(ExMod) Model for Medical Image Segmentation Using Scribble Annotations

Hateef Alshewaier[12*], Yipeng Qin[3], Xianfang Sun[4]

[1] Cardiff University, Cardiff, United Kingdom AlshewaierH@cardiff.ac.uk *
[2] Department of Computer, Applied College, Shaqra University, Saudi Arabia Halshewaier@su.edu.sa
[3] Cardiff University, Cardiff, United Kingdom qiny16@cardiff.ac.uk
[4] Cardiff University, Cardiff, United Kingdom sunx2@cardiff.ac.uk

Abstract. Medical image segmentation presents significant challenges due to the high cost of acquiring precise annotations. The task becomes even more difficult when using weak annotations, such as scribbles, as these annotations provide only limited information about the region of interest. Scribble annotations, however, are easier to acquire in practice, making them a more feasible option. Despite this, training neural networks for segmentation based solely on scribble annotations remains complex. We propose an innovative Expansion and Modification (ExMod) neural network architecture to tackle the challenges inherent in weakly supervised medical image segmentation. While scribble-based supervision has been explored in prior works, ExMod introduces a unique set of enhancements tailored to overcome the limitations of existing methods. Built upon the U-Net framework, our architecture stands out by incorporating multiple advancements designed to boost segmentation accuracy under weak supervision. ExMod introduces additional convolutional layers for richer feature extraction and batch normalization layers to improve training stability and convergence. These modifications lead to superior segmentation performance, particularly when using only scribble annotations. Compared to existing scribble-based methods, ExMod captures intricate image structures more effectively, offering better accuracy with fewer annotations and setting a new benchmark for weakly supervised segmentation. The proposed method was tested on two datasets, i.e., MSCMRseg and ACDC.

Keywords: Machine learning · deep learning · weakly supervised learning · medical image segmentation · scribble.

1 Introduction

Weakly supervised learning requires labeling only a subset of pixels, making it suitable for various domains, including medical image analysis. This approach has gained attention in recent years due to the challenges of fully supervised learning, which demands extensive pixel-wise annotations that are time-consuming and labor-intensive. Over the past decade, weakly supervised neural networks have been proposed for semantic segmentation in various domains using bounding boxes, scribbles, points, image-level labels, and multiple instance learning (MIL). However, few are tailored for medical imaging. Advancements in neural

networks focus on improving efficiency, interpretability, and performance. Fine-tuning pre-trained models on new data is common. Hardware acceleration with GPUs and TPUs enables faster training, though it is costly. Alternatively, model expansion and modification by adding layers or altering architecture can enhance performance without expensive hardware.

To address these challenges, we propose a model expansion and modification designed for medical image segmentation using only scribble annotations. Our novel framework under weak supervision demonstrates promising results and includes the following contributions:

• A new weakly supervised segmentation framework relying solely on scribble annotations, expansion of neural network capabilities with updated architecture incorporating the latest advancements.

2 Related Work

Scribble annotations represent one of the key approaches in weakly supervised learning, which has been developed to leverage weak forms of supervision, such as noisy annotations, image-level labels, and sparse annotations. Among these, the scribble annotation method has emerged as one of the most widely utilized due to its simplicity and the reduced burden it places on annotators [2]. However, the effectiveness of scribble annotations in achieving optimal image segmentation performance depends on several factors, including the complexity of the segmentation task [22], the capabilities of the segmentation algorithm [10], and the required level of accuracy. In general, superior segmentation results are more likely to be obtained when the scribble annotations are both detailed and precise.

