabeling, several works utilized conditional random field to refine the segmentation results in post-processing [4,6] or as trainable layer [26, 36]. However, these methods could not provide better supervision for model training. Other works [28,35] included a new module to evaluate the quality of segmentation masks, which encourages the predictions to be realistic. For example, Gabriele *et al.* [28] proposed the multi-scale attention gates in adversarial training, Zhang *et al.* [35] used PatchGAN discriminator [12] to leverage shape priors. However, these methods required additional data source of complete masks.

2.2. Mixup augmentations

Data augmentation plays a vital role in preventing models from overfitting to the limited training data and enhancing the generalization ability of neural networks. Mixup augmentations refer to a line of strategies which combine two images and corresponding labels [8, 15, 16, 33, 34]. Compared with conventional augmentation methods, *i.e.*, rotation and flipping, mixup approaches can increase scribble annotations of augmented image through mix operation. Zhang *et al.* [34] introduced MixUp, which performed linear interpolation between two images and their labels. Manifold MixUp in [29] extended the mixup operation of input images to hidden features. Cutout in [8] randomly dropped out the square regions of images, and CutMix in [33] replaced the dropped areas with patches from other images. Puzzle Mix in [16] introduced a new mixup method based on saliency and local statistics. Co-mixup in [16] extended the mixup between two images to multiple images, and encouraged the supermodular diversity of mixed images.

In medical imaging, mixup augmentation has been applied to semi-supervised image segmentation [5] and object detection tasks [30]. Chaitanya *et al.* [5] concluded that mixup could lead to an impressive performance gain on semi-supervised segmentation. Although the mixed images might not look realistic, the mixed soft labels can provide more information to facilitate the training of models [5, 11].

2.3. Consistency regularization

Consistency strategies take advantage of the fact that if the same image is perturbed, the segmentation results should remain consistent. Consistency regularization

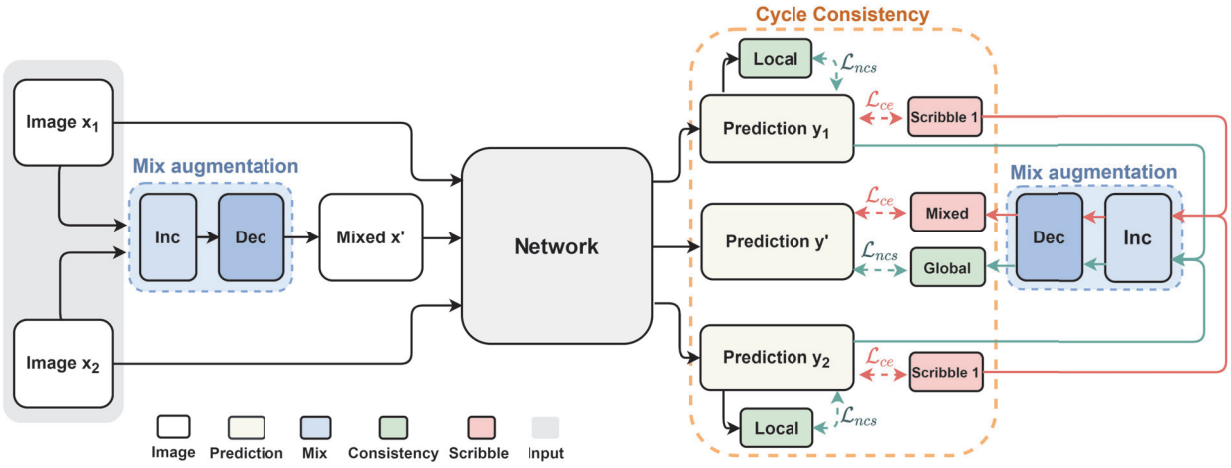

Figure 2. The pipeline of CycleMix, where two blue dashed boxes represent the same mix augmentation operation. The loss functions include the losses of scribble supervision (annotated with red) and the losses of consistency (annotated with green).

has been widely applied in image-translation and semi-supervised learning. CycleGAN [37] leveraged forward-backward consistency to enhance the ability of image-to-image translation. In semi-supervised setting, consistency is enforced over two augmentation versions of input images to obtain stable predictions of unlabeled images [17,21,27]. In this work, we propose to utilize consistency at both of the global and local levels to leverage the mix-invariant property and interconnected fact of segmentation structures.

3. Method

The proposed CycleMix is composed of two new strategies, *i.e.*, mix augmentation of scribble supervision and cycle consistency for regularization of supervision. The former is aimed to achieve the increments and decrements of scribbles by two-step mix-based image combination and random occlusion; the latter is designed to regularize the supervision in model training via two-level consistency penalty. Figure. 2 presents framework of neural network implementation of CycleMix.

