ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240509

Decoupled Competitive Framework for Semi-Supervised Medical Image Segmentation

Jiahe Chen[a,1], **Jiahe Ying**[a,1], **Shen Wang**[a] **and Jianwei Zheng**[a,*]

[a]Zhejiang University of Technology

Abstract. Confronting the critical challenge of insufficiently annotated samples in medical domain, semi-supervised medical image segmentation (SSMIS) emerges as a promising solution. Specifically, most methodologies following the Mean Teacher (MT) or Dual Students (DS) architecture have achieved commendable results. However, to date, these approaches face a performance bottleneck due to two inherent limitations, *e.g.*, the over-coupling problem within MT structure owing to the employment of exponential moving average (EMA) mechanism, as well as the severe cognitive bias between two students of DS structure, both of which potentially lead to reduced efficacy, or even model collapse eventually. To mitigate these issues, a Decoupled Competitive Framework (DCF) is elaborated in this work, which utilizes a straightforward competition mechanism for the update of EMA, effectively decoupling students and teachers in a dynamical manner. In addition, the seamless exchange of invaluable and precise insights is facilitated among students, guaranteeing a better learning paradigm. The DCF introduced undergoes rigorous validation on three publicly accessible datasets, which encompass both 2D and 3D datasets. The results demonstrate the superiority of our method over previous cutting-edge competitors. Code will be available at https://github.com/Fly-away20/DCF.

1 Introduction

Medical image segmentation (MIS) is critical in modern healthcare, offering clinicians vital information to monitor disease progression and develop treatment strategies. The advent of neural networks, especially supervised deep learning methods, has significantly advanced this field, leading to unparalleled performance in various segmentation tasks [25, 47, 31, 12, 41]. However, the practical model efficacy largely depends on the availability of massive laboriously annotated datasets, which not only demands specialized knowledge but is also extremely costly and time-consuming [9, 29]. To save considerable resources, recent years have witnessed an increasing focus on exploring methods like semi-supervised learning (SSL) to reduce the annotation burden in MIS domain [4, 1, 22].

The advantage of semi-supervised image segmentation lies particularly in leveraging unlabeled data to favor better segmentation. Within this framework, two key strategies, *i.e.*, pseudo-label supervision [43, 14, 3] and consistency regularization [24, 19, 23], have been intensively investigated. Meanwhile, co-training or mutual learning paradigm, which can be regarded as a combination of the above two methods, has achieved promising results [27, 38, 5]. Typically, the

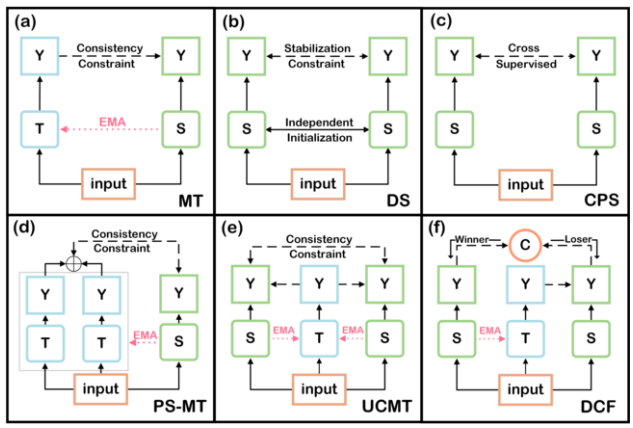


Figure 1. Different SSL structures. (a) Mean Teacher. (b) Dual Student. (c) Cross Pseudo Supervision. (d) Perturbed and Strict Mean Teachers. (e) Uncertainty-guided Collaborative Mean Teacher. (f) Decoupled Competitive Framework (ours).

Mean Teacher model [30] has dominated in this field for a long time, inspiring numerous followers seeking notable advancements [42, 19, 27, 39]. Yet, Ke et al. [13] and Zhao et al. [46] have demonstrated that MT-based methods encounter performance limitations arising from the treatment of exponential moving average (EMA). Besides, employing one coupled EMA teacher is insufficient to adequately support the student model. To break the performance bottleneck, it is often necessary to integrate additional modules with an intricate architecture into MT-based methodologies, such as MagicNet [4] and BCP [1].

