Self-supervised Correction Learning for Semi-supervised Biomedical Image Segmentation

Ruifei Zhang[1], Sishuo Liu[2], Yizhou Yu[2,3], and Guanbin Li[1,4(✉)]

[1] Sun Yat-sen University, Guangzhou, China
liguanbin@mail.sysu.edu.cn
[2] The University of Hong Kong, Pokfulam, Hong Kong
[3] Deepwise AI Lab, Beijing, China
[4] Shenzhen Research Institute of Big Data, Shenzhen, China

Abstract. Biomedical image segmentation plays a significant role in computer-aided diagnosis. However, existing CNN based methods rely heavily on massive manual annotations, which are very expensive and require huge human resources. In this work, we adopt a coarse-to-fine strategy and propose a self-supervised correction learning paradigm for semi-supervised biomedical image segmentation. Specifically, we design a dual-task network, including a shared encoder and two independent decoders for segmentation and lesion region inpainting, respectively. In the first phase, only the segmentation branch is used to obtain a relatively rough segmentation result. In the second step, we mask the detected lesion regions on the original image based on the initial segmentation map, and send it together with the original image into the network again to simultaneously perform inpainting and segmentation separately. For labeled data, this process is supervised by the segmentation annotations, and for unlabeled data, it is guided by the inpainting loss of masked lesion regions. Since the two tasks rely on similar feature information, the unlabeled data effectively enhances the representation of the network to the lesion regions and further improves the segmentation performance. Moreover, a gated feature fusion (GFF) module is designed to incorporate the complementary features from the two tasks. Experiments on three medical image segmentation datasets for different tasks including polyp, skin lesion and fundus optic disc segmentation well demonstrate the outstanding performance of our method compared with other semi-supervised approaches. The code is available at https://github.com/ReaFly/SemiMedSeg.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-87196-3_13) contains supplementary material, which is available to authorized users.

M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12902, pp. 134–144, 2021.
https://doi.org/10.1007/978-3-030-87196-3_13

1 Introduction

Medical image segmentation is an essential step in computer-aided diagnosis. In practice, clinicians use various types of images to locate lesions and analyze diseases. An automated and accurate medical image segmentation technique is bound to greatly reduce the workload of clinicians.

With the vigorous development of deep learning, the FCN [15], UNet [19] and their variants [12,23] have achieved superior segmentation performance for both natural images and medical images. However, these methods rely heavily on labeled data, which is time-consuming to acquire especially for medical images. Therefore, many studies adopt semi-supervised learning to alleviate this issue, including GAN-based methods [9,24], consistency training [17,20], pseudo labeling [11] and so on. For instance, Mean Teacher (MT) [20] and its variants [13,22] employ the consistency training for labeled data and unlabeled data by updating teacher weights via an exponential moving average of consecutive student models. Recently, some works [1,14] integrate self-supervised learning such as jigsaw puzzles [16] or contrastive learning [4] to semi-supervised segmentation and achieve competitive results. However, few of them try to dig deeply into the context and structural information of unlabeled images to supplement the semantic segmentation.

In this work, we also consider introducing self-supervised learning to semi-supervised segmentation. In contrast to [1,14], we make full use of massive unlabeled data to exploit image internal structure and boundary characteristics by utilizing pixel-level inpainting as an auxiliary self-supervised task, which is combined with semantic segmentation to construct a dual-task network. As the inpainting of normal non-lesion image content will only introduce additional noise for lesion segmentation, we design a coarse-to-fine pipeline and then enhance the network's representations with the help of massive unlabeled data in the correction stage by only masking the lesion area for inpainting based on the initial segmentation result. Specifically, in the first phase, only the segmentation branch is used to acquire a coarse segmentation result, while in the second step, the masked and original images are sent into the network again to simultaneously perform lesion region inpainting and segmentation separately. Since the two tasks rely on similar feature information, we also design a gated feature fusion (GFF) module to incorporate complementary features for improving each other. Compared with the most related work [2] which introduces a reconstruction task for unlabeled data, their two tasks lack deep interaction and feature reuse, thus cannot collaborate and facilitate each other. Besides, our network not only makes full use of massive unlabeled data, but also explores more complete lesion regions for limited labeled data through the correction phase, which can be seen as "image-level erase [21]" or "reverse attention [3]".