3.1. Mix augmentation of scribble supervision

In this section, we extend the mixup strategy to the two-stage augmentation of scribble supervision. In the first stage, we increase the amount of scribbles by image combination, referred to increments of scribbles, which is to mixup two images to maximize the saliency. In the second stage, we perform an operation of random occlusion, by replacing certain area containing scribbles with background, which results in decrements of scribbles. Finally, the augmentation of supervision is achieved via a dedicated loss function from the generated mixup images.

3.1.1 Increments of scribbles

We surmise that increasing scribbles will benefit from finer gradient flow through larger proportions of annotated pixels. Furthermore, we observe that the scribble-annotated area generally has high saliency. Therefore, we propose to maximize the scribble annotation of mixed images to efficiently obtain the maximization of saliency of mixed training images. Here, we adopt the Puzzle Mix in [16] to utilize saliency and local statistic features. Note that the proposed method is applicable to other mixup strategies, such as MixUp [34], CutMix [33] and Co-mixup [15]. Readers could refer to the supplementary material for a comparison study.

We apply Puzzle Mix to both images and their corresponding scribble labels. Let two d-dimensional images with annotations be (x_1, y_1), (x_2, y_2). The mixed result transported from the two training data, denoted as (x^m_{12}, y^m_{12}), is computed by:

$$x^m_{12} = M(x_1, x_2) \text{ and } y^m_{12} = M(y_1, y_2), \quad (1)$$

$$M(a_1, a_2) = (1 - z) \odot \prod_1^T a_1 + z \odot \prod_2^T a_2, \quad (2)$$

where $M(a_1, a_2)$ is the mixup function on a_1 and a_2; $\prod_1$ and $\prod_2$ represent the transportation matrix of dimension $d \times d$; z denotes a mask in $[0, 1]$ of dimension d; $\odot$ refers to the element-wise multiplication. The parameter set, $\{\prod_1, \prod_2, z\}$, is aimed to maximize the saliency of mixed image, which is computed by,

$$\{\prod_1, \prod_2, z\} = \underset{\prod_1, \prod_2, z}{\arg\max} [(1-z) \odot \prod_1^T s(x_1) + z \odot \prod_2^T s(x_2)],$$

where $s(x)$ is the saliency of image x and is computed by taking the l_2 norm of the gradient value. For this optimization one could refer to [16] for more details.

3.1.2 Decrements of scribbles

To further augment scribble supervision, we propose to randomly occlude a region containing scribbles from the mixed images, to generate more training images. This strategy results in decrements of scribbles in the mixed image, and has been proved to be effective in enhancing performance of object localization [33].

Let (x^o, y^o) be the pair of new training data generated from (x^m, y^m). We apply a randomly rotated rectangular area to occlude the image and turns the occluded scribbles into background,

$$x^o_{12} = (1 - \mathbb{1}_O) \odot x^m_{12} \tag{3}$$

$$y^o_{12} = (1 - \mathbb{1}_O) \odot y^m_{12} \tag{4}$$

where $\mathbb{1}_O$ is a binary rectangular mask of dimension $n \times n$. In our experiment, we chose a rectangle with size of 32×32.

3.1.3 Scribble supervision

For scribble supervision, we apply the cross-entropy function *solely on the annotated pixels, ignoring the unlabeled pixels whose ground truth labels are unknown*. Hence, the loss $\mathcal{L}_{unmix}$ for unmixed samples (x_1, y_1) and (x_2, y_2) is formulated as:

$$\mathcal{L}_{unmix} = \frac{1}{2}\left[\mathcal{L}_{ce}(\hat{y}_1, y_1) + \mathcal{L}_{ce}(\hat{y}_2, y_2)\right], \tag{5}$$

where, $\hat{y} = S(x)$ is the predicted segmentation of x, and,

$$\mathcal{L}_{ce}(\hat{y}, y) = \sum_{i \in \Omega_L} \sum_{k \in K} -y[i,k] \log(\hat{y}[i,k]), \tag{6}$$

where, K is the index set of labels, $[i,k]$ indicate the k-element of label vector of the i-th pixel, $y[i,k]$ equals the probability of i-th pixel belongs to the k-th class, and Ω_L refers to the set of pixels with scribble annotation, to which $\mathcal{L}_{ce}$ loss is applied.

