



Semi-supervised Pathological Image Segmentation via Cross Distillation of Multiple Attentions

Lanfeng Zhong¹, Xin Liao², Shaoting Zhang^{1,3}, and Guotai Wang^{1,3}(✉)

¹ University of Electronic Science and Technology of China, Chengdu, China
guotai.wang@uestc.edu.cn

² Department of Pathology, West China Second University Hospital, Sichuan University, Chengdu, China

³ Shanghai Artificial Intelligence Laboratory, Shanghai, China

Abstract. Segmentation of pathological images is a crucial step for accurate cancer diagnosis. However, acquiring dense annotations of such images for training is labor-intensive and time-consuming. To address this issue, Semi-Supervised Learning (SSL) has the potential for reducing the annotation cost, but it is challenged by a large number of unlabeled training images. In this paper, we propose a novel SSL method based on Cross Distillation of Multiple Attentions (CDMA) to effectively leverage unlabeled images. Firstly, we propose a Multi-attention Tri-branch Network (MTNet) that consists of an encoder and a three-branch decoder, with each branch using a different attention mechanism that calibrates features in different aspects to generate diverse outputs. Secondly, we introduce Cross Decoder Knowledge Distillation (CDKD) between the three decoder branches, allowing them to learn from each other's soft labels to mitigate the negative impact of incorrect pseudo labels in training. Additionally, uncertainty minimization is applied to the average prediction of the three branches, which further regularizes predictions on unlabeled images and encourages inter-branch consistency. Our proposed CDMA was compared with eight state-of-the-art SSL methods on the public DigestPath dataset, and the experimental results showed that our method outperforms the other approaches under different annotation ratios. The code is available at <https://github.com/HiLab-git/CDMA>.

Keywords: Semi-supervised learning · Knowledge distillation · Attention · Uncertainty

1 Introduction

Automatic segmentation of tumor lesions from pathological images plays an important role in accurate diagnosis and quantitative evaluation of cancers. Recently, deep learning has achieved remarkable performance in pathological image segmentation when trained with a large and well-annotated dataset [6, 13,

20]. However, obtaining dense annotations for pathological images is challenging and time-consuming, due to the extremely large image size (e.g., 10000×10000 pixels), scattered spatial distribution, and complex shape of lesions.

Semi-Supervised Learning (SSL) is a potential technique to reduce the annotation cost via learning from a limited number of labeled data along with a large amount of unlabeled data. Existing SSL methods can be roughly divided into two categories: consistency-based [9, 14, 23] and pseudo label-based [2] methods. The consistency-based methods impose consistency constraints on the predictions of an unlabeled image under some perturbations. For example, Mean Teacher (MT)-based methods [14, 23] encourage consistent predictions between a teacher and a student model with noises added to the input. Xie et al. [21] introduced a pairwise relation network to exploit semantic consistency between each pair of images in the feature space. Luo et al. [9] proposed an uncertainty rectified pyramid consistency between multi-scale predictions. Jin et al. [7] proposed to encourage the predictions of auxiliary decoders and a main decoder to be consistent under perturbed hierarchical features. Pseudo label-based methods typically generate pseudo labels for labeled images to supervise the network [4]. Since using a model’s prediction to supervise itself may over-fit its bias, Chen et al. [2] proposed Cross Pseudo Supervision (CPS) where two networks learn from each other’s pseudo labels generated by *argmax* of the output prediction. MC-Net+ [19] utilized multiple decoders with different upsampling strategies to obtain slightly different outputs, and each decoder’s probability output was sharpened to serve as pseudo labels to supervise the others. However, the pseudo labels are not accurate and contain a lot of noise, using *argmax* or sharpening operation will lead to over-confidence of potentially wrong predictions, which limits the performance of the models. Additionally, some related works advocated the entropy-minimization methods. Typical entropy Minimization (EM) [15] that aims to reduce the uncertainty or entropy in a system. Wu et al. [17] directly applied entropy minimization to the segmentation results.