Model expansion refers to enhancing or extending an existing machine learning or deep learning model. This can include increasing the model's capacity by adding more layers or parameters to the neural network architecture. Another approach is transfer learning, where a pre-trained model is adapted and fine-tuned for a specific task. For example, [21] extended the U-Net architecture to develop UNet++, specifically designed for medical image segmentation. Their method involves incorporating U-Net models of different sizes into a single framework to improve segmentation accuracy. Similarly, [6] introduced the AMO-Net architecture, which follows an encoding-decoding structure. In their model, the singular encoder-decoder structure was extended to two layers, resulting in improved performance and more effective outcomes. Data augmentation prevents model overfitting and enhances neural network generalization. Mix-up augmentation, introduced by [18], combines two images and their labels using a blending procedure, expanding the training dataset [4, 7, 8, 18]. This method, similar to traditional techniques like flipping and rotation, creates mixed data samples through linear interpolation. [8] proposed Puzzle Mix, which leverages saliency and local statistics for better augmentation, extending the process to multiple images for increased diversity. This extended method, Co-Mixup, generates varied mixed images. Despite appearing unrealistic, these mixed labels provide additional information, improving model training [3]. Consistency regularization is commonly applied in weakly supervised segmentation to improve performance [16]. This technique leverages the advantage of maintaining consistent segmentation results despite variations or perturbations in the input images. By enforcing consistency, the model ensures that segmentation predictions re-

main stable, even when the images undergo transformations or augmentations, thereby enhancing the overall robustness of the segmentation process [13].

3 Method

This section details an enhanced U-Net method for segmenting medical images using scribble annotations. Key improvements include adding convolutional layers in both the encoder and decoder paths to extract richer features and using batch normalization to stabilize and speed up training. These structural changes enhance the convergence rate and final segmentation performance, especially with scribble annotations. Compared to existing scribble-based methods, ExMod captures intricate structures more effectively, yielding more accurate results. The overall framework is shown in Figure 2.

3.1 Mix Augmentation

In this study, we applied the Mix-up augmentation technique in two stages, integrating it with scribble supervision to enhance the segmentation process. The first stage involves augmenting the data by combining two images with their annotations (x_0, y_0), (x_1, y_1). This step aims to maximize the saliency within the mixed image and leverage a finer gradient flow across a larger portion of the labeled pixels, thereby enriching the supervision provided by the scribbles.

$$x_{01}^m = M(x_0, x_1), \quad y_{01}^m = M(y_0, y_1)$$

Where x_0, x_1 The two input images. y_0, y_1 The corresponding labels (or segmentation masks) for the images x_0 and x_1. x_{01}^m The resulting mixed image obtained from combining x_0 and x_1 using the mixing function M. y_{01}^m The mixed label corresponding to the mixed image. $M(x_0, x_1)$ The mixing function that combines the two images pixel by pixel. Similarly, $M(y_0, y_1)$ mixes the labels.

Here is how the mixed result transported from two images that were denoted as (x_{01}^m, y_{01}^m) and computed as:

$$M(a_0, a_1) = (1 - z) \odot \Pi_1^T a_0 + z \odot \Pi_2^T a_1$$

$$\{\Pi_0, \Pi_1, z\} = \arg \max_{\Pi_0, \Pi_1, z} \left[(1 - z) \odot \Pi_0^T s(x_0) + z \odot \Pi_1^T s(x_1)\right]$$

Where a_0, a_1 two inputs being mixed. z a mask controlling the proportion of each input that is mixed. $\odot$ Element-wise multiplication. Π_0^T, Π_1^T Transportation matrices performing spatial transformations on the inputs a_0 and a_1.

$$M(S(x_0), S(x_1)) = S(M(x_0, x_1))$$

After Applying a randomly rotated rectangular region to obscure part of the image, converting the occluded scribbles into the background knowing that the mask has dimension $n \times n$. In our experiment, the size is set as 32×32, and the previous equation becomes:

$$(1 - \mathbb{1}_O) \odot M(\hat{y}_0, \hat{y}_1) = S((1 - \mathbb{1}_O) \odot x_{01}^m)$$

Where $S(\cdot)$ represents the segment and the corresponding segmentation is denoted as $\hat{y} = S(x)$. $\mathbb{1}_O$ denotes a binary mask representing the randomly rotated rectangular region.

