Alternate Diverse Teaching for Semi-supervised Medical Image Segmentation

Zhen Zhao[1], Zicheng Wang[3], Longyue Wang[2⋆]
Dian Yu[2], Yixuan Yuan[4], and Luping Zhou[3]

[1] Shanghai AI Lab [2] Tencent AI Lab
[3] University of Sydney [4] The Chinese University of Hong Kong
https://github.com/zhenzhao/AD-MT

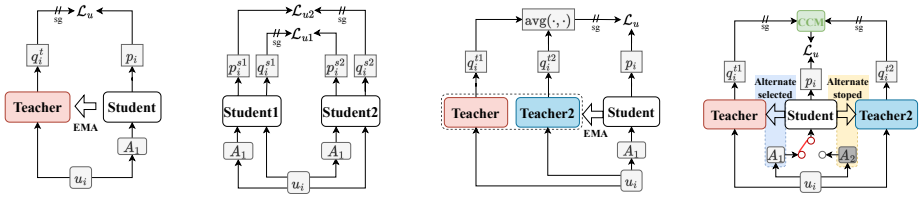
Abstract. Semi-supervised medical image segmentation has shown promise in training models with limited labeled data. However, current dominant teacher-student based approaches can suffer from the confirmation bias. To address this challenge, we propose AD-MT, an alternate diverse teaching approach in a teacher-student framework. It involves a single student model and two non-trainable teacher models that are momentum-updated periodically and randomly in an alternate fashion. To mitigate the confirmation bias via the diverse supervision, the core of AD-MT lies in two proposed modules: the Random Periodic Alternate (RPA) Updating Module and the Conflict-Combating Module (CCM). The RPA schedules an alternating diverse updating process with complementary unlabeled data batches, distinct data augmentation, and random switching periods to encourage diverse reasoning from different teaching perspectives. The CCM employs an entropy-based ensembling strategy to encourage the model to learn from both the consistent and conflicting predictions between the teachers. Experimental results demonstrate the effectiveness and superiority of AD-MT on the 2D and 3D medical segmentation benchmarks across various semi-supervised settings.

Keywords: Semi-supervised Learning · Medical Image Segmentation · Alternate Diverse Teaching · Random Periodic Alternate

1 Introduction

Medical image segmentation is a pressing task in computer-aided diagnosis, as it plays a crucial role in medical image reasoning [4,27,36]. Current methods for medical image segmentation heavily rely on deep neural networks, which necessitates a substantial amount of annotated data for training. However, annotating pixel-wise medical images is challenging and time-consuming, often requiring expert annotators [33,39]. In response to this challenge, several studies have focused on the development of semi-supervised medical image segmentation (SSMIS) techniques, aiming to train models using a limited amount of labeled data

⋆ Corresponding author. The authors also thank Hong Kong Research Grants Council (RGC) General Research Fund 14204321 for the support of this work.



(a) mean-teacher **(b)** two-student co-training **(c)** two-teacher ensembling **(d)** AD-MT

Fig. 1: Frameworks of different SSMIS methods. Crucial distinctions arise from how the unlabeled data is leveraged. a) the plain teacher-student framework, b) a two-student co-training paradigm that enforces mutual learning, c) a two-teacher ensemble framework where two differently initialized and updated teacher models supervise the training of the student model in a average manner, d) our proposed alternate diverse mean-teacher (**AD-MT**) framework. Our framework involves two teacher models that are updated periodically and randomly, using complementary unlabeled data batches, distinct data augmentation strategies, and randomized switchable periods, to enlarge their disagreement. Additionally, our Conflict-Combating Module (CCM) encourages the student model to learn from the conflict predictions of the teacher models rather than dropping conflicts directly. "sg" denotes "stop gradient".

and a larger amount of unlabeled data [2, 17, 31]. It is evident that the key lies in effectively leveraging the unlabeled data to assist the labeled data for model training [10, 13, 15].

Recent SSMIS studies are dominated by consistency regularization (CR) based approaches, which encourage the model to generate consistent predictions from disagreements on the same unlabeled input [12, 32, 40]. Various works adopt a teacher-student framework with weak and strong data augmentation strategies and encourage the exponential moving average (EMA) teacher model to provide supervision for unlabeled training. However, as discussed in many semi-supervised studies [2, 18, 29], a single model can inevitably produce noisy and even wrong pseudo-labels, resulting in the model suffering from the so-called confirmation bias issue [1].

In the literature, early studies like DTC [17], SASS-Net [12], and SS-Net [31], tend to introduce extra training constraints to tackle the confirmation bias in an indirect manner. Differently, recent works turn to increasing diverse supervision signals to alleviate the bias directly. Studies along this line can be divided into two categories, *i.e.*, the multi-student co-training method [5,18,32] and the multi-teacher ensembling method [15]. The co-training framework aims at introducing another diverse student model and encouraging both models to supervise each other mutually, as shown in Figure 1b. However, in addition to introducing extra training efforts and losing the EMA stability, these methods using the same network structure cannot produce a sufficient discrepancy between co-training models by only using different initialization and learning rates. On the other hand, the multi-teacher ensembling methods encourage the student model to update different teacher models iteratively, avoiding extra training costs, as

shown in Figure 1c. However, a key issue in such methods is the teacher model updating strategy, which has not been carefully designed to generate diverse supervision. Besides, existing ensembling methods typically adopt an average strategy and train the model only from their consistent predictions. Few studies have explored the potential benefits of learning from the conflicts.