In this work, we propose a novel and efficient method based on Cross Distillation with Multiple Attentions (CDMA) for semi-supervised pathological image segmentation. Firstly, a Multi-attention Tri-branch Network (MTNet) is proposed to efficiently obtain diverse outputs for a given input. Unlike MC-Net+ [19] that is based on different upsampling strategies, our MTNet uses different attention mechanisms in three decoder branches that calibrate features in different aspects to obtain diverse and complementary outputs. Secondly, inspired by the observation that smoothed labels are more effective for noise-robust learning found in recent studies [10, 22], we propose a Cross Decoder Knowledge Distillation (CDKD) strategy to better leverage the diverse predictions of unlabeled images. In CDKD, each branch serves as a teacher of the other two branches using soft label supervision, which reduces the effect of noise for more robust learning from inaccurate pseudo labels than *argmax* [2] and sharpening-based [19] pseudo supervision in existing methods. Differently from typical Knowledge Distillation (KD) methods [5, 24] that require a pre-trained teacher to generate soft predictions, our method efficiently obtains the teacher and student’s soft

predictions simultaneously in a single forward pass. In addition, we apply an uncertainty minimization-based regularization to the average probability prediction across the decoders, which not only increases the network’s confidence, but also improves the inter-decoder consistency for leveraging labeled images.

The contribution of this work is three-fold: 1) A novel framework named CDMA based on MTNet is introduced for semi-supervised pathological image segmentation, which leverages different attention mechanisms for generating diverse and complementary predictions for unlabeled images; 2) A Cross Decoder Knowledge Distillation method is proposed for robust and efficient learning from noisy pseudo labels, which is combined with an average prediction-based uncertainty minimization to improve the model’s performance; 3) Experimental results show that the proposed CDMA outperforms eight state-of-the-art SSL methods on the public DigestPath dataset [3].

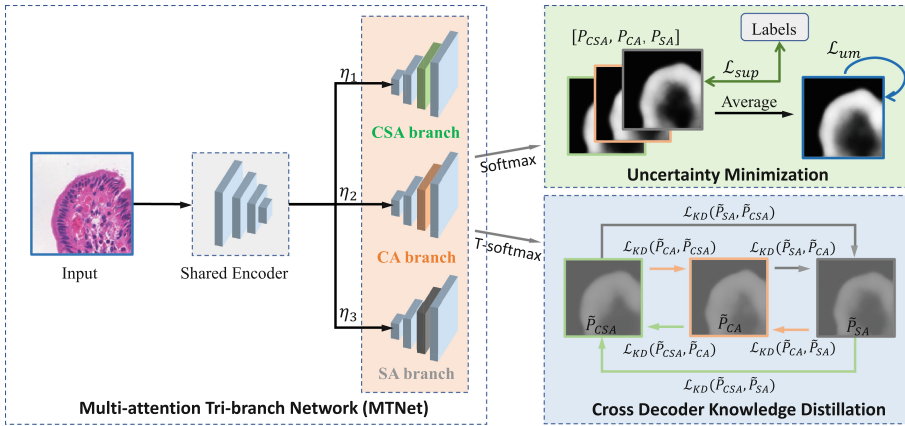


Fig. 1. Our CDMA for semi-supervised segmentation. Three decoder branches use different attentions to obtain diverse outputs. Cross Decoder Knowledge Distillation (CDKD) is proposed to better deal with noisy pseudo labels, and an uncertainty minimization is applied to the average probability prediction of the three branches. \mathcal{L}_{sup} is only for labeled images.

2 Methods

As illustrated in Fig. 1, the proposed Cross Distillation of Multiple Attentions (CDMA) framework for semi-supervised pathological image segmentation consists of three core modules: 1) a tri-branch network MTNet that uses three different attention mechanisms to obtain diverse outputs, 2) a Cross Decoder Knowledge Distillation (CDKD) module to reduce the effect of noisy pseudo labels based on soft supervision, and 3) an average prediction-based uncertainty minimization loss to further regularize the predictions on unlabeled images.

