Deep Mutual Distillation for Semi-Supervised Medical Image Segmentation

Yushan Xie, Yuejia Yin, Qingli Li, and Yan Wang(✉)

Shanghai Key Laboratory of Multidimensional Information Processing,
East China Normal University, Shanghai 200241, China
{10192100433@stu,10182100267@stu,qlli@cs,ywang@cee}.ecnu.edu.cn

Abstract. In this paper, we focus on semi-supervised medical image segmentation. Consistency regularization methods such as initialization perturbation on two networks combined with entropy minimization are widely used to deal with the task. However, entropy minimization-based methods force networks to agree on all parts of the training data. For extremely ambiguous regions, which are common in medical images, such agreement may be meaningless and unreliable. To this end, we present a conceptually simple yet effective method, termed Deep Mutual Distillation (DMD), a high-entropy online mutual distillation process, which is more informative than a low-entropy sharpened process, leading to more accurate segmentation results on ambiguous regions, especially the outer branches. Furthermore, to handle the class imbalance and background noise problem, and learn a more reliable consistency between the two networks, we exploit the Dice loss to supervise the mutual distillation. Extensive comparisons with all state-of-the-art on LA and ACDC datasets show the superiority of our proposed DMD, reporting a significant improvement of up to 1.15% in terms of Dice score when only 10% of training data are labelled in LA. We compare DMD with other consistency-based methods with different entropy guidance to support our assumption. Extensive ablation studies on the chosen temperature and loss function further verify the effectiveness of our design. The code is publicly available at https://github.com/SilenceMonk/Dual-Mutual-Distillation.

Keywords: Semi-supervised learning · Segmentation · Knowledge distillation · Consistency regularization.

1 Introduction

Supervised learning for medical image segmentation requires a large amount of per-voxel annotated data [18, 8, 4, 20, 9]. Since both expertise and time are needed to produce accurate contouring annotations, the labelled data are very expensive to acquire, especially in 3D volumetric images [19] such as MRI. Semi-supervised medical image segmentation becomes an important topic in recent years, where costly per-voxel annotations are available for a subset of training

data. In this study, we focus on semi-supervised LA segmentation by exploring both labelled and unlabelled data.

Consistency regularization methods are widely studied in semi-supervised segmentation models. Consistent predictions are enforced by perturbing input images [12, 24], learned features [15], and networks [25, 3, 21, 23]. Other consistency-based methods adopt adversarial losses to learn consistent geometric representations in the dataset [10, 27], enforcing local and global structural consistency [6], and building task-level regularization [13]. Among these methods, initialization perturbation [3] combined with entropy minimization [5] demonstrates outstanding performances. These methods [3, 23] require two segmentation networks/streams with different initialization to be consistent between the two predictions by pseudo labeling/sharpening from the other network/stream.

However, entropy minimization [5] based methods [3, 23] give up a great amount of information contained in network predictions, forcing networks to agree with each other even in ambiguous regions. But such cross guidance on ambiguous regions may be meaningless and unreliable [25]. More concretely, many parts of the target in medical images can be extremely confusing. *E.g.*, some boundaries like the outer branches, can even confuse radiologists. In this case, it may be difficult to train two reliable classifiers to simultaneously distinguish the confusing foreground from the background by entropy minimization-based methods. This is because the penalties for misclassifications on the confusing region and the confident region are equal. Meanwhile, it also makes networks inevitably plagued with confirmation bias [2]. In the early optimization stage, the pseudo labels are not stable. Thus, as the training process goes on, the two segmentation networks are prone to overfit the erroneous pseudo labels.

Motivated by Knowledge Distillation (KD) [7], we propose **Deep Mutual Distillation** (DMD), advocating to generalize the original Deep Mutual Learning (DML) [26] by introducing temperature scaling, and reformulate a symmetric online mutual distillation process to combat the clear drawback in entropy minimization [5] under medical image tasks. With the temperature scaling, the high-entropy distilled probabilities are more informative than low-entropy sharpened probabilities, therefore offering more meaningful mutual guidance, especially on ambiguous regions. Furthermore, due to the class imbalance problem in medical images, *i.e.*, targets are usually very small compared with the whole volume, we exploit the Dice loss [1] as the consistency regularization to supervise the mutual distillation of two networks. To the best of our knowledge, KD [7] is overlooked in the semi-supervised medical image segmentation field. Our DMD is conceptually simple yet computationally efficient. Experiments on MICCAI 2018 Atrial Segmentation Challenge and ACDC datasets show that DMD works favorably especially when annotated data is very small. Without bells and whistles, DMD achieves 89.70% in terms of Dice score on LA when only 10% training data are labelled, with a significant **1.15%** improvement compared with state-of-the-arts.