Facing these issues, in this paper, we propose a novel alternate diverse teaching approach in a teacher-student framework, dubbed AD-MT, for SSMIS. As shown in Figure 1d, AD-MT involves a single trainable student model and two non-trainable teacher models that are updated directly by the EMA of the student weights. In specific, two teachers are updated periodically and randomly in an alternate fashion. To encourage diverse reasoning from different teaching perspectives, AD-MT enlarges the discrepancy by using complementary data batches, distinct data augmentation strategies, and randomized switchable periods. This alternating diverse updating process is scheduled by our proposed Random Periodic Alternate Updating (RPA) Module. Furthermore, instead of discarding conflicting predictions of the teacher models, we design an entropy-based Conflict-Combating Module (CCM) that separates the consistent and conflicting predictions and also encourages the model to learn from the disagreements between the teachers. Thanks to our proposed two modules, AD-MT, as shown in Figure 2, can consistently outperform current state-of-the-art (SOTA) BCP by a large margin, achieving a 6.38% Dice improvement on the Pancrease with 10% labeled data. Our contributions are summarized as follows,

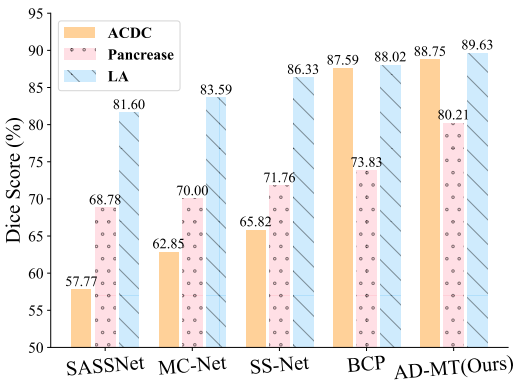


Fig. 2: We compare our proposed AD-MT with recent SSIMS methods in terms of the Dice score on 2D ACDC, 3D LA and Pancreas datasets with 3, 4, and 6 labeled instances, respectively. Our end-to-end AD-MT can consistently outperform the current state-of-the-art BCP (*which requires an additional pre-training stage*).

- We propose AD-MT, an alternate diverse teacher-student approach for SSMIS, which enforces diverse teaching models to mitigate the impact of confirmation bias.
- We design a novel Random Periodic Alternate updating module to enlarge the diversity of the two teacher models and a Conflict-combating module to learn from both consistent and conflicting predictions of the two teachers.
- Without introducing additional constraints and extra training costs, AD-MT achieves the new SOTA performance on the 2D and 3D SSMIS benchmarks.

2 Related work

As highlighted by various works [7,9,25,44–46], the effective utilization of unlabeled data plays a crucial role in tackling the semi-supervised problem. Among the existing studies, consistency regularization (CR) based approaches have emerged as the dominant direction in SSMIS. These methods aim to produce disagreements on the same unlabeled inputs, thereby training the model to generate consistent predictions [14,15,37,38,41,43]. Previous methods along this line have focused on introducing perturbations and generating pseudo-labels using stable predictions as supervision for unstable ones [21,38,42]. Mean-teacher [26], as shown in Figure 1a, is a widely adopted semi-supervised learning framework in SSMIS studies. Latter studies explored the significance of weak and strong augmentation strategies to produce sufficient prediction disagreement [2,25,34,43]. However, the supervision signal derived from the predictions of the unlabeled data is inherently noisy, which can lead to an issue known as confirmation bias [1]. The bias issue can negatively impact the training stability and hinder the model's recognition ability. On top of Mean-teacher, UAMT [40] proposed an uncertainty-aware training scheme to alleviate the bias. SASS-Net [12] posed a geometric shape constraint upon the segmentation outputs, and SSNet [31] designed additional contrastive losses to enhance the model's discriminative ability, which both tended to improve the SSMIS in an indirect manner.

Differently, recent works propose increasing the supervision signals to mitigate the potential bias and improve the training process directly. These approaches can be broadly categorized into two main categories: the multi-student co-training methods and the multi-teacher ensembling methods. As illustrated in Figure 1b, two-student co-training methods [18,29,32] tackle the noisy supervision problem by introducing an additional branch to the learning framework. The incorporation of an extra branch involves another student model, which can mutually supervise each other. Such an approach encourages diverse reasoning from different perspectives, thereby mitigating the impact of confirmation bias. Existing studies along this line typically adopt the different model initialization and learning rates to maintain the discrepancy between student models [5,16]. Differently, our proposed AD-MT enlarges the discrepancy by using complementary sets of unlabeled data batches, distinct data augmentation strategies, and randomized switchable periods while reserving the benefits of exponential moving averaging teacher models. Besides, these methods also come at the cost of increased training costs, as they introduce additional training parameters.

On the other hand, some works adopt the multi-teacher ensembling framework [15,22], as shown in Figure 1c. Along this line, the student model iteratively updates multiple teacher models with different updating strategies, leveraging their differing perspectives. By utilizing multi-teacher models, these methods ensure supervision diversity without introducing additional training parameters. However, it is worth noting that existing methods have not carefully considered the updating strategy to enforce the teacher models to be sufficiently different. For example, some works only utilize different initialization or updating at different epochs to encourage teacher differences. It is important to develop more

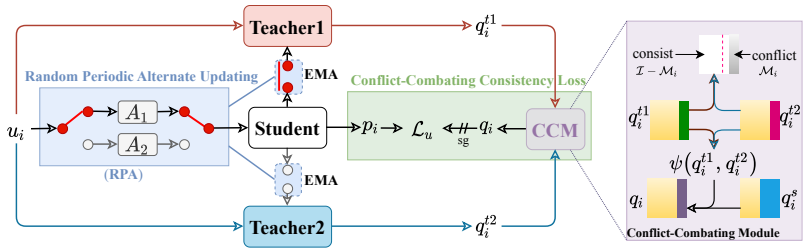


Fig. 3: The diagram of our proposed AD-MT. Our method consists of two main modules: the Random Periodic Alternate Updating Module (RPA) and the Conflict-combating Module (CCM). Specifically, two teacher models T1 and T2 are updated in turn periodically and randomly. At each iteration, only one certain teacher model T_m, $(m = 1, 2)$ will be updated, using complementary unlabeled data batches and different strong data augmentation strategies A_m accordingly. Furthermore, the switchable period of two teachers is randomly generated by the RPA module, aiming to increase the disagreement between the two teacher models. Meanwhile, the CCM module separates the consistent and conflicting predictions of two teacher models, and encourage the model to learning from instead of dropping the conflicts. $q_i^s, q_i^{t1}, q_i^{t2}$ represent the generated pseudo-labels from the student and two teachers models, respectively.

effective and robust updating strategies that can maintain the diversity of the teacher models while ensuring training stability. More importantly, it is also crucial to address the issue of conflicting supervision when tackling the ensembled predictions. Conflicting supervision naturally occurs when different sources of supervision provide contradictory guidance to the student model, potentially leading to training instability and suboptimal segmentation results. Despite the significance, most of the existing methods tend to drop these conflicts but have not explicitly tackled the issue of conflicting supervision.