2.1 Multi-attention Tri-Branch Network (MTNet)

Attention is an effective network structure design in fully supervised image segmentation [12, 16]. It can calibrate the feature maps for better performance by paying more attention to the important spatial positions or channels with only a few extra parameters. However, it has been rarely investigated in semi-supervised segmentation tasks. To more effectively exploit attention mechanisms for semi-supervised pathological image segmentation, our proposed MTNet consists of a shared encoder and three decoder branches that are based on Channel Attention (CA), Spatial Attention (SA) and simultaneous Channel and Spatial Attention (CSA), respectively. The encoder consists of multiple convolutional blocks that are sequentially connected to a down-sampling layer, and each decoder has multiple convolutional blocks that are sequentially connected by an up-sampling layer. For a certain decoder, it uses CA, SA or SCA at the convolutional block at each resolution level to calibrate the features.

CA branch uses channel attention blocks to calibrate the features in the first decoder. A channel attention block highlights important channels in a feature map and it is formulated as:

$$F_c = F \cdot \sigma \left(MLP(Pool_{avg}^S(F)) + MLP(Pool_{max}^S(F)) \right) \quad (1)$$

where F represents an input feature map. $Pool_{avg}^S$ and $Pool_{max}^S$ represent average pooling and max-pooling across the spatial dimension, respectively. MLP and σ denote multi-layer perception and the sigmoid activation function respectively. F_c is the output feature map calibrated by channel attention.

SA branch leverages spatial attention to highlight the most relevant spatial positions and suppress the irrelevant regions in a feature map. An SA block is:

$$F_s = F \cdot \sigma \left(Conv(Pool_{avg}^C(F) \oplus Pool_{max}^C(F)) \right) \quad (2)$$

where $Conv$ denotes a convolutional layer. $Pool_{avg}^C$ and $Pool_{max}^C$ are average and max-pooling across the channel dimension, respectively. \oplus means concatenation.

CSA branch calibrates the feature maps using a CSA block for each convolutional block. A CSA block consists of a CA block followed by an SA block, taking advantage of channel and spatial attention simultaneously.

Due to the different attention mechanisms, the three decoder branches pay attention to different aspects of feature maps and lead to different outputs. To further improve the diversity of the outputs and alleviate over-fitting, we add a dropout layer and a feature noise layer η [11] before each of the three decoders. For an input image, the logit predictions obtained by the three branches are denoted as Z_{CA} , Z_{SA} and Z_{CSA} , respectively. After using a standard Softmax operation, their corresponding probability prediction maps are denoted as P_{CA} , P_{SA} and P_{CSA} , respectively.

2.2 Cross Decoder Knowledge Distillation (CDKD)

Since the three branches have different decision boundaries, using the predictions from one branch as pseudo labels to supervise the others would avoid each branch over-fitting its bias. However, as the predictions for unlabeled training images are noisy and inaccurate, using hard or sharpened pseudo labels [2, 19] would strengthen the confidence on incorrect predictions, leading the model to overfit the noise [10, 22]. To address this problem, we introduce CDKD to enhance the ability of our MTNet to leverage unlabeled images and eliminate the negative impact of noisy pseudo labels. It forces each decoder to be supervised by the other two decoders' soft predictions. Following the practice of KD [5], a temperature calibrated Softmax (T-Softmax) is used to soften the probability maps:

$$\tilde{\mathbf{p}}_c = \frac{\exp(\mathbf{z}_c/T)}{\sum_c \exp(\mathbf{z}_c/T)} \quad (3)$$

where \mathbf{z}_c represents the logit prediction for class c of a pixel, and $\tilde{\mathbf{p}}_c$ is the soft probability value for class c . Temperature T is a parameter to control the softness of the output probability. Note that $T = 1$ corresponds to a standard Softmax function, and a larger T value leads to a softer probability distribution with higher entropy. When $T < 1$, Eq. 3 is a sharpening function.