Furthermore, since the operation of Puzzle Mix is not symmetric, namely $M(x_1, x_2) \neq M(x_2, x_1)$, we use a symmetrical loss, referred to as mixed loss $\mathcal{L}_{mix}$, for the generated samples (x^o_{12}, y^o_{12}) and (x^o_{21}, y^o_{21}),

$$\mathcal{L}_{mix} = \frac{1}{2}\left[\mathcal{L}_{ce}(\hat{y}^o_{12}, y^o_{12}) + \mathcal{L}_{ce}(\hat{y}^o_{21}, y^o_{21})\right]. \tag{7}$$

The loss for augmented scribble supervision is given by,

$$\mathcal{L}_{sup} = \lambda_1 \mathcal{L}_{unmix} + \lambda_2 \mathcal{L}_{mix}, \tag{8}$$

where λ_1,λ_2 are the balancing parameters.

3.2. Regularization of supervision via cycle consistency

In this section, we introduce two regularization terms, *i.e.*, the global consistency loss and the local consistency loss.

3.2.1 Global consistency

The objective of global consistency is to leverage the mix-invariant property, which requires the same image patch to behave consistently in two scenarios, *i.e.*, the original image and the mixed image. Therefore, we propose the global consistency loss to penalize the inconsistent segmentation.

For images x_1, x_2 and their mixed image $x^m_{12} = M(x_1, x_2)$, the corresponding segmentation is represented as $\hat{y} = S(x)$, where $S(\cdot)$ is the segmentor. Assume the parameters of mixing function, *i.e.*, Π_1, Π_2, and z in in Eq. (2), remain unchanged, one should have,

$$M(S(x_1), S(x_2)) = S(M(x_1, x_2)). \tag{9}$$

This is the global consistency requiring the mixed segmentation of image x_1 and x_2 to be consistent with the segmentation of the mixed image x^m_{12} after the same mixing operation. Taking the random occlusion operation into account, we modify Eq. (9) as follows,

$$(1 - \mathbb{1}_O) \odot M\left(\hat{y}_1, \hat{y}_2\right) = S\left((1 - \mathbb{1}_O) \odot x^m_{12}\right). \tag{10}$$

We propose to use a symmetrical metric based on the negative cosine similarity between two segmentation results as the global consistency loss [7, 10],

$$\mathcal{L}_{con\text{-}g} = \frac{1}{2}\left[\mathcal{L}_{ncs}(p_{12}, q_{12}) + \mathcal{L}_{ncs}(p_{21}, q_{21})\right], \tag{11}$$

where, $p_{12} \triangleq (1 - \mathbb{1}_o) \odot M(\hat{y}_1, \hat{y}_2)$ and $q_{12} \triangleq S((1 - \mathbb{1}_o) \odot x^m_{12})$ are respectively the mixed segmentation and segmentation of mixed image, and likewise for p_{21} and q_{21}; $\mathcal{L}_{ncs}(\cdot, \cdot)$ is the negative cosine similarity and is defined as,

$$\mathcal{L}_{ncs}(p, q) = -\frac{p \cdot q}{||p||_2 \cdot ||q||_2}. \tag{12}$$

3.2.2 Local consistency

For a target object, the mixup operation often causes disconnected structure in the mixed image. This phenomenon makes it particularly difficult for a segmentation model to learn the shape priors of target objects.

Leveraging the fact that the target structure can be interconnected in many medical applications, we propose the local consistency to eliminate the discrete results. For unmixed images x_1 and x_2, the local consistency loss $\mathcal{L}_{con\text{-}l}$ is formulated as:

$$\mathcal{L}_{con\text{-}l} = \frac{1}{2}\left[\mathcal{L}_{ncs}\left(\hat{y}_1, C\left(\hat{y}_1\right)\right) + \mathcal{L}_{ncs}\left(\hat{y}_2, C\left(\hat{y}_2\right)\right)\right], \tag{13}$$

where, $C(\cdot)$ is a morphological operation on a segmentation result, which outputs the largest connected area of each non-background class in the input segmentation. The purpose of Eq. (13) is to minimize the distance between segmentation

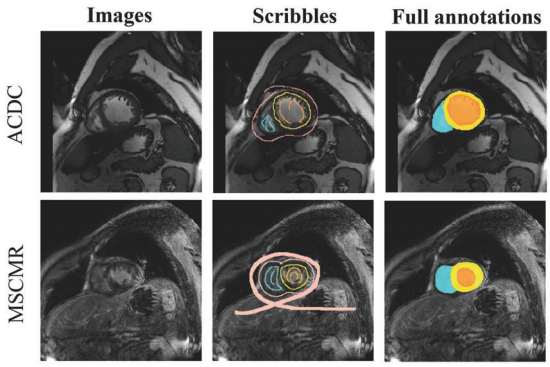

Figure 3. Examples from ACDC and MSCMRseg datasets. Since boundaries of structures in MSCMRseg images are general more difficult to distinguish, a thicker scribble is included to annotate the background for more supervision.

results and their largest connected areas. As formulated in Eq. (11), we use the symmetrical negative cosine similarity as the metric of distance.