3 Method

In this section, we provide an overview of our AD-MT, followed by a detailed description of its two main components: the Random Periodic Alternate Updating Module (RPA) and the Conflict-Combating Module (CCM).

3.1 Overview

In the context of semi-supervised medical image segmentation (SSMIS), the available data comprises both labeled samples $\mathcal{X}$ and unlabeled samples $\mathcal{U}$, with the number of labeled samples typically being much smaller than that of unlabeled ones (*i.e.*, $|X| \ll |U|$). During the training process, given a batch of labeled samples $\mathcal{B}_x = \{(x_i, y_i)\}_{i=1}^{B}$ and a batch of unlabeled samples $\mathcal{B}_u = \{u_i\}_{i=1}^{\mu B}$, SSMIS methods aim to obtain a good segmentation model by leveraging both labeled and unlabeled data effectively. Different from widely-adopted co-training

methods, our proposed method, as illustrated in Figure 3, utilizes a single student model, parameterized by θ_s, that receives model back-propagating gradient. Besides, two auxiliary teacher models, parameterized by θ_{t_1} and θ_{t_2}, are alternatively updated by the exponential moving average (EMA) of the student model weights. Similar to the plain teacher-student methods [26], the student model can be directly trained on the labeled data via a standard supervised loss $\mathcal{L}_x$,

$$\mathcal{L}_x = \frac{1}{|\mathcal{B}_x|} \sum_{i=1}^{B} \frac{1}{H \times W} \sum_{j=1}^{H \times W} \ell(\hat{y}_i(j), y_i(j)), \tag{1}$$

where $\hat{y}_i$ denotes the student model's prediction on the i-th labeled data x_i, *i.e.*, $\hat{y}_i = f(x_i; \theta_s)$, and j represents the j-th pixel on the image or the corresponding segmentation mask with a resolution of $H \times W$. Following [2, 17], ℓ represents the loss function, calculated by an average of the dice and cross-entropy loss.

Though Teacher1 (T1) and Teacher2 (T2) are updated in turn, both models are involved in generating pseudo-labels for unlabeled data at each iteration simultaneously. Using $a(\cdot)$ denote the weak augmentations, which include random cropping and flipping operations [31, 40], we can obtain pseudo-labels for each unlabeled instance u_i, *i.e.*, $q_i^{t_1} = f(a(u_i); \theta_{t_1})$, $q_i^{t_2} = f(a(u_i); \theta_{t_2})$, and $q_i^s = f(a(u_i); \theta_s)$ from two teacher models as well as the student model, respectively. Our proposed AD-MT can then exploit all this information and obtain an ultimate pseudo-label q_i,

$$q_i = \phi(q_i^{t_1}, q_i^{t_2}, q_i^s, \tau), \tag{2}$$

where $\phi(\cdot)$ represents a function of our proposed **Conflict-Combating Module**, and τ denotes a pre-defined high-confidence threshold. Subsequently, our method employs a conflict-combating consistency loss, $\mathcal{L}_u$, on unlabeled data,

$$\mathcal{L}_u = \frac{1}{|\mathcal{B}_u|} \sum_{i=1}^{\mu B} \frac{1}{H \times W} \sum_{j=1}^{H \times W} \ell(p_i(j), q_i(j)), \tag{3}$$

where $p_i = f(A_m(u_i); \theta_s)$ is the student model's prediction on strongly-augmented unlabeled data $A_m(u_i)$. $A_m \in \{A_1, A_2\}, m = 1, 2.$ represents two different strong data augmentation strategies, corresponding to the alternate turn of T1 and T2, respectively. The whole updating strategy is performed by our **Random Periodic Alternate Updating Module**. In summary, the total training loss is,

$$\mathcal{L} = \mathcal{L}_x + \lambda_t \mathcal{L}_u, \tag{4}$$

where λ_t denotes an iteration-dependent function to adjust the importance of consistency loss $\mathcal{L}_u$. Apart from the overall straightforward structure, as shown in Figure 3, the core of AD-MT lies in two aspects. First, two teachers are updated periodically and randomly in an alternate fashion. During the training

procedure, we utilize complementary sets of unlabeled data, different data augmentation strategies, and randomized switching periods to ensure that the two teacher models are updated in a distinct manner, complementing each other throughout the process. Second, instead of discarding conflicting predictions of the two teacher models, we design a Conflict-Combating strategy that encourages the model to learn from the disagreements between the teachers. This strategy proves highly beneficial in improving segmentation performance at the latter stages of training. These two aspects correspond to our proposed two novel components, the RPA and the CCM, which are detailed in the following two sections.

3.2 Random Periodic Alternate Updating Module

Maintaining two different teacher models can benefit the pseudo-labeling and further reduce the confirmation bias in semi-supervised learning. However, to maximize the benefits of multiple models, it is essential to ensure that they are as diverse as possible. To this end, in AD-MT, we propose the Random Periodic Alternate Updating Module, a novel approach to updating two teacher models in a way that maximizes their diversity with little additional training efforts. In specific, the RPA module involves the following strategies.