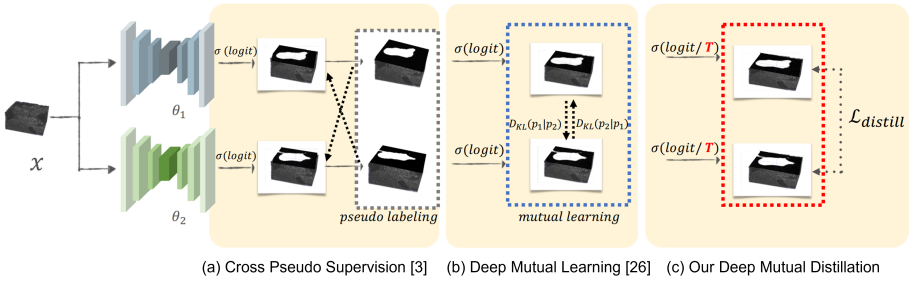

Fig. 1. Visualizations of some consistency-based methods with different entropy guidance. From (a) to (c), the entropy of network guidance increases, enabling the networks to learn an increasing amount of information from the data x. σ means sigmoid.

2 Method

2.1 Overview

As shown in Fig. 1, we illustrate some consistency-based methods with different entropy guidance. From Fig. 1 (a) to (c), the entropy used as guidance for network learning is increasing. In Cross Pseudo Supervision (CPS) [3], the hard pseudo segmentation map is used as guidance to supervise the other segmentation network. In DML [26], a two-way KL mimicry loss is applied directly to the probability distribution learned by the softmax layer. In our proposed DMD, we generalize DML [26] by introducing a temperature scaling strategy and further increasing the entropy of the probability distribution. Considering the class imbalance between foreground and background pixels under medical image segmentation tasks [17], we design a Dice [1]-based distillation loss.

2.2 Deep Mutual Learning

The original DML [26] deals with the standard M-class classification problem. Given two initialization perturbed networks $f_{\theta_j}, j \in \{1, 2\}$, we obtain their raw logit predictions on the same sample point $x_i \in \mathcal{X}$ in parallel as $z_j^m = f_{\theta_j}(x_i)$ for class m. The probability of class m from f_{θ_j} is given by standard softmax function:

$$p_j^m = \frac{exp(z_j^m)}{\sum_{m=1}^{M} exp(z_j^m)} \tag{1}$$

The critical part of mutual learning contains a 2-way KL mimicry loss:

$$\mathcal{L}_{ml} = D_{KL}(p_2||p_1) + D_{KL}(p_1||p_2) \tag{2}$$

where $\mathcal{L}_{ml}$ is obtained on both labelled and unlabelled sample points, p_1, p_2 being posterior probability predictions of corresponding networks. Together with

standard supervised loss obtained on labelled sample points, and the trade-off weight λ, we get the final DML [26] objective:

$$\mathcal{L}_{DML} = \mathcal{L}_{sup} + \lambda \cdot \mathcal{L}_{ml} \tag{3}$$

where λ is set to 1 in DML [26].

2.3 Deep Mutual Distillation

The original p_j^m of DML [26] can be considered as a special case of online knowledge distillation [7] with temperature T set to 1, where each network serves as both teacher and student symmetrically. However, $T = 1$ makes a great amount of information from both networks still masked within p_j^m. Therefore, we generalize DML [26] to **Deep Mutual Distillation**(DMD) by setting T greater than 1 as in KD [7]. In DMD, the distilled probability $p_{j,T}^m$ is obtained by:

$$p_{j,T}^m = \frac{exp(z_j^m/T)}{\sum_{m=1}^{M} exp(z_j^m/T)} \tag{4}$$

In the case of the binary segmentation task, we replace softmax with sigmoid to get the distilled per-pixel probability mask $p_{j,T}$ from f_{θ_j}:

$$p_{j,T} = \frac{1}{1 + exp(-z_j/T)} \tag{5}$$

However, using KL-divergence-based loss in KD [7] cannot handle class imbalance between foreground and background pixels [17]. Hence, we replace the original 2-way KL-divergence mimicry loss with Dice loss [1] to alleviate this problem, obtaining our new distillation loss:

$$\mathcal{L}_{distill} = Dice(p_{1,T}, p_{2,T}) \tag{6}$$

Together with standard supervised loss obtained on labelled sample points, and the trade-off weight λ, we get our final DMD objective:

$$\mathcal{L}_{DMD} = \mathcal{L}_{sup} + \lambda \cdot \mathcal{L}_{distill} \tag{7}$$

where $\mathcal{L}_{distill}$ is obtained on both labelled and unlabelled sample points. Here, we also adopt Dice loss [1] for $\mathcal{L}_{sup}$, under the context of highly class imbalanced medical image segmentation tasks [1].