The second stage introduces random occlusion, where specific areas of the mixed images containing scribbles are replaced with the background. This procedure reduces the number of scribbles in the training data, simulating incomplete annotations. As demonstrated by [17], this technique has been shown to improve performance in object localization tasks. For the mixing strategy, we adopted the Puzzle Mix method proposed by [8], which effectively augments images while preserving critical structural information. This method was successfully integrated into our framework, contributing to the overall performance of the segmentation model.

3.2 Consistency Regularization

In this section, we adopt two *consistency regularization* methods in our experiment, both of which have demonstrated strong performance in prior work [19] and have proven effective in our proposed approach.

First, *global consistency*, designed to enforce mix-invariant features. This process requires the segmentation of an image to remain consistent under two conditions: (1) the original image and (2) the mixed image. Specifically, this regularization applies consistency between the segmentations of the original images $\{x_0, x_1\}$ and the segmentation of the mixed image $\{x_0^1\}$. Additionally, the random occlusion operation is considered during the mixing process, which is explained earlier.

The next loss function was applied to penalize inconsistent segmentations and calculated using the *negative cosine similarity* between the segmentation of the original image and the mixed image. The loss function for global consistency is expressed as:

$$L_{\text{con-g}} = \frac{1}{2}[L_{\text{ncs}}(p_{01}, q_{01}) + L_{\text{ncs}}(p_{10}, q_{10})]$$

Here, $p_{01} = (1 - \mathbb{1}_O) \odot M(\hat{y}_0, \hat{y}_1)$ and $q_{01} = S((1 - \mathbb{1}_O) \odot x_0^1)$ represent the mixed segmentation and the segmentation of the mixed image, respectively, with similar expressions for p_{10} and q_{10}. The term $L_{\text{ncs}}(\cdot, \cdot)$ denotes the negative cosine similarity between the segmentations.

Second, *local consistency*, is introduced to eliminate disconnected results arising from unlinked regions in the mixed image. This loss function is also calculated using negative cosine similarity, serving as a metric for the distance between the predicted segmentation and its refined version after applying a *morphological operation*. The morphological operation extracts the largest connected area for each non-background class in the input image. We applied this operation when computing the negative cosine similarity for images $\{x_0\}$ and $\{x_1\}$, as proposed by [19], which has proven effective in improving segmentation results during the image mixing process.

The local consistency loss is defined as:

$$L_{\text{con-l}} = \frac{1}{2}[L_{\text{ncs}}(\hat{y}_0, C(\hat{y}_0)) + L_{\text{ncs}}(\hat{y}_1, C(\hat{y}_1))]$$

Where $C(\cdot)$ represents the morphological operation applied to the segmentation result, which outputs the largest connected region for each class in the image.

3.3 Model Expansion and Modification

In this section, we present the proposed enhancements to the U-Net architecture, originally introduced in 2015 [12], which has been widely recognized for its effectiveness in medical image segmentation. While the original design has proven robust, the challenges associated with weakly supervised learning necessitate modifications to enhance model predictions. Our updated U-Net structure demonstrates significant improvements, surpassing state-of-the-art accuracy performance.

Key modifications to improve accuracy performance by doubling the number of convolutional layers in the encoder and decoder paths, enhancing feature extraction and segmentation accuracy [12]. We also incorporated batch normalization layers, which stabilize and accelerate training, improving stability, efficiency, and generalization capabilities, making it more resilient to challenges encountered during training [11]. Additionally, we replaced constant padding with dynamic padding to prevent information loss and reduce computational overhead. These adjustments resulted in smoother image outputs and less noise compared to the CycleMix method [19]. Our method was tested against CycleMix using the same datasets and annotation types, demonstrating superior performance.

Ultimately, our modifications resulted in smoother image outputs, facilitating label generation with reduced noise compared to the CycleMix method [19] as shown in Figure 1. We compared our proposed method and CycleMix, using the same datasets and data annotation types to evaluate performance.