- **Alternate Updating.** At each iteration, only one of the two teacher models is updated. Consequently, throughout a complete training epoch, unique and complementary batches of unlabeled data are utilized to refine the two teacher models. This strategy ensures that the updates to the teacher models are distinct, allowing them to complement each other's learning across the training process. We denote the alternate updating period (in the unit of iterations) of each teacher model by $\mathcal{T}_m$, where $m \in \{1, 2\}$.
- **Distinct augmentation strategies.** To further increase the diversity between the two teacher models, we also employ distinct augmentation strategies in addition to using different data batches. Specifically, we apply the color-jitting [6] operation for the turn of T1 while the copy-paste [8] augmentation for the turn of T2.
- **Randomized switching periods.** Instead of adopting a fixed switchable period rigidly for each teacher model, we consider increasing the randomness of the switchable pattern. Given a pre-defined maximum value of the period, denoted by $\mathcal{T}_{max}$, the alternating period for each teacher is randomly generated when switching occurs, *i.e.*,

$$\mathcal{T}_m \leftarrow \text{random.randomint}(0, \mathcal{T}_{max}), \quad m \in \{1, 2\}. \tag{5}$$

In this way, our RPA module cannot only increase the diversity between the two teacher models but also lower the risks of data over-perturbation discussed in [41, 43]. This is simply because these different strong augmentations are not applied simultaneously in our method.

3.3 Conflict-combating Module

Having two distinct teacher models derived from the RPA, our proposed Conflict-Combating Module, abstracted as a function of $\phi(\cdot)$, is carefully designed to address the issue of conflicting predictions between the two teacher models. Instead of discarding these conflicts directly, the CCM module encourages the model to further learn from the disagreements with the help of the student model.

Specifically, the CCM module first separates the consistent and conflicting predictions of the two teacher models. On the one hand, it applies the entropy-based teacher ensemble to obtain an ensembled prediction, ψ_i, on the unlabeled instance u_i, which is directly used for the consistent supervision. Given the prediction entropy (denoted by $H(\cdot)$) of two teacher models,

$$H_{t_1} = H(q_i^{t_1}) = -\sum_{i=1}^{C} q_i^{t_1} \log_2 q_i^{t_1}, \tag{6}$$

$$H_{t_2} = H(q_i^{t_2}) = -\sum_{i=1}^{C} q_i^{t_2} \log_2 q_i^{t_2}, \tag{7}$$

we can obtain the entropy-based ensembled prediction,

$$\psi_i = \psi(q_i^{t_1}, q_i^{t_2}) = \frac{w_1 q_i^{t_1} + w_2 q_i^{t_2}}{w_1 + w_2}, \tag{8}$$

with $w_1 = e^{-H_{t_1}}, w_2 = e^{-H_{t_2}}$. On the other hand, to account for the increasing improvement of the student model, we compare the entropy of the student's prediction with the entropy of the ensembled prediction. We then use the lower-entropy one as the final supervision for the conflicting prediction. This strategy ensures that the conflicting supervision benefits from the strengths of both the teacher models and student model. In summary, the final prediction q_i is,

$$q_i(j) = \begin{cases} \mathbb{1}(\max(q_i^s(j)) \geq \tau) q_i^s(j), & q_i^{t_1} \neq q_i^{t_2} \;\&\; H_{\psi_i(j)} > H_{q_i^s(j)} \\ \mathbb{1}(\max(\psi_i(j)) \geq \tau) \psi_i(j), & \text{otherwise} \end{cases} \tag{9}$$

where $\mathbb{1}(\cdot)$ only selects the high-confidence predictions for the unlabeled supervision. Additionally, we examine more ensembling strategies in the experiment section. As the training process progresses, the student model learns from both diverse teacher models efficiently and effectively. In the inference phase, the student model is used as the final segmentation model, providing accurate and reliable segmentation results on new medical images.

4 Experiments

4.1 Datasets and Evaluation Metrics

Following the previous works [2, 30, 31], we adopt the widely used benchmarks, the **Pancreas-NIH** dataset [24], the Left Atrium (**LA**) dataset [35], and the

Table 1: Performance comparison with the SOTA methods on the **LA**, with 5% and 10% labeled data. † denotes that BCP [2] requires an additional pre-training stage before the semi-supervised training. The best is highlighted in **Bold**.

Method	Left atrium (5% / 4 labeled data)				Left atrium (10% / 8 labeled data)			
	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓
VNet (SupOnly)	52.55	39.60	47.05	9.87	82.74	71.72	13.35	3.26
UA-MT [40] (MICCAI'19)	82.26	70.98	13.71	3.82	86.28	76.11	18.71	4.63
SASSNet [12] (MICCAI'20)	81.60	69.63	16.16	3.58	85.22	75.09	11.18	2.89
DTC [17] (AAAI'21)	81.25	69.33	14.90	3.99	87.51	78.17	8.23	2.36
URPC [19] (MedIA'22)	82.48	71.35	14.65	3.65	85.01	74.36	15.37	3.96
SS-Net [31] (MICCAI'22)	86.33	76.15	9.97	2.31	88.55	79.62	7.49	1.90
MC-Net+ [32] (MedIA'22)	83.59	72.36	14.07	2.70	88.96	80.25	7.93	1.86
PS-MT [15] (CVPR'22)	88.49	79.13	8.12	2.78	89.72	81.48	6.94	1.92
MCF [28] (CVPR'23)	-	-	-	-	88.71	80.41	6.32	1.90
BCP [2]† (CVPR'23)	88.02	78.72	7.90	2.15	89.62	81.31	6.81	1.76
AD-MT (Ours)	**89.63**	**81.28**	**6.56**	**1.85**	**90.55**	**82.79**	**5.81**	**1.70**

Table 2: Performance comparison with the SOTA methods on the **ACDC**, in the semi-supervised setting of 5% and 10% labeled data.