Finally, the training objective $\mathcal{L}$ is formulated as:

$$\mathcal{L} = \underbrace{(\lambda_1 \mathcal{L}_{unmix} + \lambda_2 \mathcal{L}_{mix})}_{sup} + \underbrace{(\lambda_3 \mathcal{L}_{con\text{-}g} + \lambda_4 \mathcal{L}_{con\text{-}l})}_{unsup}, \tag{14}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are hyperparameters to leverage the relative importance of different loss components.

4. Experiments

4.1. Data and evaluation metric

CycleMix is evaluated on two open datasets, *i.e.*, ACDC and MSCMRseg, on which rich results have been reported in literature for comparisons. In addition, we use ACDC dataset for extensive parameter studies.

ACDC [3] dataset is composed of 2-dimensional cine-MRI images from 100 patients. The cine-MRI images were obtained using two MRI scanners of various magnetic strengths and different resolutions. For each patient, manual annotations of right ventricle (RV), left ventricle (LV) and myocardium (MYO) are provided for both the end-diastolic (ED) and end-systolic (ES) phase. Following [28], the 100 subjects in ACDC dataset is randomly divided into 3 sets of 70 (training), 15(validation), 15 (test) subjects for experiments. To compare with the previous state-of-the-art methods, which use unpaired masks to learn shape priors, we further divided the training set into two halves, 35 training images with scribble labels and 35 mask images with heart segmentation. *Unless specified, we only used 35 training images when training the proposed CycleMix and baselines*.

MSCMRseg [38, 39] contains late gadolinium enhancement (LGE) MRI images collected from 45 patient underwent cardiomyopathy, which represents more challenges for automatic segmentation than the unenhanced cardiac MRI. Gold standard segmentation of LV, MYO, RV of these images has also been released by the organizers. Following [32], we randomly divided the images from 45 patients into 3 sets, including 25 for training, 5 for validation and 20 for test.

Scribble annotations. For ACDC dataset, we used the released expert-made scribble annotations [28]. To obtain realistic scribble annotations, we further manually annotate the MSCMRseg dataset, following the principles in [28]. The average image coverages of scribbles for background, RV, MYO, LV are 3.4%, 27.7%, 31.3%, and 24.1%, respectively. Figure. 3 presents two exemplar images and their annotations from the two datasets. Please refer to supplementary material for more details of scribble annotations.

Evaluation. We adopted the Dice coefficient [9] to evaluate the performance of each method, which gauges the similarity of two segmentation masks.

4.2. Experimental setup

Implementation Details. We adopted the 2D variant of UNet [2], denoted as UNet$^+$, as the network architecture of CylceMix for all experiments, which was implemented using Pytorch. Since the provided images have different resolutions, We first resampled them and their annotations into a common in-plane resolution of 1.37×1.37 mm. Then, all images were cropped or padded to the same image size of 212×212 pixel. During training, we normalized the intensity of each image to zero mean and unit variance. The learning rate was fixed to 0.0001. We empirically set $\lambda_1 = \lambda_2 = \lambda_4 = 1$ and $\lambda_3 = 0.05$ in Eq.(14). All models were trained using one single NVIDIA 3090Ti 24GB GPU for 1000 epochs.

Baseline settings. The proposed CycleMix was trained with *scribble annotations*. Firstly, we compared it with baselines trained on scribble-annotated datasets. Recently, there are several works leveraged GAN networks to learn shape priors. We also compared with these challenging benchmarks which require *extra unpaired segmentation masks* to train GAN networks. Finally, we considered several *supervised methods* as upper bounds, which were trained on fully-annotated datasets.

- *Baselines*: We first compared to UNet$^+_{pce}$ trained with cross entropy loss of annotated pixels in [26]. Then, we applied different mix-up augmentation strategy to UNet$^+_{pce}$, *i.e.*, MixUp [34], CutMix [33], Puzzle Mix [16], Co-mixup [15]. Finally, we included the experiment results on ACDC dataset reported in [28] for reference, *i.e.,* UNet$_{pce}$ [25],UNet$_{wpce}$ [28], UNet$_{CRF}$ [36].