Let \tilde{P}_{CA} , \tilde{P}_{SA} and \tilde{P}_{CSA} represent the soft probability map obtained by T-Softmax for the three branches, respectively. With the other two branches being the teachers, the KD loss for the CSA branch is:

$$\mathcal{L}_{kd}^{CSA} = \mathbf{KL}(\tilde{P}_{CSA}, \tilde{P}_{CA}) + \mathbf{KL}(\tilde{P}_{CSA}, \tilde{P}_{SA}) \quad (4)$$

where $\mathbf{KL}()$ is the Kullback-Leibler divergence function. Note that the gradient of \mathcal{L}_{kd}^{CSA} is only back-propagated to the CSA branch, so that the knowledge is distilled from the teachers to the student. Similarly, the KD losses for the CA and SA branches are denoted as \mathcal{L}_{kd}^{CA} and \mathcal{L}_{kd}^{SA} , respectively. Then, the total distillation loss is defined as:

$$\mathcal{L}_{cdkd} = \frac{1}{3}(\mathcal{L}_{kd}^{CSA} + \mathcal{L}_{kd}^{CA} + \mathcal{L}_{kd}^{SA}) \quad (5)$$

2.3 Average Prediction-Based Uncertainty Minimization

Minimizing the uncertainty (e.g., entropy) [15] has been shown to be an effective regularization for predictions on unlabeled images, which increases the model's confidence on its predictions. However, applying uncertainty minimization to each branch independently may lead to inconsistent predictions between the decoders where each of them is very confident, e.g., two branches predict the foreground probability of a pixel as 0.0 and 1.0 respectively. To avoid this problem and further encourage inter-decoder consistency for regularization, we propose an average prediction-based uncertainty minimization:

$$\mathcal{L}_{um} = -\frac{1}{N} \sum_{i=0}^N \sum_{c=0}^C \bar{P}_i^c \log(\bar{P}_i^c) \quad (6)$$

where $\bar{P} = (P_{CSA} + P_{CA} + P_{SA})/3$ is the average probability map. C and N are the class number and pixel number respectively. \bar{P}_i^c is the average probability for class c at pixel i . Note that when \mathcal{L}_{um} for a pixel is close to zero, the average probability for class c of that pixel is close to 0.0 (1.0), which drives all the decoders to predict it as 0.0 (1.0) and encourages inter-decoder consistency.

Finally, the overall loss function for our CDMA is:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_1 \mathcal{L}_{cdkd} + \lambda_2 \mathcal{L}_{um} \quad (7)$$

where $\mathcal{L}_{sup} = (\mathcal{L}_{sup}^{CSA} + \mathcal{L}_{sup}^{CA} + \mathcal{L}_{sup}^{SA})/3$ is the average supervised learning loss for the three branches on the labeled training images, and the supervised loss for each branch calculates the Dice loss and cross entropy loss between the probability prediction (P_{CSA} , P_{CA} and P_{SA}) and the ground truth label. λ_1 and λ_2 are the weights of \mathcal{L}_{cdkd} and \mathcal{L}_{um} respectively. Note that \mathcal{L}_{cdkd} and \mathcal{L}_{um} are applied on both labeled and unlabeled training images.

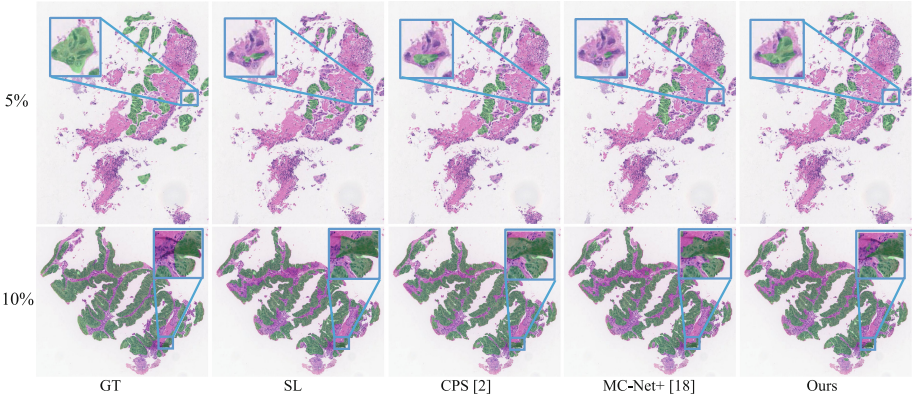


Fig. 2. Visual comparison between our proposed CDMA with state-of-the-art methods for semi-supervised semantic segmentation of WSIs. The green regions are lesions.