Our contribution is summarized as follows. (1) We propose a novel self-supervised semi-supervised learning paradigm for general lesion region segmentation of medical imaging, and verify that the pretext self-supervised learning task of inpainting the lesion region at the pixel level can effectively enhance the feature learning and greatly reduce the algorithm's dependence on large-scale dense

annotation. (2) We propose a dual-task framework for semi-supervised medical image segmentation. Through introducing the inpainting task, we create supervision signals for unlabeled data to enhance the network's representation learning of lesion regions and also exploit additional lesion features for labeled data, thus effectively correct the initial segmentation results. (3) We evaluate our method on three tasks, including polyp, skin lesion and fundus optic disc segmentation, under a semi-supervision setting. The experimental results demonstrate that our method achieves superior performance compared with other state-of-the-art semi-supervised methods.

2 Methodology

2.1 Overview

In this work, we adopt a coarse-to-fine strategy and propose a self-supervised correction learning paradigm for semi-supervised biomedical image segmentation. Specifically, we introduce inpainting as the pretext task of self-supervised learning to take advantage of massive unlabeled data and thus construct a dual-task network, as shown in Fig. 1. Our proposed framework is composed of a shared encoder, two decoders and five GFF modules placed on each layer of both decoders. We utilize ResNet34 [8] pretrained on the ImageNet [5] as our encoder, which consists of five blocks in total. Accordingly, the decoder branch also has five blocks. Each decoder block is composed of two Conv-BN-ReLU combinations. For the convenience of expression, we use E_{seg}, D_{seg} to represent the encoder and decoder of the segmentation branch, and E_{inp} and D_{inp} for those of the inpainting branch.

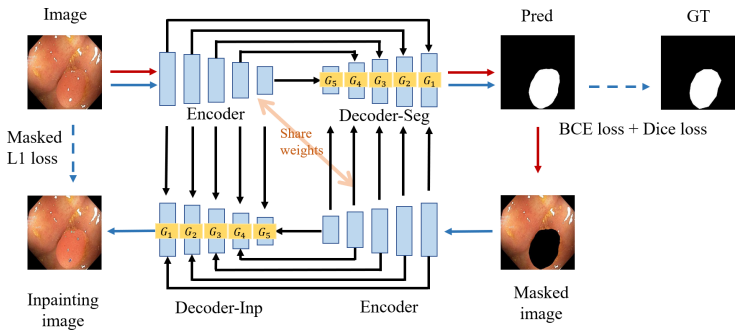

Fig. 1. The overview of our network. Both encoders share weights. G_1–G_5 represent five GFF modules. The red and blue arrows denote the input and output of our network in the first and second stage respectively. (Color figure online)

In the first step, given the image $x \in \mathbb{R}^{H \times W \times C}$, in which H, W, C are the height, width and channels of the image respectively, we use the segmenta-

tion branch E_{seg}, D_{seg} with skip-connections, the traditional U-shape encoder-decoder structure, to obtain a coarse segmentation map $\hat{y}_{coarse}$ and then mask the original input based on its binary result $\overline{y}_{coarse}$ by the following formulas:

$$\hat{y}_{coarse} = D_{seg}(E_{seg}(x)) \tag{1}$$

$$x_{mask} = x \times (1 - \overline{y}_{coarse}) \tag{2}$$

In the second phase, the original image x and the masked image x_{mask} are sent into E_{seg} and E_{inp} simultaneously to extract features e_{seg} and e_{inp}. Obviously, e_{seg} is essential for the inpainting task, and since the initial segmentation is usually inaccurate and incomplete, e_{inp} may also contain important residual lesion features for the correction of the initial segmentation. In order to adaptively select the useful features of e_{inp} and achieve complementary fusion of e_{seg} and e_{inp}, we design the GFF modules (G_1–G_5) and place them on each decoder layer of both branches. Specifically, for the i^{th} layer, the features e^i_{seg} and e^i_{inp} are delivered into G_i through skip-connections to obtain the fusion $e^i = G_i(e^i_{seg}, e^i_{inp})$, and then sent to the corresponding decoder layer. Thus, both G_i of the two branches shown in Fig. 1 actually share parameters, taking the same input and generating the identical output. To enhance the learning of the GFF modules, we adopt a deep supervision strategy and each layer of the two decoder branches generate a segmentation result and an inpainting result respectively by the following formulas:

$$\hat{y}^i_{fine} = \begin{cases} D^i_{seg}([e^i, d^{i+1}_{seg}]), & i = 1, 2, 3, 4 \\ D^i_{seg}(e^i), & i = 5 \end{cases} \tag{3}$$

$$\hat{x}^i = \begin{cases} D^i_{inp}([e^i, d^{i+1}_{inp}]), & i = 1, 2, 3, 4 \\ D^i_{inp}(e^i), & i = 5 \end{cases} \tag{4}$$

Where $[\cdot, \cdot]$ denotes the concatenation process, and d^{i+1}_{seg}, d^{i+1}_{inp} represent the features from previous decoder layers. The deep supervision strategy can also avoid D_{inp} directly copying the features of the low-level e_{seg} to complete the inpainting task without in-depth lesion feature mining. The output of the last layer $\hat{y}^1_{fine}$ is the final segmentation result of our network.

2.2 Gated Feature Fusion (GFF)

To better incorporate complementary features and filter out the redundant information, we design the GFF modules placed on each decoder layer to integrate the features delivered from the corresponding encoder layer of two branches. The details are shown in Fig. 2. Our GFF module consists of a reset gate and a select gate. Specifically, for the G_i placed on the i^{th} decoder layer, the value of two gates is calculated as follows:

$$r_i = \sigma(W_r \left[e^i_{seg}, e^i_{inp}\right]) \tag{5}$$

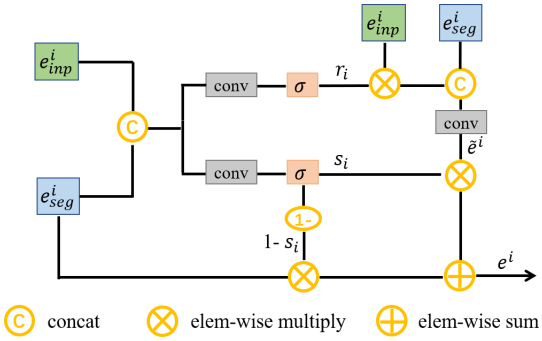

Fig. 2. Gated feature fusion module

$$s_i = \sigma(W_s \left[e^i_{seg}, e^i_{inp}\right]) \tag{6}$$

Where W_r, W_s denote the convolution process, taking the concatenation of e^i_{seg} and e^i_{inp} as input. σ represents the Sigmoid function. r_i and s_i represent the value of the reset gate and the select gate, respectively. Since the input of the inpainting branch is the masked image, the reset gate is necessary to suppress massive invalid background information. And then the select gate achieves adaptive and complementary feature fusion between the reintegrated features $\tilde{e}^i$ and the original segmentation feature e^i_{seg} by the following operations:

$$\tilde{e}^i = W \left[r_i \times e^i_{inp}, e^i_{seg}\right]) \tag{7}$$

$$e^i = s_i \times \tilde{e}^i + (1 - s_i) \times e^i_{seg} \tag{8}$$

where W also represents the convolution process to make the reintegrated features $\tilde{e}^i$ have the same dimension with e^i_{seg}.