- *Challenging benchmarks*: The above baselines do

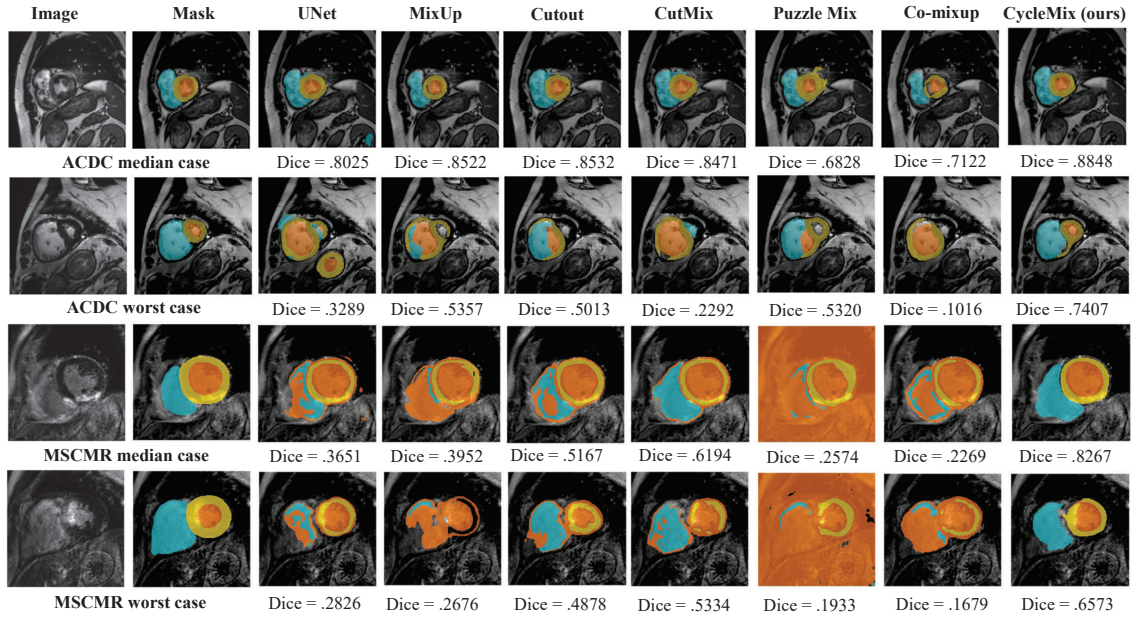

Figure 4. Quantitative comparison of the proposed method on ACDC and MSCMRseg datasets. The selected subjects were the median and the worst cases with regard to the Dice scores of the results of fully-supervised segmentation by $\text{UNet}_{\text{F}}^{+}$.

not leverage additional unpaired segmentation masks during training. For more challenging benchmarks, we compared with four works using extra unpaired data to learn shape priors, including PostDAE [18], UNet_D [28], ACCL [35], MAAG [28]. We refered to their segmentation results reported in [28] on ACDC dataset for comparison.

- *Supervised methods*: Finally, we performed the comparison in fully-supervised segmentation. Firstly, we applied UNet^{+} in [2] to the training data of full annotations using conventional cross entropy loss, referred to as UNet_F^{+}. Then, we applied Puzzle Mix augmentation strategy to UNet_F^{+}, and obtained the Puzzle Mix_F. Finally, we trained CycleMix with fully annotated data, denoted as CycleMix_F, and compared with UNet_F^{+} and Puzzle Mix_F on both ACDC and MSCMRseg datasets.

4.3. Comparison with different mix-up strategies

Table. 1 presents the performance of CycleMix on ACDC and MSCMRseg datasets. We compared with different data augmentation methods, *i.e.*, Mixup, Cutout, CutMix, Puzzle Mix, Comix-up as strong baselines. Here, we used 35 subjects for training, and the results using 70 training images are presented in supplementary material.

When only scribble annotations are available, Puzzle Mix achieved poor performance, with average Dice Scores of 62.4% on ACDC dataset and 24.1% on MSCM[illegible]. Therefore, it is more difficult for the segmentation

When with our proposed augmentation and regularization of supervision, CycleMix boosted the performance to reach Dice of 84.8% and 80.0% for the two datasets, respectively, demonstrating improvements of 22.4% and 55.9%.

Furthermore, the average Dice Score of CycleMix not only surpassed all weakly-supervised baselines by a large margin, but also exceeded the two fully-supervised methods. Particularly on the challenging task of MSCMRseg dataset, CycleMix achieved average Dice 0.800, with 14.6% increment than CutMix which ranks the second in the scribble supervision leader board. For the fully-supervised methods, one can observe CycleMix (marginally) outperformed both UNet_F and Puzzle Mix_F in Table. 1. Specifically, CycleMix with scribble supervision obtained an average improvement of 0.8% (84.8% vs 84.0%) and 1.1% (80.0% vs 78.9%) on MSCMRseg and ACDC dataset, respectively.