With the temperature scaling, each network under DMD learns from each other through the distilled high-entropy probabilities, which are more informative, especially on ambiguous regions. The distillation [7] also makes $p_{j,T}$ become soft labels [7], which reduces the influence of confirmation bias [2] throughout the training process. Both advantages make DMD outperforms current state-of-the-art methods. We also carry out comprehensive ablation studies to demonstrate the effectiveness of DMD design in section 3.3.

Table 1. Comparisons with previous state-of-the-art methods on the LA dataset. "↑" and "↓" indicate the larger and the smaller the better, respectively. **Bold** denotes the best results.

Method	#Scans used		Metrics			
	labelled	Unlabelled	Dice(%)↑	Jaccard(%)↑	95HD(voxel)↓	ASD(voxel)↓
V-Net	8(10%)	0	79.99	68.12	21.11	5.48
V-Net	16(20%)	0	86.03	76.06	14.26	3.51
V-Net	80(All)	0	91.14	83.82	5.75	1.52
DAP [27]	8(10%)	72	81.89	71.23	15.81	3.80
UA-MT [25]	8(10%)	72	84.25	73.48	13.84	3.36
SASSNet [11]	8(10%)	72	87.32	77.72	9.62	2.55
LG-ER-MT [6]	8(10%)	72	85.54	75.12	13.29	3.77
DUWM [21]	8(10%)	72	85.91	75.75	12.67	3.31
DTC [13]	8(10%)	72	86.57	76.55	14.47	3.74
MC-Net [23]	8(10%)	72	87.71	78.31	9.36	2.18
SS-Net [22]	8(10%)	72	88.55	79.62	7.49	1.90
DMD (Ours)	8(10%)	72	**89.70**	**81.42**	**6.88**	**1.78**
DAP [27]	16(20%)	64	87.89	78.72	9.29	2.74
UA-MT [25]	16(20%)	64	88.88	80.21	7.32	2.26
SASSNet [11]	16(20%)	64	89.54	81.24	8.24	2.20
LG-ER-MT [6]	16(20%)	64	89.62	81.31	7.16	2.06
DUWM [21]	16(20%)	64	89.65	81.35	7.04	2.03
DTC [13]	16(20%)	64	89.42	80.98	7.32	2.10
MC-Net [23]	16(20%)	64	90.34	82.48	**6.00**	1.77
DMD (Ours)	16(20%)	64	**90.46**	**82.66**	6.39	**1.62**

3 Experiments and Results

3.1 Experimental Setup

Dataset: We evaluated our proposed DMD on the 2018 Atria Segmentation Challenge (LA)[1], which provides a 80/20 split for training/validation on 3D MR imaging scans and corresponding LA segmentation mask, with an isotropic resolution of $0.625 \times 0.625 \times 0.625 \text{mm}^3$. We also extended our experiments on the Automated Cardiac Diagnosis Challenge (ACDC)[2]. We report the performance on the validation set, following the same settings from previous methods [27, 25, 10, 6, 21, 23, 22] for fair comparisons.

Evaluation Metric: The performance of our method is quantitatively evaluated in terms of Dice, Jaccard, the average surface distance (ASD), and the 95% Hausdorff Distance (95HD) as previous methods [27, 25, 10, 6, 21, 23, 22].

Implementation Details: We implement DMD using PyTorch [16]. We adopt VNet [1] as the backbone for both of the segmentation networks. We first randomly initialize two networks, then we train both networks under the scheme

[1] https://www.cardiacatlas.org/atriaseg2018-challenge/
[2] https://www.creatis.insa-lyon.fr/Challenge/acdc/#phase/5966175c6a3c770dff4cc4fb

Table 2. Comparisons on the ACDC dataset under the settings of [22]. "↑" and "↓" indicate the larger and the smaller the better, respectively.