2.3 Loss Function

We only calculate loss in the second stage. The labeled dataset and unlabeled dataset are denoted as D_l and D_u. For the labeled data $x_l \in D_l$, y_l is the Ground Truth. Since we adopt the deep supervision strategy, the overall loss is the sum of the combination of Binary CrossEntropy (BCE) Loss and Dice loss between each output and the Ground Truth:

$$\mathcal{L}_{seg}(x_l) = \sum_{i=1}^{5} L^i_{BCE}(\hat{y}^i_l, y^i_l) + L^i_{Dice}(\hat{y}^i_l, y^i_l) \tag{9}$$

where $\hat{y}^i_l$, y^i_l denote the segmentation map $\hat{y}^i_{fine}$ of the i^{th} decoder layer and the corresponding down-sampling Ground Truth y_l.

For unlabeled data $x_u \in D_u$, the inpainting loss is the sum of $L1$ loss between each inpainting image and the original image in the masked region:

$$\mathcal{L}_{inp}(x_u) = \sum_{i=1}^{5} \overline{y}_u^i \times \left| \hat{x}_u^i - x_u^i \right| \tag{10}$$

where $\hat{x}_u^i$, x_u^i and $\overline{y}_u^i$ represent the inpainting image, down-sampling original image and binary segmentation result of the i^{th} decoder layer, respectively. In the end, the total loss function is formulated as follows:

$$\mathcal{L} = \lambda_1 \sum_{x_l \in D_l} \mathcal{L}_{seg}(x_l) + \lambda_2 \sum_{x_u \in D_u} \mathcal{L}_{inp}(x_u) \tag{11}$$

where λ_1, λ_2 are weights balancing the segmentation loss and the inpainting loss. And we set $\lambda_1 = 2$ and $\lambda_2 = 1$ in our experiments.

3 Experimental Results

3.1 Dataset and Evaluation Metric

We conduct experiments on a variety of medical image segmentation tasks to verify the effectiveness and robustness of our approach, including polyp, skin lesion and fundus optic disc segmentation, respectively.

Polyp Segmentation. We use the publicly available kvasir-SEG [10] dataset containing 1000 images, and randomly select 600 images as the training set, 200 images as the validation set, and the remaining as the test set.

Skin Lesion Segmentation. We utilize the ISBI 2016 skin lesion dataset [7] to evaluate our method performance. This dataset consists of 1279 images, among which 900 are used for training and the others for testing.

Optic Disc Segmentation. The Rim-one r1 dataset [6] is utilized in our experiments, which has 169 images in total. We randomly split the dataset into a training set and a test set with the ratio of 8:2.

Evaluation Metric. Referring to common semi-supervised segmentation settings [13,22], for all datasets, we randomly use 20% of the training set as the labeled data, 80% as the unlabeled data and adopt five metrics to quantitively evaluate the performance of our approach and other methods, including "Dice Similarity Coefficient (Dice)", "Intersection over Union (IoU)", "Accuracy (Acc)", "Recall (Rec)" and"Specificity (Spe)".

3.2 Implementation Details

Data Pre-processing. In our experiments, since the image resolution of all datasets varies greatly, we uniformly resize all images into a fixed size of 320×320 for training and testing. And in the training stage, we use data augmentation,

including random horizontal and vertical flips, rotation, zoom, and finally all the images are randomly cropped to 256×256 as input.

Training Details. Our method is implemented using PyTorch [18] framework. We set batch size of the training process to 4, and use SGD optimizer with a momentum of 0.9 and a weight decay of 0.00001 to optimize the model. A poly learning rate police is adopted to adjust the initial learning rate, which is $lr = init\_lr \times (1 - \frac{iter}{max\_iter})^{power}$, where $init\_lr = 0.001$, $power = 0.9$). The total number of epochs is set to 80.