Figure. 4 visualizes results on the worst and median cases selected using the fully-supervised UNet. It is observed that Puzzle Mix could fail in the scribble supervision-based segmentation, especially on the challenging task of MSCMRseg. This may be due to its transportation strategy of image patches, which is more likely to change the shape of the target structure than other mix-up strategies based on linear interpolation or local replacement. Similar behavior could be seen from Co-mixup which adopts the similar transportation strategy to that of Puzzle

Table 1. The performance (Dice Scores) on ACDC and MSCMRseg dataset of CycleMix compared with different mixup strategies. **Bold** denotes the best performance, underline denotes the second best performance.

Methods	Data	ACDC				MSCMRseg			
		LV	MYO	RV	Avg	LV	MYO	RV	Avg
35 scribbles									
UNet$^{+}_{pce}$	scribbles	.785±.196	.725±.151	.746±.203	.752	.494±.082	.583±.067	.057±.022	.378
MixUp [34]	scribbles	.803±.178	.753±.116	.767±.226	.774	.610±.144	.463±.147	.378±.153	.484
Cutout [8]	scribbles	.832±.172	.754±.138	.812±.129	.800	.459±.077	.641±.136	.697±.149	.599
CutMix [33]	scribbles	.641±.359	.734±.144	.740±.216	.705	.578±.063	.622±.121	.761±.105	.654
Puzzle Mix [16]	scribbles	.663±.333	.650±.231	.559±.343	.624	.061±.021	.634±.084	.028±.012	.241
Co-mixup [15]	scribbles	.622±.304	.621±.214	.702±.211	.648	.356±.075	.343±.067	.053±.022	.251
CycleMix(ours)	scribbles	**.883±.095**	.798±.075	.863±.073	**.848**	**.870±.061**	.739±.049	**.791±.072**	**.800**
35 masks									
UNet$^{+}_{F}$	masks	.849±.152	.792±.140	.817±.151	.820	.857±.055	.720±.075	.689±.120	.755
Puzzle Mix$_F$ [16]	masks	.849±.182	**.807±.088**	**.865±.089**	.840	.867±.042	**.742±.043**	.759±.039	.789

model to learn the shape priors, especially in the case of a small training dataset. CycleMix overcomes this disadvantage by combining losses of mixed images and unmixed images, *i.e.*, $\mathcal{L}_{mix}$ and $\mathcal{L}_{unmix}$, and leveraging consistency regularization to preserve shape priors, which will be further explored in the ablation study.

4.4. Comparison with weakly-supervised methods

Table. 2 presents the results on the ACDC dataset. The previous best method, MAMG [28] exploited the unpaired masks from 35 additional subjects, and achieved 81.6% Dice Score with the assistance of multi-scale GAN. Without these masks, CycleMix still achieved a new state-of-the-art (SOTA) Dice of 84.8% average, with a promising margin over MAMG. For the RV structure with more shape variation, CycleMix obtained remarkable gains of 11.1% over MAMG (86.3% vs 75.2%). For the other methods, CycleMix demonstrated more significant performance improvements. We concluded that despite the additional masks, the models could learn very limited prior shapes through GAN when the number of training images is small. Thanks to the mix augmentation and consistency regularization for scribble supervision, CycleMix learned robust shape priors and set a new SOTA of segmentation.

Moreover, as one can observed from the upper part of Table. 2, CycleMix consistently outperformed all the other scribble supervision-based methods. Particularly, CycleMix obtained average performance gain up to 8.2% than UNet$_{pce}$ which ranks the 2nd.

4.5. Ablation study

This section studies the effectiveness of our proposed strategies, including the usage of unmix loss ($\mathcal{L}_{unmix}$), mixed loss ($\mathcal{L}_{mix}$), global consistency loss ($\mathcal{L}_{con\text{-}g}$), random occlusion ($\mathbb{1}_O$), and local consistency loss ($\mathcal{L}_{con\text{-}l}$). Table. 3 presents the details.

Table 2. The performance (Dice Scores) on ACDC dataset of CycleMix compared with state-of-the-art weakly-supervised methods. We referred to their segmentation results reported in [28] on ACDC dataset for comparison.

Methods	Data	LV	MYO	RV	Avg
35 scribbles					
UNet$_{pce}$ [25]	scribbles	.842	.764	.693	.766
UNet$_{wpce}$ [28]	scribbles	.784	.675	.563	.674
UNet$_{CRF}$ [36]	scribbles	.766	.661	.590	.672
CycleMix(ours)	scribbles	**.883**	.798	**.863**	**.848**
35 scribbles + 35 unpaired masks					
UNet$_D$ [28]	scribbles+masks	.404	.597	.753	.585
PostDAE [18]	scribbles+masks	.806	.667	.556	.676
ACCL [35]	scribbles+masks	.878	.797	.735	.803
MAAG [28]	scribbles+masks	.879	**.817**	.752	.816

Effectiveness of global consistency: UNet^{+} (#1) with cross entropy loss of annotated pixels could achieve the average Dice Score of 75.2%. When we added mixed loss $\mathcal{L}_{mix}$ as additional segmentation loss, the average performance increased by 5.7% (75.2% to 80.9%); and when the global consistency ($\mathcal{L}_{con\text{-}g}$) was included for regularization, the average Dice was further boosted to 83.0%. This was attribute to the fact that the combination of global consistency could encourage segmentation model to learn the mix-invariant property, and enhance the ability of model to learn robust shape priors.