Method	ACDC (5% / 3 labeled data)				ACDC (10% / 7 labeled data)			
	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓
U-Net (SupOnly)	47.83	37.01	31.16	12.62	79.41	68.11	9.35	2.70
UA-MT [40] (MICCAI'19)	46.04	35.97	20.08	7.75	81.65	70.64	6.88	2.02
SASSNet [12] (MICCAI'20)	57.77	46.14	20.05	6.06	84.50	74.34	5.42	1.86
DTC [17] (AAAI'21)	56.90	45.67	23.36	7.39	84.29	73.92	12.81	4.01
CPS [5] (CVPR'21)	70.15	61.17	5.96	2.14	86.91	78.11	5.72	1.92
URPC [19] (MedIA'22)	55.87	44.64	13.60	3.74	83.10	72.41	4.84	1.53
SS-Net [31] (MICCAI'22)	65.82	55.38	6.67	2.28	86.78	77.67	6.07	1.40
MC-Net+ [32] (MedIA'22)	62.85	52.29	7.62	2.33	87.10	78.06	6.68	2.00
PS-MT [15] (CVPR'22)	86.94	77.90	4.65	2.18	88.91	80.79	4.96	1.83
BCP [2] (CVPR'23)	87.59	78.67	1.90	0.67	88.84	80.62	3.98	1.17
AD-MT (Ours)	**88.75**	**80.41**	**1.48**	**0.50**	**89.46**	**81.47**	**1.51**	**0.44**

Automated Cardiac Diagnosis Challeng (**ACDC**) dataset [3] to validate the effectiveness of our proposed AD-MT. The Pancreas-NIH and LA are two 3D datasets, consisting of 82 contrast-enhanced abdominal CT volumes and 100 3D gadolinium-enhanced magnetic resonance image scans, respectively. During the training stages, the 3D images are randomly cropped into 112x112x80 and 96x96x96 for Pancrease and LA, respectively. The ACDC dataset is a 2D benchmark, which contains 100 cardiac MR imaging samples, which are resized into 256×256 pixels and normalized into [0, 1].

Consistent with previous studies [17, 19, 31], we adopt the Dice Score (%), Jaccard Score (%), 95% Hausdorff Distance in voxel (95HD), and Average Surface Distance in voxel (ASD) as our evaluation metrics to compare the segmentation performance under the different semi-supervised partition protocols. A higher Dice Score and Jaccard Score indicate better segmentation performance, while a lower 95HD and ASD indicate better agreement between the predicted segmentation and ground truth.

Table 3: Performance comparison with SOTA methods on the **Pancreas**, in the semi-supervised setting of 10% and 20% labeled data.

Method	Pancreas (10% / 6 labeled data)				Pancreas (20% / 12 labeled data)			
	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓
VNet (SupOnly)	55.60	41.74	45.33	18.63	72.38	58.26	19.35	5.89
UA-MT [40] (MICCAI'19)	66.34	53.21	17.21	4.57	76.10	62.62	10.84	2.43
SASSNet [12] (MICCAI'20)	68.78	53.86	19.02	6.26	77.66	64.08	10.93	3.05
DTC [17] (AAAI'21)	69.21	54.06	17.21	5.95	78.27	64.75	8.36	2.25
ASE-Net [11] (TMI'22)	71.54	56.82	16.33	5.73	79.03	66.57	8.62	2.30
SS-Net [31] (MICCAI'22)	71.76	57.05	17.56	5.77	78.98	66.32	8.86	2.01
MC-Net+ [32] (MedIA'22)	70.00	55.66	16.03	3.87	79.37	66.83	8.52	1.72
PS-MT [15](CVPR'22)	76.94	62.37	13.12	3.66	80.74	68.15	7.41	2.06
MCF [28] (CVPR'23)	-	-	-	-	75.00	61.27	11.59	3.27
BCP [2] (CVPR'23)	73.83	59.24	12.71	3.72	**82.91**	**70.97**	6.43	2.25
AD-MT (Ours)	**80.21**	**67.51**	**7.18**	**1.66**	82.61	70.70	**4.94**	**1.38**

4.2 Implementation Details

Following other SSMIS studies [2,30,40], we adopt the U-Net [23] and V-Net [20] as the backbones for the experiments on 2D and 3D datasets, respectively. For the 2D ACDC dataset, we train the segmentation model with a batch size of 24 (12 labeled and 12 unlabeled instances) for 30,000 iterations. On the LA and Pancreas datasets, we follow existing studies and adopt a batch size of 4 (2 labeled and 2 unlabeled instances) for training 15,000 iterations. We use an SGD optimizer to train the student model with a polynomial learning-rate decay where the initial learning rate, 0.01, is multiplied by $(1-\text{iter}/\text{max\_iter})^{0.9}$. The momentum and the weight decay are set as 0.9 and 0.0001, respectively. The two teacher models are randomly initialized and updated with an exponential parameter of 0.99. By default, we set the maximum loss weight $\lambda_u = 2.0$, and the maximum period $\mathcal{T}_{max} = 0.5$ epoch for all runs.

4.3 Comparison with SOTAs

In this section, we compare our method with the most recent SSMIS methods, including UA-MT [40], SASSNet [12], DTC [17], ASE-Net [11], SS-Net [31], MC-Net+ [30], MCF [28], PS-MT [15] and BCP [2]. Note that BCP requires an additional pre-training stage while other methods do not.

3D LA. As shown in Table 1, our proposed AD-MT approach achieves the highest Dice Score and Jaccard Score on both the 5% and 10% labeled data settings for the LA dataset. With only 4 labeled instances, our approach achieves a Dice Score of 89.63%, outperforming the previous state-of-the-art method BCP [2] by 1.61% in Dice Score and 2.56% in Jaccard Score, without introducing an extra pre-training stage. Also note that our AD-MT outperforms MC-Net+ by over 6% in terms of the Dice score with 4 labeled data available, despite the fact that MC-Net+ introduces far more training parameters. Meanwhile, AD-MT also obtains the lowest 95HD and ASD on both labeled data settings, indicating that our approach produces better segmentation results that are closer to the ground

Table 4: Ablation studies on different components of our proposed AD-MT, when using 3 cases as labeled data on the ACDC. Three different classes of RV, Myo, LV represent the right ventricle, myocardium, and left ventricle, respectively.