Table 1. Comparison with other state-of-the-art methods and ablation study on the Kvasir-SEG dataset

Methods	Data	*Dice*	*IoU*	*Acc*	*Rec*	*Spe*
Supervised	600L (All)	89.48	83.69	97.34	91.06	98.58
Supervised	120L	84.40	76.18	96.09	85.35	98.55
DAN [24]	120L + 480U	85.77	78.12	96.37	86.86	98.53
MT [20]	120L + 480U	85.99	78.84	96.21	86.81	**98.79**
UA-MT [22]	120L + 480U	85.70	78.34	96.38	88.51	98.40
TCSM_V2 [13]	120L + 480U	86.17	79.15	96.38	87.14	98.76
MASSL [2]	120L + 480U	86.45	79.61	96.34	89.18	98.32
Ours	120L + 480U	**87.14**	**80.49**	**96.42**	**90.78**	97.89
Ours (add)	120L + 480U	85.59	78.56	96.12	87.98	98.26
Ours (concat)	120L + 480U	86.09	78.98	96.21	90.54	97.63

3.3 Comparisons with the State-of-the-Art

In our experiments, ResNet34 [8] based UNet [19] is utilized as our baseline, which is trained using all training set and our selected 20% labeled data separately in a fully-supervised manner. Besides, we compare our method with other state-of-the-art approaches, including DAN [24], MASSL [2], MT [20] and its variants (UA-MT [22], TCSM_V2 [13]). All comparison methods adopt ResNet34UNet as the backbone and use the same experimental settings for a fair comparison. On the **Kvasir-SEG dataset**, Table 1 shows that our method obtains the outstanding performance compared with other semi-supervised methods, with Dice of 87.14%, which is 2.74% improvement over the baseline only using the 120 labeled data, outperforming the second best method by 0.69%. On the **ISBI 2016 skin lesion dataset**, we obtain a 90.95% Dice score, which is superior to other semi-supervised methods and very close to the score of 91.38% achieved by the baseline using all training set images. On the **Rim-one r1 dataset**, we can conclude that our method achieves the best performance over five metrics, further demonstrating the effectiveness of our method. Note

that detailed results on the latter two datasets are listed in the supplementary material due to the space limitation. Some visual segmentation results are shown in Fig. 3 (col. 1–8).

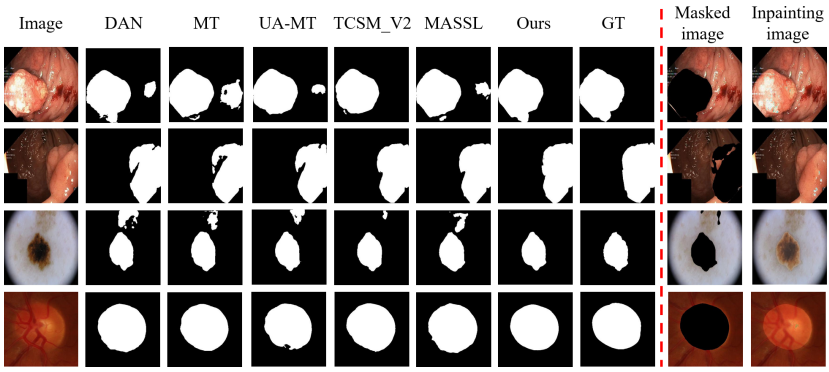

Fig. 3. Visual comparison of various lesion segmentation from state-of-the-art methods. Our proposed method consistently produces segmentation results closest to the ground truth. The inpainting result is shown in the rightmost column.

3.4 Ablation Study

Effectiveness of Our Approach with Different Ratio of Labeled Data. We draw the curves of Dice score under three settings in Fig. 4. To verify that our proposed framework can mine residual lesion features and enhance the lesion representation by GFF modules in the second stage, we conduct experiments and draw the blue line. The blue line denotes that our method uses the same labeled data with the baseline (the red line) to perform the two-stage process, without utilizing any unlabeled data. Note that we only calculate the segmentation loss for the labeled data. The performance gains compared with the baseline show that our network mines useful lesion information in the second stage. The green line means that our method introduces the remaining as unlabeled data for the inpainting task, further enhancing the feature representation learning of the lesion regions and improving the segmentation performance, especially when only a small amount of labeled data is used. When using 100% labeled data, the green line is equivalent to the blue line since no additional unlabeled data is utilized to do the inpainting task, thus maintaining the same results.