Effectiveness of random occlusion: For model #4, we observed that random occlusion ($\mathbb{1}_O$) brought a convincing average Dice Score improvement of 1.3% (84.3% vs 83.0%), demonstrating its effectiveness to enhance the localization ability of model via additional augmentation of scribble supervision.

Table 3. Ablation study: CycleMix for image segmentation with different settings, including loss of unmixed samples ($\mathcal{L}_{unmix}$), loss of mixed samples ($\mathcal{L}_{mix}$), global consistency loss ($\mathcal{L}_{con\text{-}g}$), random occlusion ($\mathbb{1}_O$), local consistency loss ($\mathcal{L}_{con\text{-}l}$). Symbol * indicates statistically significant improvement given by a Wilcoxon signed-rank test with $p \leq 0.05$.

Methods	$\mathcal{L}_{unmix}$	$\mathcal{L}_{mix}$	$\mathcal{L}_{con\text{-}g}$	$\mathbb{1}_O$	$\mathcal{L}_{con\text{-}l}$	LV	MYO	RV	Avg
#1	✓	×	×	×	×	.785±.196	.725±.151	.746±.203	.752
#2	✓	✓	×	×	×	.863±.104*	.783±.086*	.782±.173	.809*
#3	✓	✓	✓	×	×	.867±.130	.786±.114	.837±.097*	.830*
#4	✓	✓	✓	✓	×	**.898±.059***	.786±.078	.847±.132*	.843*
#5	✓	✓	✓	✓	✓	.883±.095	**.798±.075***	**.863±.073**	**.848**

Table 4. Data sensitivity study: the performance of CycleMix with different ratio of scribbles to full annotations.

Methods	scribble: full	LV	MYO	RV	Avg
1	35:00	.883±.095	.798±.075	.863±.073	.848
2	70:00	.880±.115	.825±.072	.860±.089	.855
3	56:14	.898±.075	.842±.072	.876±.112	.872
4	42:28	.911±.063	.854±.056	.883±.076	.883
5	28:42	.902±.080	.851±.065	**.899±.058**	.884
6	14:56	.906±.065	.856±.066	.893±.083	.885
7	00:70	**.919±.065**	**.858±.058**	.882±.088	**.886**
UNet_F^+	00:70	.883±.130	.831±.093	.870±.096	.862

Table 5. Comparisons on fully-supervised segmentation.

Methods	LV	MYO	RV	Avg
ACDC dataset				
UNet_F^+	.883±.130	.831±.093	.870±.096	.862
Puzzle Mix_F [16]	.912±.082	.842±.081	**.887±.066**	.880
CycleMix_F	**.919±.065**	**.858±.058**	.882±.088	**.886**
MSCMRseg dataset				
UNet_F^+	.857±.055	.720±.075	.689±.120	.755
Puzzle Mix_F [16]	**.867±.042**	.742±.043	.759±.039	.789
CycleMix_F	.864±.034	**.785±.042**	**.781±.066**	**.810**

Effectiveness of local consistency: When local consistency ($\mathcal{L}_{con\text{-}l}$) was adopted for shape regularization, model #5 performed marginally better than model #4, with an increase of 0.8% average Dice Score (84.8% vs 84.0%). Particularly on MYO structure, $\mathcal{L}_{con\text{-}l}$ helped obtaining a statistically significant improvement of 1.2% Dice, indicating the benefit of local consistency in shape regularization for segmentation of challenging structures.

4.6. Data sensitivity study

This study investigates the performance of CycleMix with different training images of scribble annotation and full annotation. For this study, we included all the 70 training images from ACDC and altered the ratio between the two sets of annotations. Table. 4 presents the results.

Interestingly, one can observe that when the ratio of full annotation reaches 20% (56:14), CycleMix outperformed the fully-supervised UNet_F^+ by a margin of 1.0% (87.2% vs 86.2%) on the average Dice. As expected, the performance of CycleMix tended to increase as the ratio of fully-annotated subjects increases. One can observe that the general performance of CycleMix converged when the ratio of fully-annotated data reaches 40%. This confirms that CycleMix could achieve a satisfactory segmentation result with a relatively small amount of full annotations.