The teacher-less architecture is yet another research branch, which offers a promising avenue to mitigate the issue of overly tight coupling. Without a teacher model, the main challenge lies in facilitating efficient knowledge extraction and exchange between two autonomous models. Ke et al. [13] introduced Dual Student (DS) as a remedy, replacing the teacher with another student, and integrating a stabilization constraint during training stage. Following this line, Zhao et al. [46] additionally integrated region-level uncertainty estimation to ensure better performance.

While the potential of DS is widely acknowledged, successful practices again require the integration of intricate constraints within student models; otherwise, it may lead to model collapse due to the abnormal exchange of erroneous information, which remains as an emergent issue to be solved. Upon this aspiration, we propose a straightforward yet potent solution, *i.e.*, a Decoupled Competitive Framework (DCF), whose disparity against current architecture is

* Corresponding Author. Email: zjw@zjut.edu.cn
[1] Equal contribution.

given in Figure 1. Note that the real-time performance of two student models can be iteratively assessed. On that basis, the superior-performing student model shall be leaned to update the EMA of teacher model. The teacher model then further provides pseudo-labels to the inferior-performing student to favor an improvement. Through this dynamic mechanism, both student models have the opportunity to contribute to EMA update of the teacher model, which naturally reduces the coupling between a single teacher-student pair. Technically, our work makes three primary contributions.

- With the deficiencies of MT-based and DS-based methods deeply scrutinized, we engineer a novel Decoupled Competitive Framework, effectively surmounting the bottleneck of existing models.
- An efficient competition and mentoring mechanism is crafted to mitigate the tight coupling between the teacher's parameters and the individual students, thereby augmenting students' capacity to acquire valuable knowledge.
- Yet without any sophisticated modules employed, our DCF sets new state-of-the-art scores among three benchmark datasets, namely left atrium segmentation in MRI, pancreas segmentation in CT scans, and dermoscopy images.

2 Related Work

2.1 Semi-supervised Learning

Semi-supervised learning (SSL) is a widely employed method in numerous computer vision tasks [35, 10, 18], aiming at the mitigation of performance degradation encountered in cases with limited training samples. Commonly, SSL relies on three core assumptions: (1) Smoothness assumption that ensures similar inputs would yield similar outputs and vice versa; (2) Cluster assumption, which suggests that instances of the same class tend to be clustered together in the feature space. Consequently, the classification boundary should traverse sparsely populated regions while avoiding densely populated areas on either side. (3) Manifold assumption, which considers that samples residing within a compact neighborhood in a low-dimensional manifold are likely to share similar labels. Currently, two semi-supervised learning branches, *i.e.*, pseudolabel-based and consistency-based, have been extensively investigated.

Pseudolabel-based SSL: Pseudo-labeling methods adopt a supervised paradigm that simultaneously learns from labeled and unlabeled data. The essence of this branch lies in the reliable generation of pseudo-labels [16]. For instance, Fixmatch [28] employed a fixed threshold to determine the trustworthiness of the samples, ensuring high quality and reliability of the pseudo-labels. Moreover, Ref. [14] utilized an auxiliary error localization network, identifying pixels with potentially erroneous labels. Additionally, Freematch [35] dynamically adjusted the confidence threshold based on the learning states of the involved model.

Consistency-based SSL: According to the assumptions of smoothness and clustering, model predictions are expected to exhibit similarity when specific perturbations are applied, which may involve adjustments to input data, features, or networks. Drawing on this inspiration, Laine et al. introduced a Pi-Model and a temporal ensembling model [15], aiming to exploit both data-level and model-level consistencies. Subsequently, Tarvainen et al. presented the MT model [30], in which the student network utilizes EMA to update the parameters of teacher network, thereby reducing model-level inconsistencies. To take advantage of unlabeled data, CCT [24] employed the training of multiple auxiliary decoders, each receiving distinct perturbations of the output generated by the shared encoder.