Effectiveness of the GFF Modules. To verify the effectiveness of the GFF modules, we also design two variants, which merge features by directly addition and concatenation, denoting as Ours (add) and Ours (concat) respectively. In Table 1, we can observe performance degradation by both approaches compared with our method, proving that the GFF module plays a significant role in filtering redundant information and improving the model performance.

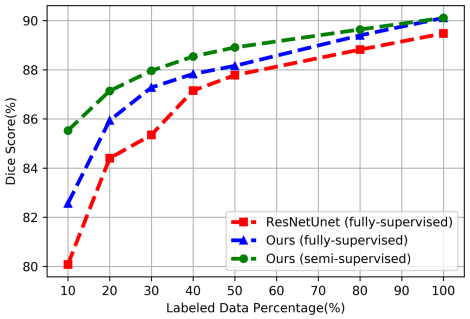

Fig. 4. The performance of our method with different ratio of labeled data on the Kvasir-SEG dataset.

4 Conclusions

In this paper, we believe that massive unlabeled data contains rich context and structural information, which is significant for lesion segmentation. Based on this, we introduce the self-supervised inpainting branch for unlabeled data, cooperating with the main segmentation task for labeled data, to further enhance the representation for lesion regions, thus refine the segmentation results. We also design the GFF module for better feature selection and aggregation from the two tasks. Experiments on various medical datasets have demonstrated the superior performance of our method.

Acknowledgement. This work is supported in part by the Key-Area Research and Development Program of Guangdong Province (No. 2020B0101350001), in part by the Guangdong Basic and Applied Basic Research Foundation (No. 2020B1515020048), in part by the National Natural Science Foundation of China (No. 61976250) and in part by the Guangzhou Science and technology project (No. 202102020633).

References

1. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations. arXiv preprint arXiv:2006.10511 (2020)
2. Chen, S., Bortsova, G., García-Uceda Juárez, A., van Tulder, G., de Bruijne, M.: Multi-task attention-based semi-supervised learning for medical image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 457–465. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_51
3. Chen, S., Tan, X., Wang, B., Hu, X.: Reverse attention for salient object detection. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11213, pp. 236–252. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01240-3_15
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML, pp. 1597–1607. PMLR (2020)

5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR, pp. 248–255. IEEE (2009)
6. Fumero, F., Alayón, S., Sanchez, J.L., Sigut, J., Gonzalez-Hernandez, M.: RIM-ONE: an open retinal image database for optic nerve evaluation. In: 24th International Symposium on Computer-Based Medical Systems (CBMS), pp. 1–6. IEEE (2011)
7. Gutman, D., et al.: Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1605.01397 (2016)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
9. Hung, W.C., Tsai, Y.H., Liou, Y.T., Lin, Y.Y., Yang, M.H.: Adversarial learning for semi-supervised semantic segmentation. arXiv preprint arXiv:1802.07934 (2018)
10. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. In: Ro, Y.M., De Neve, W., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 451–462. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_37
11. Lee, D.H.: Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, ICML, vol. 3, p. 896 (2013)
12. Li, G., Yu, Y.: Deep contrast learning for salient object detection. In: CVPR, pp. 478–487 (2016)
13. Li, X., Yu, L., Chen, H., Fu, C.W., Xing, L., Heng, P.A.: Transformation-consistent self-ensembling model for semisupervised medical image segmentation. IEEE Trans. Neural Netw. Learn. Syst. **32**(2), 523–534 (2020)
14. Li, Y., Chen, J., Xie, X., Ma, K., Zheng, Y.: Self-loop uncertainty: a novel pseudo-label for semi-supervised medical image segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 614–623. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_60
15. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440 (2015)
16. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving Jigsaw puzzles. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 69–84. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_5
17. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: CVPR, pp. 12674–12684 (2020)
18. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. In: NeurIPS, pp. 8026–8037 (2019)
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
20. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS, pp. 1195–1204 (2017)
21. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: a simple classification to semantic segmentation approach. In: CVPR, pp. 1568–1576 (2017)

22. Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 605–613. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_67
23. Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., Yu, Y.: Adaptive context selection for polyp segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 253–262. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_25
24. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 408–416. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_47