4.7. Experiments on fully-annotated data

Table. 5 provides the Dice Score of fully-supervised segmentation on ACDC and MSCMRseg datasets. With fully-annotated labels, Puzzle Mix demonstrated competitive performance, improving the average Dice of Unet_F^+ from 86.2% to 88.0% on ACDC, and from 75.5% to 78.9% on MSCMRseg. By contrast, CycleMix could improve more, but the margins were not so exciting as it did in the scribble supervision. This indicates that CycleMix can excel in both scribble supervision-based and fully-supervised segmentation, but its advantage could be more evident in the former applications, for which CycleMix has been specifically designed.

5. Conclusions

In this paper, we have investigated a novel weakly-supervised learning framework, CycleMix, to learn segmentation from scribble supervision. The proposed method utilizes mix augmentation of supervision and cycle consistency of segmentation to enhance the generalization ability of segmentation models. CycleMix was evaluated on two open datasets, *i.e.*, ACDC and MSCMRseg, and achieved new state-of-the-art performance.

References

[1] Wenjia Bai, Hideaki Suzuki, Chen Qin, Giacomo Tarroni, Ozan Oktay, Paul M Matthews, and Daniel Rueckert. Recurrent neural networks for aortic image sequence segmentation with sparse annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 586–594. Springer, 2018. 2

[2] Christian F Baumgartner, Lisa M Koch, Marc Pollefeys, and Ender Konukoglu. An exploration of 2d and 3d deep learning techniques for cardiac mr image segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 111–119. Springer, 2017. 5, 6

[3] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Išgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018. 5

[4] Yigit Baran Can, Krishna Chaitanya, Basil Mustafa, Lisa M. Koch, Ender Konukoglu, and Christian F. Baumgartner. Learning to segment medical images with scribble-supervision alone. In *DLMIA/ML-CDS@MICCAI*, 2018. 1, 2

[5] Krishna Chaitanya, Neerav Karani, Christian F Baumgartner, Anton Becker, Olivio Donati, and Ender Konukoglu. Semi-supervised and task-driven data augmentation. In *International conference on information processing in medical imaging*, pages 29–41. Springer, 2019. 2

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020. 4

[8] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2, 7

[9] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 5

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. 4

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[13] Zhanghexuan Ji, Yan Shen, Chunwei Ma, and Mingchen Gao. Scribble-based hierarchical weakly supervised learning for brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 175–183. Springer, 2019. 1, 2

[14] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 1

[15] JangHyun Kim, Wonho Choo, Hosan Jeong, and Hyun Oh Song. Co-mixup: Saliency guided joint mixup with supermodular diversity. In *International Conference on Learning Representations*, 2021. 2, 3, 5, 7

[16] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International Conference on Machine Learning (ICML)*, 2020. 2, 3, 5, 7, 8

[17] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 3

[18] Agostina J. Larrazabal, C'esar Mart'inez, Ben Glocker, and Enzo Ferrante. Post-dae: Anatomically plausible segmentation via post-processing with denoising autoencoders. *IEEE Transactions on Medical Imaging*, 39:3813–3820, 2020. 2, 6, 7

[19] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016. 1, 2

[20] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1

[21] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 3

[22] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1796–1804, 2015. 1

[23] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *Proceedings of the IEEE international conference on computer vision*, pages 5688–5696, 2017. 1

[24] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020. 1, 2

[25] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1818–1827, 2018. 5, 7

[26] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *ECCV*, 2018. 2, 5

[27] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 1, 3

[28] Gabriele Valvano, Andrea Leo, and Sotirios A. Tsaftaris. Learning to segment from scribbles using multi-scale adversarial attention gates. *IEEE Transactions on Medical Imaging*, pages 1–1, 2021. 2, 5, 6, 7

[29] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019. 2

[30] Dong Wang, Yuan Zhang, Kexin Zhang, and Liwei Wang. Focalmix: Semi-supervised learning for 3d medical image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3951–3960, 2020. 2

[31] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2016. 1

[32] Qian Yue, Xinzhe Luo, Qing Ye, Lingchao Xu, and Xiahai Zhuang. Cardiac segmentation from lge mri using deep neural network incorporating shape and spatial priors. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 559–567. Springer, 2019. 5

[33] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 4, 5, 7

[34] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018. 2, 3, 5, 7

[35] Pengyi Zhang, Yunxin Zhong, and Xiaoqiong Li. Accl: Adversarial constrained-cnn loss for weakly supervised medical image segmentation, 2020. 2, 6, 7

[36] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537, 2015. 2, 5, 7

[37] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3

[38] Xiahai Zhuang. Multivariate mixture model for cardiac segmentation from multi-sequence mri. In *MICCAI*, 2016. 5

[39] Xiahai Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):2933–2946, 2019. 5