Components				**RV**		**Myo**		**LV**		**Mean**	
T1	T2	RPA	CCM	Dice(%)	95HD	Dice(%)	95HD	Dice(%)	95HD	Dice(%)↑	95HD↓
✓				85.30	2.18	84.44	1.53	90.76	4.25	86.83	2.65
	✓			84.59	3.08	83.51	1.86	90.56	2.33	86.22	2.43
✓	✓	✓		85.80	1.98	85.63	1.29	92.20	2.83	87.88	2.03
✓	✓	✓	✓	**86.63**	**1.92**	**86.78**	**1.17**	**92.86**	**1.36**	**88.75**	**1.48**

truth than the other SOTA methods. These results demonstrate the effectiveness of our proposed approach in improving the accuracy of SSMIS, even in scenarios where labeled data is scarce.

2D ACDC. Table 2 shows the results of our AD-MT compared to the current SOTA methods on the ACDC dataset. The ACDC dataset is a challenging dataset for SSMIS due to fine-grained multiple classes and the variability in heart anatomy and pathology, making it a good benchmark for evaluating the effectiveness of our approach. Similar to the results on the LA, AD-MT achieves the best performance on the ACDC dataset in both the 5% and 10% labeled data settings. In the 5% labeled data setting, AD-MT achieves a Dice Score of 88.75%, which is 1.16% higher than BCP's Dice Score of 87.59%. A notable advantage of our proposed AD-MT approach is that it does not require any additional pre-training stage before the semi-supervised training, unlike the BCP method [2]. This makes AD-MT a more efficient and practical approach. AD-MT achieves a significant improvement in Dice Score compared to other end-to-end methods without pre-training, with more than a 20% improvement observed in the 5% labeled setting.

3D Pancreas. The results on the Pancreas-NIH dataset are reported in Table 3. It can be seen from the table that our method demonstrates great recognition performance when the number of labeled data is small, *e.g.*, our AD-MT surpasses the baseline method and UA-MT by a large margin of more than 20% and 10% in terms of the Dice score with 6 labeled data available, respectively. In the 10% labeled setting, AD-MT achieves a Dice Score of 80.21%, which is 6.38% higher than the BCP's Dice Score of 73.83%. In the 20% labeled setting, AD-MT achieves a Dice Score of 82.61%, which is only slightly lower than the BCP's Dice Score of 82.91%. However, AD-MT achieves the lowest 95HD and ASD, indicating that our approach produces more dedicated segmentation results. Similar to the observation on the LA and ACDC, our AD-MT produces highly accurate segmentation on the Pancreas with limited labeled data, which is a significant advantage in real-world scenarios where labeled data is scarce and expensive to obtain.

Table 5: Ablation studies on the threshold τ with 5% labeled data. It is set as 0.95 and 0.75 for the 2D and 3D datasets, respectively.

τ	0.7	0.75	0.8	0.85	0.9	0.95
ACDC	86.58	87.90	87.97	88.07	88.42	**88.75**
LA	89.59	**89.63**	89.35	88.89	88.63	86.94

4.4 Ablations Studies

Impact of different components. In Table 4, we investigate the effectiveness of the main components of our proposed AD-MT on the ACDC dataset with 5% labeled data. The components studied include T1-only, T2-only, the RPA, and the CCM. The evaluation metrics used are Dice Score and 95HD, and we perform a category-wise examination for three classes: the right ventricle (RV), the myocardium (Myo), and the left ventricle (LV). Recall that our AD-MT method iteratively updates two teacher models, named T1 and T2, where T1 is updated by the student model applying color augmentations and T2 is updated by the student model applying mix augmentations. It can be seen from the table that the T1-only obtains slightly better performance than the T2-only, indicating the superiority of the color or intensity based perturbation compared to the copy-paste augmentation. The third row shows the results when both T1 and T2 are involved, along with our proposed RPA module for alternate diverse updating. Although we only use the average prediction of two teacher models at the current stage, we observe a significant improvement in Dice Score for all three classes (RV, Myo, and LV), resulting in a mean improvement of more than 1%. It suggests that our RPA module can indeed generate diverse reasoning and consequently improve the segmentation performance. Furthermore, as discussed in Section 1, two teacher models will inevitably come across conflicting supervision, and it is critical to address the conflicts. As shown in the fourth row of the table, the complete AD-MT with all components, obtains the most accurate segmentation, with the lowest 95HD and the highest Dice Score. Overall, the results demonstrate the importance of each component in our proposed AD-MT approach. Particularly, the RPA and CCM modules are effective in encouraging diverse reasoning and leveraging all involved models to improve the accuracy of the segmentation. The results highlight the importance of leveraging diverse information from multiple sources and effectively combining them to improve segmentation accuracy.

Impact of the threshold τ. Table 5 shows the results of an ablation study on the pre-defined high-confidence threshold (τ) for our AD-MT approach on the ACDC and LA in terms of the Dice Score. For the ACDC, increasing the threshold from 0.7 to 0.95 leads to a gradual improvement in Dice Score, with the highest Dice Score of 88.75% achieved at a threshold of 0.95. Differently, for the LA dataset, increasing the threshold beyond 0.75 leads to a drop in Dice Score. The highest Dice Score of 89.63% is achieved at a threshold of 0.75. It suggests that, on the 3D dataset, a higher threshold may filter out too many predictions,

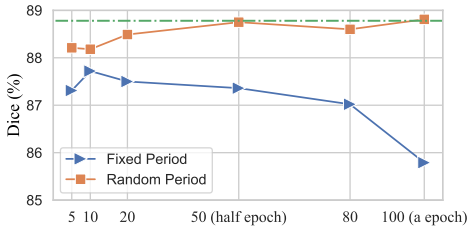


Fig. 4: Impact of different alternating periodic updating strategies in the RPA with varying values of $\mathcal{T}_{max}$ on the ACDC with 5% labeled data. By default, we adopt the random switching periods and set $\mathcal{T}_{max}$ as the half-epoch iterations.

Table 6: Compare different ensembling strategies on the ACDC with 5% labeled data, whenever conflicts occur. "Drop" denotes dropping the conflicts completely. "Avg." and "Ent." represent to use the mean and entropy-based ensembling of two teachers.