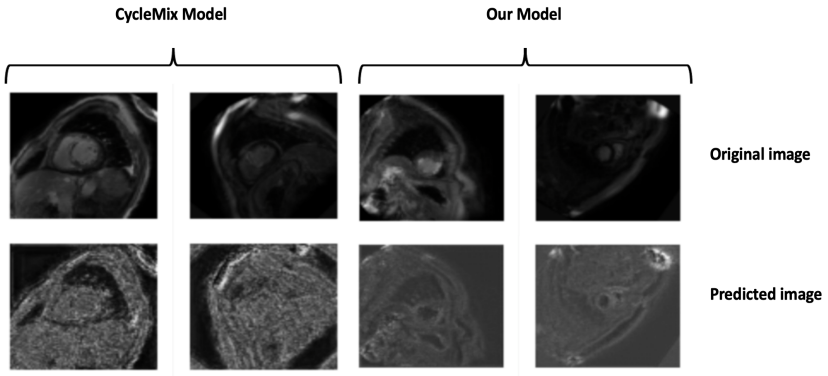

Fig. 1. Comparison between our proposed method and CycleMix method

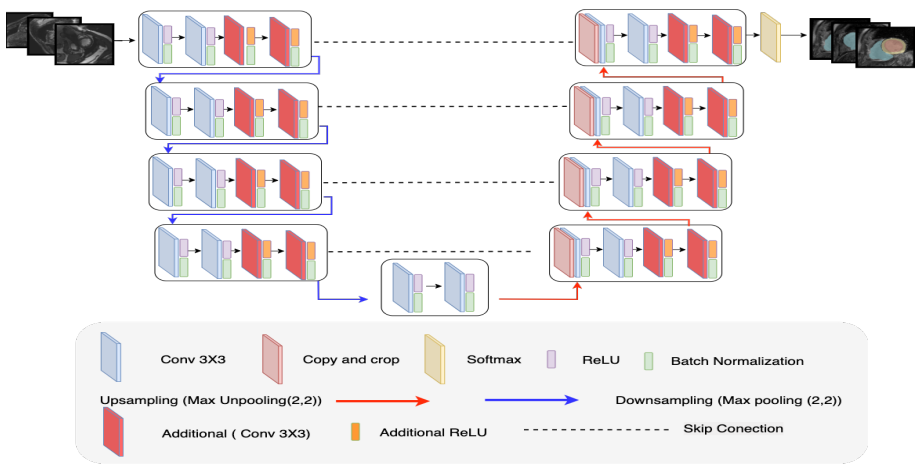

Fig. 2. The framework for the proposed method

4 Experiment

4.1 Dataset Explain

MSCMRseg dataset [23, 24] consists of late gadolinium enhancement (LGE) MRI images from 45 patients with cardiomyopathy. It includes three segmentation classes: right ventricle (RV), left ventricle (LV), and myocardium (MYO). Manually labeled scribble masks for this task, provided by [19]. The dataset was randomly split into 25 training, 5 validation, and 20 test images. 35 images were annotated with scribbles using the same approach as in the ACDC dataset.

ACDC dataset [1] consists of two-dimensional Cine-MRI images collected from 100 patients, with manual annotations for the right ventricle (RV), left ventricle (LV), and myocardium (MYO). The dataset was randomly divided into 70 training, 15 validation, and 15 test images. Of the 100 images, 35 received scribble annotations, targeting key anatomical structures: RV, MYO, and LV. The scribbles covered 27.7% of the RV, 31.3% of the MYO, 24.1% of the LV, and 3.4% of the background, reflecting sparse but strategically placed annotations for effective segmentation.

4.2 Evaluation Metric and Implementation Details

This paper utilized the Dice Coefficient proposed by [5]. The Dice Coefficient calculates the similarity between the ground truth masks and the predicted masks. The proposed method was implemented in the PyTorch environment. The implementation was conducted in Google Colab. The optimizer used is AdamW [9], with a learning rate $1 * 10^{-4}$.