3 Experiments and Results

Dataset and Implementation Details. We used the public DigestPath dataset [3] for binary segmentation of colonoscopy tumor lesions from Whole Slide Images (WSI) in the experiment. The WSIs were collected from four medical institutions of $\times 20$ magnification ($0.475 \mu\text{m}/\text{pixel}$) with an average size of 5000×5000 . We randomly split 130 malignant WSIs into 100, 10, and 20 for training, validation and testing, respectively. For SSL, we investigated two annotation ratios: 5% and 10%, where only 5 and 10 WSIs in the training set were taken as annotated respectively. Labeled WSIs were randomly selected. For computational feasibility, we cropped the WSIs into patches with a size of 256×256 .

At inference time for segmenting a WSI, we used a sliding window of size 256×256 with a stride of 192×192 .

The CDMA framework was implemented in PyTorch, and all experiments were performed on one NVIDIA 2080Ti GPU. MTNet was implemented by extending DeepLabv3+ [1] into a tri-branch network, where the three decoders were equipped with CA, SA and CSA blocks respectively. The encoder used a backbone of ResNet50 pre-trained on ImageNet. The kernel size of *Conv* in the SA block is 7×7 . SGD optimizer was used for training, with weight decay 5×10^{-4} , momentum 0.9 and epoch number 150. The learning rate was initialized to 10^{-3} and decayed by 0.1 every 50 epochs. The hyper-parameter setting was $\lambda_1 = \lambda_2 = 0.1$, $T = 10$ based on the best results on the validation set. The batch size was 16 (8 labeled and 8 unlabeled patches). For data augmentation, we adopted random flipping, random rotation, and random Gaussian noise. For inference, only the CSA branch was used due to the similar performance of the three branches after converge and the increased inference time of their ensemble, and no post-processing was used. Dice Similarity Coefficient (DSC) and Jaccard Index (JI) were used for quantitative evaluation.

Table 1. Comparison between different SSL methods on the DigestPath dataset. * denotes p -value < 0.05 (significance level) when comparing the proposed CDMA with the others under t-test hypothesis testing.

Methods	DSC		Jaccard Index	
	5% labeled	10% labeled	5% labeled	10% labeled
SL lower bound	64.74±23.24*	68.32±21.18*	52.35±21.53*	53.62±20.32*
EM [15]	67.09±24.28*	70.01±22.24*	54.55±22.40*	56.96±21.70*
MT [14]	67.46±23.10*	70.19±21.72*	54.68±21.27*	56.38±21.21*
UAMT [23]	67.76±23.44	69.64±22.41*	55.16±22.24	57.22±22.25*
R-Drop [18]	67.22±24.05*	70.37±23.58*	54.70±22.63*	57.39±22.94*
CPS [2]	67.71±22.50*	70.46±23.75	54.73±20.92*	58.67±23.30
HCE [7]	67.34±22.32*	70.29±22.62	54.58±20.37*	58.04±21.11
CNN&Transformer [8]	67.66±25.12	70.43±18.84*	55.74±23.38	57.89±19.48*
MC-Net+ [19]	67.81±24.22*	70.09±22.07*	55.40±22.54*	57.64±21.80*
Ours (CSA branch)	69.72±22.06	72.24±21.21	57.09±21.23	60.17±21.98
Full Supervision	77.47±12.49		64.97±14.09	

Comparison with State-of-the-Art Methods. Our CDMA was compared with eight existing SSL methods: 1) Entropy Minimization (EM) [15]; 2) Mean Teacher (MT) [14]; 3) Uncertainty-Aware Mean Teacher (UAMT) [23]; 4) R-Drop [18] that introduces a dropout-based consistency regularization between two networks; 5) CPS [2]; 6) Hierarchical Consistency Enforcement (HCE) [7]; 7) CNN&Transformer [8] that introduces cross-supervision between CNN and Transformer; 8) MC-Net+ [19] that imposes mutual consistency between multiple slightly different decoders. They were also compared with the lower bound of Supervised Learning (SL) that only learns from the labeled images. All these methods used the same backbone of DeepLabv3+ [1] for a fair comparison.