Strategy	Drop	Avg.	Ent.	CCM
Dice (%)	86.69	87.88	88.11	**88.75**

leading to a loss of information and a decrease in segmentation accuracy. The default thresholds are 0.95 and 0.75 for the 2D and 3D datasets, respectively.

Impact of the switching periods $\mathcal{T}_{max}$. Figure 4 shows our examinations on different alternate updating periods in the RPA with two different strategies: the fixed and random periods. We can easily observe that the random strategies can consistently outperform the fixed ones. It is simply because the highly randomness can further enlarge the diversity between the two teacher models, resulting in better segmentation performance. In contrast, since both the student and teacher models are trained from scratch, large $\mathcal{T}_{max}$ in the fixed strategies, may enforce ensembling predictions among models with a big performance gap, which significantly reduces the ensembling effectiveness.

Impact of different Ensembling strategies. In Table 6, we investigate different ensembling strategies to address the conflicts on the ACDC with 5% labeled data. We can clearly see that directly dropping all the conflicts leads to the lowest Dice Score of 86.69%, indicating the significance of learning from the conflicts. Using the average and entropy-based ensembling strategies can clearly improve the segmentation performance, which is still limited at only involving the teacher models. As the training progresses, the student model can effectively learn from both teachers and gradually become comparable to the teachers. Our CCM exploits such information and obtains the highest Dice of 88.75%.

Visualization. Figure 5 illustrates example segmentation results on the LA (top 2 rows)and ACDC (bottom 2 rows) in the semi-supervised setting of 5% labeled data. We clearly see that our approach produces segmentation results that are closer to the ground truth than the other SOTA methods. For example, only SS-Net and our AD-MT can recognize the connected segmented region, as shown in the second row. In contrast, SS-Net segments the wrong RV region in the third row on the multi-class ACDC dataset, while our ACDC does not. Additionally, we observe that it is generally more challenging to segment three classes from a Sagittal plane (the fourth row) than from a Coronal plane (the third row) on the ACDC. The better segmentation results further highlight the importance of leveraging diverse information from multiple sources to improve

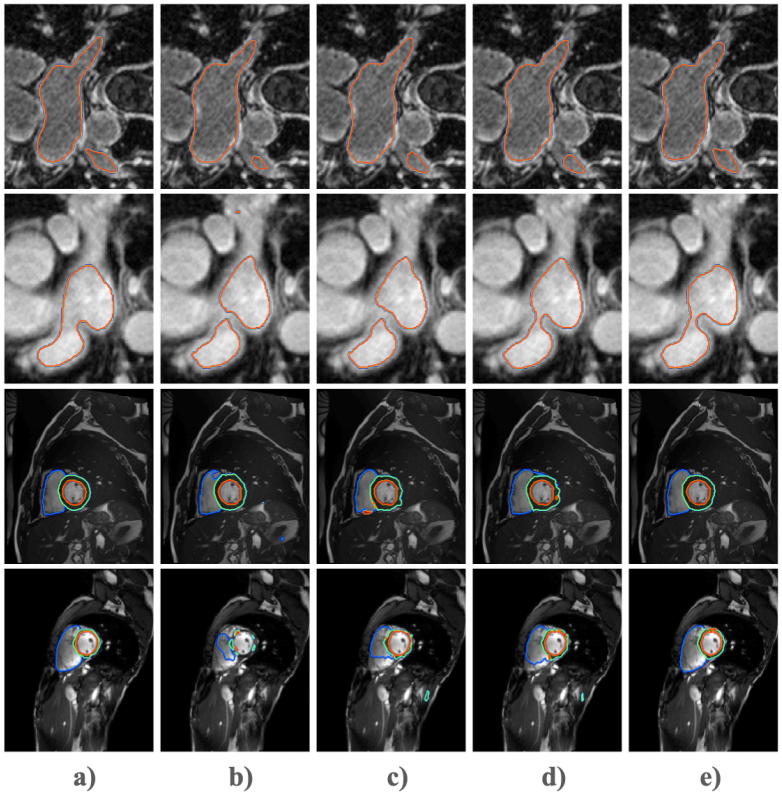


Fig. 5: Qualitative results from the 3D LA (top 2 rows) and 2D ACDC (bottom 2 rows). a) the ground-truth, b) UA-MT, c) MC-Net, d) SS-Net, and e) AD-MT.

segmentation accuracy, as our proposed AD-MT approach does by combining information from multiple modalities and teacher models. Overall, the results in Figure 5 demonstrate the effectiveness of our proposed AD-MT approach in improving the accuracy of SSMIS.

5 Conclusion

In this paper, we propose an alternate diverse teaching approach in a teacher-student framework, which boosts SSMIS via two novel modules: the Random Periodic Alternate Updating Module and the Conflict-Combating Module. With the RPA scheduling, two teacher models are momentum-updated periodically and randomly in an alternate manner to produce diverse supervision. The entropy-based CCM effectively leverages all involved models to encourage the student model to learn from the two teachers' consistent and conflicting predictions. Without introducing extra training parameters and constraints, AD-MT achieves the new SOTA performance on popular 2D and 3D SSMIS benchmarks.