Method	#Scans used		Metrics			
	labelled	Unlabelled	Dice(%)↑	Jaccard(%)↑	95HD(voxel)↓	ASD(voxel)↓
U-Net	3(5%)	0	47.83	37.01	31.16	12.62
U-Net	7(10%)	0	79.41	68.11	9.35	2.70
U-Net	70(All)	0	91.44	84.59	4.30	0.99
UA-MT [25]	3(5%)	67	46.04	35.97	20.08	7.75
SASSNet [11]	3(5%)	67	57.77	46.14	20.05	6.06
DTC [13]	3(5%)	67	56.90	45.67	23.36	7.39
URPC [14]	3(5%)	67	55.87	44.64	13.60	3.74
MC-Net [23]	3(5%)	67	62.85	52.29	7.62	2.33
SS-Net [22]	3(5%)	67	65.82	55.38	**6.67**	**2.28**
DMD (Ours)	3(5%)	67	**66.23**	**55.84**	8.66	2.40

of DMD using SGD optimizer for 6k iterations simultaneously, with an initial learning rate (LR) 0.01 decayed by 0.1 every 2.5k iterations following [23]. Other data pre-processing and augmentation details are kept the same as [23]. For other hyper-parameters in DMD, we set the trade-off weight λ to 4, and the temperature T to 2/1.93 for the 8/16 label scenario for the best performance. It is interesting to observe that with less labelled data, T is prone to be set to a bigger value since less labelled training data will usually lead to more ambiguous regions. After training, we only use one network for generating results for evaluation, without using any ensembling methods.

3.2 Quantitative Evaluation

We compare DMD with previous state-of-the-arts [27, 25, 10, 6, 21, 23, 22], following the measurements from MC-Net [23]. Table 1 shows that our method outperforms state-of-the-art methods with a significant improvement over 8 label scenarios under all 4 metrics, and achieves state-of-the-art on the 16 label scenario under almost all metrics on LA dataset. We do not compare the performance of SS-Net [22] on 16 labels as SS-Net [22] does not report this. We can see that even with an extremely small amount of labelled samples, networks in DMD are still able to formulate a certain representation of the unlabelled data and transfer such meaningful knowledge via high-entropy probabilities with each other by the efficient distilling process. Distillation is more informative and greatly benefits training on complex medical images with confusing regions. We further extended our experiments on the ACDC dataset shown in Table 2.

To study how consistency-based methods with different entropy guidance affect performances, we implement CPS [3] and DML [26] for LA segmentation. From Table 3, we can see general improvements from low-entropy methods to high-entropy methods from CPS [3] to our proposed DMD (from top to bottom in the first column in Table 3). In these entropy minimization methods, *i.e.*, CPS [3] and MC-Net [23], networks are forced to assign sharpened labels on all parts

Table 3. Comparisons with consistency-based methods with different entropy guidance on the LA dataset. "↑" and "↓" indicate the larger and the smaller the better, respectively.

Method	Entropy control	#Scans used		Metrics	
		labelled	Unlabelled	Dice(%)↑	Jaccard(%)↑
CPS [3]	pseudo-labeling	8(10%)	72	87.49	78.06
MC-Net [23]	sharpening	8(10%)	72	87.71	78.31
DML [26]	N/A	8(10%)	72	88.19	78.92
DMD (Ours)	distillation [7]	8(10%)	72	**89.70**	**81.42**

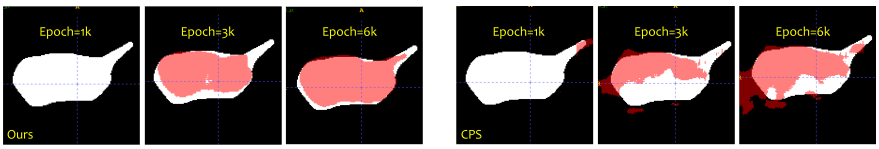

Fig. 2. Entropy-minimization methods like CPS [3] do worse in refining pseudo labels throughout the training process compared to our method. Red mask: pseudo label; White background: ground truth.

of unlabelled samples, including ambiguous areas. Thus, networks are forced to be exposed to the risk of confirmation bias [2] of each other, which limits their performances. In section 3.3, we further study how the temperature T affects DMD performances, where T controls the entropy in DMD guidance.

Furthermore, we show in Fig. 2 our method can lead to better pseudo-labels when the training process is going on, compared with entropy-minimization methods like CPS. Besides, we provide the gradient visualization for $\mathcal{L}_{distill}$ on an unlabelled sample point in Fig. 3(a). We can see that when using Dice [1] for distillation, the gradient is enhanced more on the foreground, especially on the boundary of the object predicted by the segmentation network than 2-way KL-divergence. We can also see that using Dice [1], the segmentation network better captures the shape of the object, thus providing better guidance for the other network than using 2-way KL-divergence.