Quantitative evaluation of these methods is shown in Table 1. In the existing methods, MC-Net+ [19] and CPS [2] showed the best performance for both of the two annotation ratios. Our proposed CDMA achieved a better performance than all the existing methods, with a DSC score of 69.72% and 72.24% when the annotation ratio was 5% and 10%, respectively. Figure 2 shows a qualitative comparison between different methods. It can be observed that our CDMA yields less mis-segmentation compared with CPS [2] and MC-Net+ [19].

Table 2. Ablative analysis of our proposed method.

Methods	Mean DSC		Mean JI	
	5% labeled	10% labeled	5% labeled	10% labeled
MTNet (Baseline)	65.02±23.94	68.61±22.10	52.59±22.54	55.47±21.81
MTNet + \mathcal{L}_{cdkd} (argmax)	68.20±23.42	70.61±21.03	55.46±21.49	58.71±21.23
MTNet + \mathcal{L}_{cdkd} ($T=1$)	68.22±23.55	70.32±21.67	55.48±21.57	58.45±21.32
MTNet + \mathcal{L}_{cdkd}	68.84±22.89	71.49±20.74	55.92±21.44	59.02±21.13
MTNet + \mathcal{L}_{cdkd} + \mathcal{L}'_{um}	69.11±23.43	71.56±22.02	56.57±21.49	59.52±22.46
MTNet + \mathcal{L}_{cdkd} + \mathcal{L}_{um}	69.72±22.06	72.24±21.21	57.09±21.23	60.17±21.98
MTNet(dual) + \mathcal{L}_{cdkd} + \mathcal{L}_{um}	69.49±22.42	71.65±20.48	56.96±21.85	59.13±21.10
MTNet(csa×3) + \mathcal{L}_{cdkd} + \mathcal{L}_{um}	69.24±23.57	71.50±20.54	56.93±22.34	59.04±21.25
MTNet(-atten) + \mathcal{L}_{cdkd} + \mathcal{L}_{um}	68.92±23.42	71.37±20.68	56.03±22.13	58.81±21.46
MTNet(ensb) + \mathcal{L}_{cdkd} + \mathcal{L}_{um}	69.66±22.08	72.25±21.19	57.01±21.25	60.18±21.98

Ablation Study. For ablation study, we set the baseline as using the proposed MTNet with three different decoders for supervised learning from labeled images only. It obtained an average DSC of 65.02% and 68.61% under the two annotation ratios respectively. The proposed \mathcal{L}_{cdkd} was compared with two variants: \mathcal{L}_{cdkd} (argmax) and \mathcal{L}_{cdkd} ($T=1$) that represent using hard pseudo labels and standard probability output obtained by Softmax for CDKD respectively. Table 2 shows that our \mathcal{L}_{cdkd} obtained an average DSC of 68.84% and 71.49% under the two annotation ratios respectively, and it outperformed \mathcal{L}_{cdkd} (argmax) and \mathcal{L}_{cdkd} ($T=1$), demonstrating that our CDKD based on softened probability prediction is more effective in dealing with noisy pseudo labels. By introducing our average prediction-based uncertainty minimization \mathcal{L}_{um} , the DSC was further improved to 69.72% and 72.24% under the two annotation ratios respectively. In addition, replacing our \mathcal{L}_{um} by applying entropy minimization to each branch respectively (\mathcal{L}'_{um}) led to a DSC drop by around 0.65%.

Then, we compared different MTNet variants: 1) MTNet(dual) means a dual-branch structure (removing the CSA branch); 2) MTNet(csa×3) means all the three branches use CSA blocks; 3) MTNet(-atten) means no attention block is used in all the branches; and 4) MTNet(ensb) means using an ensemble of the three branches for inference. Note that all these variants were trained with \mathcal{L}_{cdkd} and \mathcal{L}_{um} . The results in the second section of Table 2 show that using the same structures for different branches, i.e., MTNet(-atten) and MTNet(csa×3), had a lower performance than using different attention blocks, and using three

attention branches outperformed just using two attention branches. It can also be found that using CSA branch for inference had a very close performance to MTNet(ensb), and it is more efficient than the later.