5 Results

The performance of our proposed segmentation method is summarized in Table 1 the Dice Similarity Coefficient for MSCMRseg and ACDC datasets.

5.1 Class-Specific Performance

The RV class achieved a maximum Dice score of 0.94 a mean score of 0.88 on the MSCMRseg dataset, and a maximum score of 0.97 and a mean score of 0.89 on the ACDC dataset. Despite its small size, our method demonstrates high accuracy in segmenting the RV class. For the MYO class, the MSCMRseg dataset showed a maximum Dice score of 0.90 and a mean score of 0.84, while the ACDC dataset showed a maximum score of 0.94 and a mean score of 0.88. This underscores the robustness of our method in accurately segmenting the MYO class, particularly on the ACDC dataset. The LV class exhibited the highest performance, with a maximum Dice score of 0.97, a mean score of 0.93 on the MSCMRseg dataset, and a maximum score of 0.98 and a mean score of 0.93 on the ACDC dataset. These results highlight the effectiveness of our method in segmenting larger anatomical structures.

5.2 Average Performance Metrics

The average Dice scores across the three classes (RV, MYO, and LV) indicate that the MSCMRseg dataset achieved an overall mean score of 0.88, while the ACDC dataset achieved a mean of 0.90. This distinction suggests that our method performs particularly well on the ACDC dataset, contributing to its overall efficacy in medical image segmentation.

5.3 Qualitative Analysis

Figure 3 presents visual comparisons of the segmentation results from our proposed method against the ground truth masks. The qualitative assessment highlights our method's accuracy and reliability in delineating anatomical structures, supporting the quantitative results in Table 1.

Our findings show that our method excels in segmenting both small and large anatomical structures, particularly the MYO and LV classes. The consistently high Dice scores across both datasets demonstrate the robustness and adaptability of our approach, essential for clinical applications where precision and accuracy are crucial.

Table 1. The performance indicates in Dice Scores of the proposed method on MSCMRseg and ACDC datasets

	MSCMRseg		ACDC	
	Maximum	**Mean**	**Maximum**	**Mean**
RV	0.94	0.88	0.97	0.89
MYO	0.90	0.84	0.94	0.88
LV	0.97	0.93	0.98	0.93
Avg	0.94	0.88	0.96	0.90

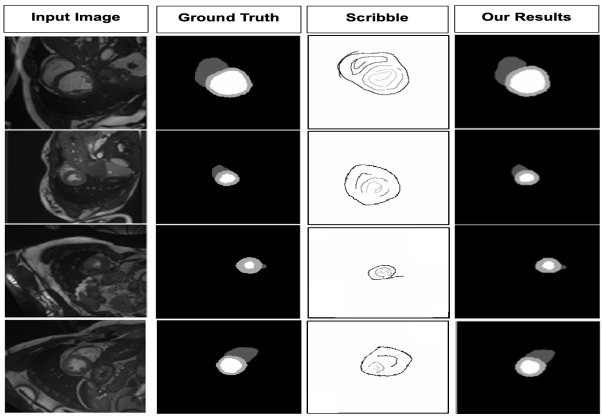

Fig. 3. Results on ACDC dataset compared with the ground truth

Table 2. The performance of Dice Scores on ACDC dataset of proposed results compared with state-of-the-art weakly-supervised methods using Scribble annotation

Method	LV	MYO	RV	Avg
UNet_CRF	0.76	0.66	0.59	0.67
UNet_wpce	0.78	0.67	0.56	0.67
UNet_pce	0.84	0.76	0.69	0.76
CycleMix	0.88	0.79	0.86	0.84
Ours (ACDC)	0.93	0.88	0.89	**0.90**

5.4 Comparison With other Weakly Supervised Methods

The performance of our proposed method was evaluated against various state-of-the-art weakly supervised learning techniques using scribble annotations. As summarized in Table 2, the results demonstrate the effectiveness of our approach on the ACDC dataset. Table 2 presents the Dice scores achieved by our method alongside those from several notable weakly supervised learning approaches, organized in ascending order for clear comparison. The methodologies included are referenced from previous studies [14, 15, 19, 20]. Notably, the CycleMix method [19], previously regarded as the benchmark, achieved an average Dice score of 84% on the ACDC dataset, while our method attained a significantly higher average score of 90%. This improvement highlights the effectiveness of our model in accurately segmenting cardiac structures.