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International Joint Conference on Neural Networks. IEEE (2020) 2, 4
2. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation. arXiv preprint arXiv:2305.00673 (2023) 2, 4, 6, 8, 9, 10, 11
3. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., 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) 9
4. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: Eur. Conf. Comput. Vis. Springer (2022) 1
5. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021) 2, 4, 9
6. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: IEEE Conf. Comput. Vis. Pattern Recog. workshops. pp. 702–703 (2020) 7
7. Duan, Y., Zhao, Z., Qi, L., Wang, L., Zhou, L., Shi, Y., Gao, Y.: Mutexmatch: semi-supervised learning with mutex-based consistency regularization. IEEE Transactions on Neural Networks and Learning Systems (2022) 4
8. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021) 7
9. Gui, G., Zhao, Z., Qi, L., Zhou, L., Wang, L., Shi, Y.: Improving barely supervised learning by discriminating unlabeled samples with super-class. Advances in Neural Information Processing Systems **35**, 19849–19860 (2022) 4
10. Kwon, D., Kwak, S.: Semi-supervised semantic segmentation with error localization network. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022) 2
11. Lei, T., Zhang, D., Du, X., Wang, X., Wan, Y., Nandi, A.K.: Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network. IEEE Transactions on Medical Imaging (2022) 10
12. Li, S., Zhang, C., He, X.: Shape-aware semi-supervised 3d semantic segmentation for medical images. In: Medical Image Computing and Computer Assisted Intervention. pp. 552–561. Springer (2020) 2, 4, 9, 10
13. Li, S., He, Y., Zhang, W., Zhang, W., Tan, X., Han, J., Ding, E., Wang, J.: Cfcg: Semi-supervised semantic segmentation via cross-fusion and contour guidance supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16348–16358 (2023) 2
14. Liu, S., Zhi, S., Johns, E., Davison, A.J.: Bootstrapping semantic segmentation with regional contrast. In: ICLR (2022) 4
15. Liu, Y., Tian, Y., Chen, Y., Liu, F., Belagiannis, V., Carneiro, G.: Perturbed and strict mean teachers for semi-supervised semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022) 2, 4, 9, 10
16. Luo, W., Yang, M.: Semi-supervised semantic segmentation via strong-weak dual-branch network. In: Eur. Conf. Comput. Vis. (2020) 4

17. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 8801–8809 (2021) 2, 6, 9, 10
18. Luo, X., Hu, M., Song, T., Wang, G., Zhang, S.: Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In: International Conference on Medical Imaging with Deep Learning. pp. 820–833. PMLR (2022) 2, 4
19. 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: Medical Image Computing and Computer Assisted Interventio. pp. 318–329. Springer (2021) 9
20. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016) 10
21. Miyato, T., Maeda, S.i., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE transactions on pattern analysis and machine intelligence **41**(8), 1979–1993 (2018) 4
22. Na, J., Ha, J.W., Chang, H.J., Han, D., Hwang, W.: Switching temporary teachers for semi-supervised semantic segmentation. arXiv preprint arXiv:2310.18640 (2023) 4
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015) 10
24. Roth, H.R., Lu, L., Farag, A., Shin, H.C., Liu, J., Turkbey, E.B., Summers, R.M.: Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: Medical Image Computing and Computer-Assisted Intervention. pp. 556–564. Springer (2015) 8
25. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685 (2020) 4
26. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780 (2017) 4, 6
27. Valanarasu, J.M.J., Patel, V.M.: Unext: Mlp-based rapid medical image segmentation network. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V. pp. 23–33. Springer (2022) 1
28. Wang, Y., Xiao, B., Bi, X., Li, W., Gao, X.: Mcf: Mutual correction framework for semi-supervised medical image segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 15651–15660 (2023) 9, 10
29. Wang, Z., Zhao, Z., Zhou, L., Xu, D., Xing, X., Kong, X.: Conflict-based cross-view consistency for semi-supervised semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023) 2, 4
30. Wu, Y., Ge, Z., Zhang, D., Xu, M., Zhang, L., Xia, Y., Cai, J.: Mutual consistency learning for semi-supervised medical image segmentation. Medical Image Analysis **81**, 102530 (2022) 8, 10
31. Wu, Y., Wu, Z., Wu, Q., Ge, Z., Cai, J.: Exploring smoothness and class-separation for semi-supervised medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V. pp. 34–43. Springer (2022) 2, 4, 6, 8, 9, 10

32. Wu, Y., Xu, M., Ge, Z., Cai, J., Zhang, L.: Semi-supervised left atrium segmentation with mutual consistency training. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. pp. 297–306. Springer (2021) 2, 4, 9, 10
33. Xiang, J., Qiu, P., Yang, Y.: Fussnet: Fusing two sources of uncertainty for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 481–491. Springer (2022) 1
34. Xie, Q., Dai, Z., Hovy, E., Luong, M.T., Le, Q.V.: Unsupervised data augmentation for consistency training. arXiv preprint arXiv:1904.12848 (2019) 4
35. Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., Vesal, S., Ravikumar, N., Maier, A., Yang, X., et al.: A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Medical image analysis **67**, 101832 (2021) 8
36. Yan, X., Tang, H., Sun, S., Ma, H., Kong, D., Xie, X.: After-unet: Axial fusion transformer unet for medical image segmentation. In: WACV (2022) 1
37. Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y.: Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7236–7246 (2023) 4
38. Yang, L., Zhuo, W., Qi, L., Shi, Y., Gao, Y.: St++: Make self-training work better for semi-supervised semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022) 4
39. You, C., Dai, W., Min, Y., Staib, L., Duncan, J.S.: Bootstrapping semi-supervised medical image segmentation with anatomical-aware contrastive distillation. In: International Conference on Information Processing in Medical Imaging. pp. 641–653. Springer (2023) 1
40. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 605–613. Springer (2019) 2, 4, 6, 9, 10
41. Yuan, J., Liu, Y., Shen, C., Wang, Z., Li, H.: A simple baseline for semi-supervised semantic segmentation with strong data augmentation. In: Int. Conf. Comput. Vis. (2021) 4, 7
42. Zhao, Z., Long, S., Pi, J., Wang, J., Zhou, L.: Instance-specific and model-adaptive supervision for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23705–23714 (2023) 4
43. Zhao, Z., Yang, L., Long, S., Pi, J., Zhou, L., Wang, J.: Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11350–11359 (2023) 4, 7
44. Zhao, Z., Zhao, M., Liu, Y., Yin, D., Zhou, L.: Entropy-based optimization on individual and global predictions for semi-supervised learning. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 8346–8355 (2023) 4
45. Zhao, Z., Zhou, L., Duan, Y., Wang, L., Qi, L., Shi, Y.: Dc-ssl: Addressing mismatched class distribution in semi-supervised learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2022) 4
46. Zhao, Z., Zhou, L., Wang, L., Shi, Y., Gao, Y.: Lassl: Label-guided self-training for semi-supervised learning. In: AAAI (2022) 4