4 Conclusion

We have presented a novel semi-supervised framework based on Cross Distillation of Multiple Attentions (CDMA) for pathological image segmentation. It employs a Multi-attention Tri-branch network to generate diverse predictions based on channel attention, spatial attention, and simultaneous channel and spatial attention, respectively. Different attention-based decoder branches focus on various aspects of feature maps, resulting in disparate outputs, which is beneficial to semi-supervised learning. To eliminate the negative impact of incorrect pseudo labels in training, we employ a Cross Decoder Knowledge Distillation (CDKD) to enforce each branch to learn from soft labels generated by the other two branches. Experimental results on a colonoscopy tissue segmentation dataset demonstrated that our CDMA outperformed eight state-of-the-art SSL methods. In the future, it is of interest to apply our method to multi-class segmentation tasks and pathological images from different organs.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (62271115).

References

1. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV, pp. 801–818 (2018)
2. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: CVPR, pp. 2613–2622 (2021)
3. Da, Q., et al.: Digestpath: a benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. *Med. Image Anal.* **80**, 102485 (2022)
4. Fan, D.P., et al.: Inf-Net: automatic covid-19 lung infection segmentation from CT images. *IEEE Trans. Med. Imaging* **39**(8), 2626–2637 (2020)
5. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NeurIPS, pp. 1–10 (2015)
6. Hou, X., et al.: Dual adaptive pyramid network for cross-stain histopathology image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 101–109. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_12
7. Jin, Q., et al.: Semi-supervised histological image segmentation via hierarchical consistency enforcement. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13432, pp. 3–13. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-16434-7_1
8. Luo, X., Hu, M., Song, T., Wang, G., Zhang, S.: Semi-supervised medical image segmentation via cross teaching between CNN and transformer. In: MIDL, pp. 820–833. PMLR (2022)

9. Luo, X., et al.: Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Med. Image Anal.* **80**, 102517 (2022)
10. Müller, R., Kornblith, S., Hinton, G.E.: When does label smoothing help? In: *NeurIPS*, pp. 1–10 (2019)
11. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: *CVPR*, pp. 12674–12684 (2020)
12. Roy, A.G., Navab, N., Wachinger, C.: Recalibrating fully convolutional networks with spatial and channel “squeeze and excitation” blocks. *IEEE Trans. Med. Imaging* **38**(2), 540–549 (2019)
13. Shen, H., et al.: Deep active learning for breast cancer segmentation on immuno-histochemistry images. In: Martel, A.L., et al. (eds.) *MICCAI 2020*. LNCS, vol. 12265, pp. 509–518. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59722-1_49
14. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: *NeurIPS*, pp. 1–10 (2017)
15. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: adversarial entropy minimization for domain adaptation in semantic segmentation. In: *CVPR*, pp. 2517–2526 (2019)
16. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: convolutional block attention module. In: *ECCV*, pp. 3–19 (2018)
17. Wu, H., Wang, Z., Song, Y., Yang, L., Qin, J.: Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images. In: *CVPR*, pp. 11666–11675 (2022)
18. , Wu, L., et al.: R-drop: regularized dropout for neural networks. In: *NeurIPS*, pp. 10890–10905 (2021)
19. Wu, Y., et al.: Mutual consistency learning for semi-supervised medical image segmentation. *Med. Image Anal.* **81**, 102530 (2022)
20. Xie, Y., Lu, H., Zhang, J., Shen, C., Xia, Y.: Deep segmentation-emendation model for gland instance segmentation. In: Shen, D., et al. (eds.) *MICCAI 2019*. LNCS, vol. 11764, pp. 469–477. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_52
21. Xie, Y., Zhang, J., Liao, Z., Verjans, J., Shen, C., Xia, Y.: Pairwise relation learning for semi-supervised gland segmentation. In: Martel, A.L., et al. (eds.) *MICCAI 2020*. LNCS, vol. 12265, pp. 417–427. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59722-1_40
22. Xu, K., Rui, L., Li, Y., Gu, L.: Feature normalized knowledge distillation for image classification. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) *ECCV 2020*. LNCS, vol. 12370, pp. 664–680. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58595-2_40
23. 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
24. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: *CVPR*, pp. 11953–11962 (2022)