In summary, our method outperformed previous state-of-the-art results, setting new benchmarks for average Dice scores on the ACDC dataset. Improvements in model architecture and the use of weakly supervised learning significantly advanced segmentation accuracy in medical image analysis.

6 Conclusion

In conclusion, this paper introduces an enhanced U-Net framework using scribble annotations for medical image segmentation, significantly improving perfor-

mance. Key modifications include increased network depth, batch normalization, and dynamic padding, enhancing both accuracy and efficiency. The proposed method outperforms current weakly supervised techniques, setting a new benchmark and demonstrating the potential of using scribble annotations in clinical settings where detailed annotations are impractical. Future work will explore data augmentation techniques and the integration of various weak annotations to further improve robustness and generalizability, ultimately enhancing medical image analysis outcomes.

References

[1] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. 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.

[2] Yigit B Can, Krishna Chaitanya, Basil Mustafa, Lisa M Koch, Ender Konukoglu, and Christian F Baumgartner. Learning to segment medical images with scribble-supervision alone. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 236–244. Springer, 2018.

[3] 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.

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

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

[6] Chao Jia and Jianjing Wei. Amo-net: abdominal multi-organ segmentation in mri with a extend unet. In *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, volume 4, pages 1770–1775. IEEE, 2021.

[7] Jang-Hyun Kim, Wonho Choo, Hosan Jeong, and Hyun Oh Song. Co-mixup: Saliency guided joint mixup with supermodular diversity. *arXiv preprint arXiv:2102.03065*, 2021.

[8] 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*, pages 5275–5285. PMLR, 2020.

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2019.

[10] Xiangde Luo, Minhao Hu, Wenjun Liao, Shuwei Zhai, Tao Song, Guotai Wang, and Shaoting Zhang. Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 528–538. Springer, 2022.

[11] Roseline Oluwaseun Ogundokun, Rytis Maskeliunas, Sanjay Misra, and Robertas Damaševičius. Improved cnn based on batch normalization and adam optimizer. In *International Conference on Computational Science and Its Applications*, pages 593–604. Springer, 2022.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[13] Wei Shen, Zelin Peng, Xuehui Wang, Huayu Wang, Jiazhong Cen, Dongsheng Jiang, Lingxi Xie, Xiaokang Yang, and Qi Tian. A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. *IEEE transactions on pattern analysis and machine intelligence*, 45(8):9284–9305, 2023.

[14] 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.

[15] Gabriele Valvano, Andrea Leo, and Sotirios A Tsaftaris. Learning to segment from scribbles using multi-scale adversarial attention gates. *IEEE Transactions on Medical Imaging*, 40(8):1990–2001, 2021.

[16] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12275–12284, 2020.

[17] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[18] H Zhang, M Cisse, Y Dauphin, and D Lopez-Paz. mixup: Beyond empirical risk management. In *6th Int. Conf. Learning Representations (ICLR)*, pages 1–13, 2018.

[19] Ke Zhang and Xiahai Zhuang. Cyclemix: A holistic strategy for medical image segmentation from scribble supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11656–11665, 2022.

[20] 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.

[21] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.

[22] Mingrui Zhuang, Zhonghua Chen, Yuxin Yang, Lauri Kettunen, and Hongkai Wang. Annotation-efficient training of medical image segmentation network based on scribble guidance in difficult areas. *International Journal of Computer Assisted Radiology and Surgery*, 19(1):87–96, 2024.

[23] Xiahai Zhuang. Multivariate mixture model for cardiac segmentation from multi-sequence mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 581–588. Springer, 2016.

[24] Xiahai Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2933–2946, 2018.