3.3 Parameter Analysis

Here, we first demonstrate the effectiveness of temperature scaling T and the choice of KL-divergence-based and Dice [1]-based $\mathcal{L}_{distill}$ on LA with 8 labelled data. In order to do so, we conduct independent experiments to study the influence of T for each choice of $\mathcal{L}_{distill}$. For a fair comparison, we choose different trade-off weights λ in Eq. 7 for each choice to get the corresponding best performance, denoted as λ_{dice} and λ_{KL}, where we set $\lambda_{dice} = 1$ and $\lambda_{KL} = 4$. Fig. 4 shows how T affects DMD performances on the validation set, with corresponding fixed λ_{dice} and λ_{KL}. Experiments show that slightly higher T improves

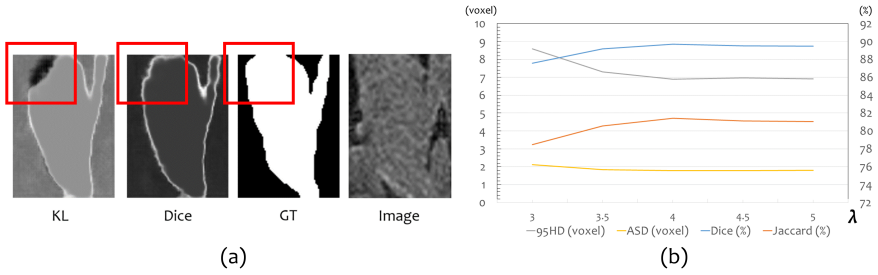

Fig. 3. (a) Gradient visualization (highlighted in outer branch) shows the difference between the choice of KL-divergence-based and Dice [1]-based $\mathcal{L}_{distill}$ on an unlabelled sample point during training. (b) Ablation study of all evaluation metrics on trade-off weight λ. Here, we set $T = 2$, and choose Dice loss [1] for $\mathcal{L}_{distill}$.

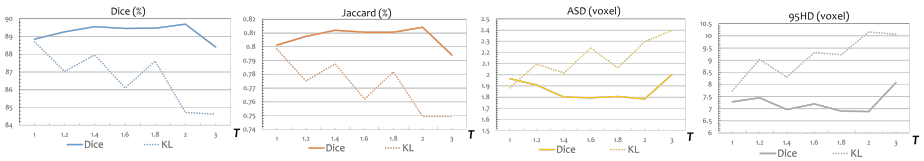

Fig. 4. Ablation study of all evaluation metrics on temperature T and the choice of $\mathcal{L}_{distill}$. The solid/dotted line denotes $\mathcal{L}_{distill}$ using Dice [1]/2-way KL-divergence. Here, we set $\lambda_{dice} = 4$ for all Dice loss [1] experiments, and $\lambda_{KL} = 1$ for all KL-divergence loss experiments. Experiments are conducted under 8 available labels for demonstration. Note that when $\lambda_{KL} = 1$ and $T = 1$, DMD degenerates to DML [26].

over the performance, and we can also see that Dice loss [1] outperforms 2-way KL-divergence loss under various T. It is interesting to observe that when T increases, the performance decreases by using the KL-divergence loss. We suspect that due to the complex background context and class imbalance problem, the learning of two networks is heavily influenced by the background noise. We also provide an ablation study on the influence of trade-off weight λ in Fig. 3(b), where we set $T = 2$, and choose Dice for $\mathcal{L}_{distill}$. Then, to see how the performance changes *w.r.t.* λ, we vary λ and fix $T = 2$. As shown in Fig. 3(a), the performance is not sensitive within the range of $\lambda \in [3.5, 5]$.

4 Conclusions

We revisit Knowledge Distillation and have presented a novel and simple semi-supervised medical segmentation method through Deep Mutual Distillation. We rethink and analyze consistency regularization-based methods with the entropy minimization, and point out that cross guidance with low entropy on extremely

ambiguous regions may be unreliable. We hereby propose to introduce a temperature scaling strategy into the network training and propose a Dice-based distillation loss to alleviate the influence of the background noise when the temperature $T > 1$. Our DMD works favorably for semi-supervised medical image segmentation, especially when the number of training data is small (*e.g.*, 10% training data are labelled in LA). Compared with all prior arts, a significant improvement up to 1.15% in the Dice score is achieved in LA dataset. Ablation studies with the consistency-based methods of different entropy guidance further verify our assumption and design.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 62101191), Shanghai Natural Science Foundation (Grant No. 21ZR1420800), and the Science and Technology Commission of Shanghai Municipality (Grant No. 22DZ2229004).

References

1. Abdollahi, A., Pradhan, B., Alamri, A.: Vnet: An end-to-end fully convolutional neural network for road extraction from high-resolution remote sensing data. IEEE Access **8**, 179424–179436 (2020)
2. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: Proc. IJCNN (2020)
3. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proc. CVPR (2021)
4. Dou, Q., Chen, H., Jin, Y., Yu, L., Qin, J., Heng, P.A.: 3d deeply supervised network for automatic liver segmentation from ct volumes. In: Proc. MICCAI (2016)
5. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. Advances in neural information processing systems **17** (2004)
6. Hang, W., Feng, W., Liang, S., Yu, L., Wang, Q., Choi, K., Qin, J.: Local and global structure-aware entropy regularized mean teacher model for 3d left atrium segmentation. In: Proc. MICCAI (2020)
7. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 **2**(7) (2015)
8. Imran, A.A.Z., Hatamizadeh, A., Ananth, S.P., Ding, X., Tajbakhsh, N., Terzopoulos, D.: Fast and automatic segmentation of pulmonary lobes from chest ct using a progressive dense v-network. Computer Methods in Biomechanics and Biomedical Engineering (2019)
9. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al.: nnu-net: Self-adapting framework for u-net-based medical image segmentation. Nature Methods (2021)
10. Li, S., Zhang, C., He, X.: Shape-aware semi-supervised 3d semantic segmentation for medical images. In: Proc. MICCAI (2020)
11. Li, S., Zhang, C., He, X.: Shape-aware semi-supervised 3d semantic segmentation for medical images. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskow-

icz, L. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part I. Lecture Notes in Computer Science, vol. 12261, pp. 552–561. Springer (2020). https://doi.org/10.1007/978-3-030-59710-8_54, https://doi.org/10.1007/978-3-030-59710-8_54
12. Li, X., Yu, L., Chen, H., Fu, C., Heng, P.: Transformation consistent self-ensembling model for semi-supervised medical image segmentation. IEEE Trans. Neural Networks and Learning Systems (2020)
13. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: Proc. AAAI (2021)
14. Luo, X., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Chen, N., Wang, G., Zhang, S.: Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In: Proc. MICCAI (2021)
15. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: Proc. CVPR (2020)
16. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
17. Rajput, V.: Robustness of different loss functions and their impact on networks learning capability. arXiv preprint arXiv:2110.08322 (2021)
18. Roth, H.R., et al.: Deeporgan: multi-level deep convolutional networks for automated pancreas segmentation. In: Proc. MICCAI (2015)
19. Wang, Y., Tang, P., Zhou, Y., Shen, W., Fishman, E.K., Yuille, A.L.: Learning inductive attention guidance for partially supervised pancreatic ductal adenocarcinoma prediction. IEEE Trans. Medical Imaging **40**(10), 2723–2735 (2021)
20. Wang, Y., Wei, X., Liu, F., Chen, J., Zhou, Y., Shen, W., Fishman, E.K., Yuille, A.L.: Deep distance transform for tubular structure segmentation in CT scans. In: Proc. CVPR (2020)
21. Wang, Y., Zhang, Y., Tian, J., Zhong, C., Shi, Z., Zhang, Y., He, Z.: Double-uncertainty weighted method for semi-supervised learning. In: Proc. MICCAI (2020)
22. Wu, Y., Wu, Z., Wu, Q., Ge, Z., Cai, J.: Exploring smoothness and class-separation for semi-supervised medical image segmentation. In: Proc. MICCAI (2022)
23. Wu, Y., Xu, M., Ge, Z., Cai, J., Zhang, L.: Semi-supervised left atrium segmentation with mutual consistency training. In: Proc. MICCAI (2021)
24. Xia, Y., Yang, D., Yu, Z., Liu, F., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A.L., Roth, H.: Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. Medical Image Anal. **65**, 101766 (2020)
25. Yu, L., Wang, S., Li, X., Fu, C., Heng, P.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: Proc. MICCAI (2019)
26. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4320–4328 (2018)
27. Zheng, H., Lin, L., Hu, H., Zhang, Q., Chen, Q., Iwamoto, Y., Han, X., Chen, Y., Tong, R., Wu, J.: Semi-supervised segmentation of liver using adversarial learning with deep atlas prior. In: Proc. MICCAI (2019)