2.2 Semi-supervised medical image segmentation

In contrast to natural semantic segmentation, medical images often suffer from limited data availability while requiring higher prediction accuracy. Consequently, there is an urgent need to explore efficient semi-supervised methods to alleviate data requirements and improve accuracy. Through an examination of pseudo-labeling methods, the practical quality can be refined using techniques such as uncertainty knowledge [32], and random propagation [7], among others. Additionally, Lyu et al. [21] suggested generating synthetic images aligned with the retained pseudo-labels. In methods employing consistency regularization, Yu et al. [42] proposed an uncertainty-aware mean teacher model for left atrium segmentation, while Wang et al. [34] introduced a double-uncertainty weighted method for semi-supervised applications. Moreover, Huang et al. [11] developed a two-stage learning scheme for neuron segmentation, which fully extracts useful information from unlabeled data. Furthermore, numerous other practices are also available to support semi-supervised medical image segmentation. Bai et al. [1] utilized bidirectional copy-paste to prompt unlabeled data to assimilate comprehensive semantics from labeled data, thus mitigating the experience mismatch problem between labeled and unlabeled data. Additionally, [2, 45, 36, 44] have incorporated contrastive learning into SSMIS, with the aim of learning representations of distinct features and emphasizing differences in feature spaces across various categories.

2.3 Different Structures for SSL

As illustrated in Figure 1, five SSL architectures currently dominate this field, namely MT [30], DS [13], CPS [5], PS-MT [19], and UCMT [27]. Additionally, our DCF is also swept in generalization.

Mean Teacher: In Figure 1 (a), MT model mainly comprises two networks with identical architectures: the teacher network and the student network. While the parameters of the student network are updated through backpropagation, the teacher network undergoes updating via Exponential Moving Average. Nevertheless, this treatment is currently stuck in a performance bottleneck.

Dual Student: DS model involves two students with shared architectures, which utilizes stable samples to impose effective constraints between the two counterparts, thereby mitigating the problem of over-coupling that often encountered within EMA computation. Please refer to Figure 1 (b) for a visual depiction.

CPS: As shown in Figure 1 (c), the CPS structure integrates both self-training and consistency learning methodologies, wherein one-hot pseudo labels derived from the outcomes of both models serve as supervision signals, mutually guiding and supervising each other's learning processes.

PS-MT: PS-MT in Figure 1 (d) employs two teachers. To produce pseudo labels, the prediction outcomes of both are merged using an ensemble approach, bolstering the stability of the pseudo labels. Furthermore, during each training epoch, only one of the teachers undergoes updating, adjusting the model parameters to augment the diversity between two subassemblies.

UCMT: UCMT integrates collaborative mean teacher techniques and uncertainty-guided region mixture to concurrently maintain model inconsistency and high-confidence labels, resulting in promising outcomes. Please see Figure 1(e) for a visual representation.

DCF (ours): Due to EMA computation, MT-based techniques inevitably result in the over-coupling issue, yet the no-teacher alternatives often lack a direct method to enforce consistency constraints. Therefore, we introduce DCF to mitigate these challenges. For an in-depth analysis, please refer to Section 3.

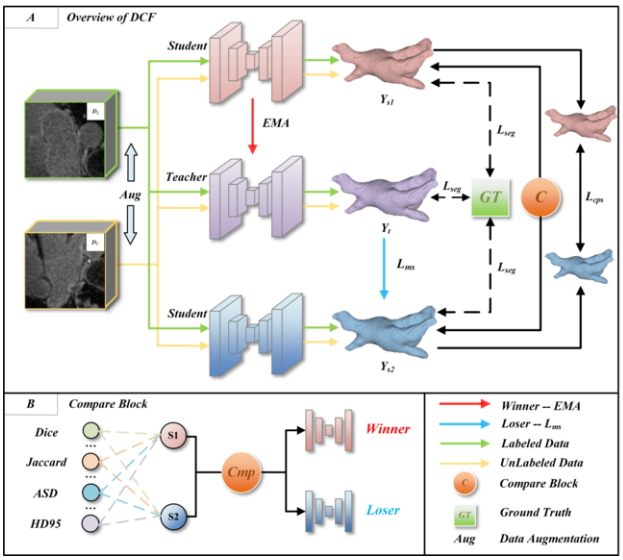


Figure 2. Block A introduces our proposed DCF. Upon receiving input data, DCF undergoes two random data augmentations. Then, three separate networks follow: a student network and two teacher networks. Operating on a co-teaching scheme, DCF fosters cross-pseudo supervision between two student results. In Block B, a competitive mechanism employing metrics such as Dice, Cross-entropy, and 95HD is elaborated during training to compare the performance of the two students and determine a winner.

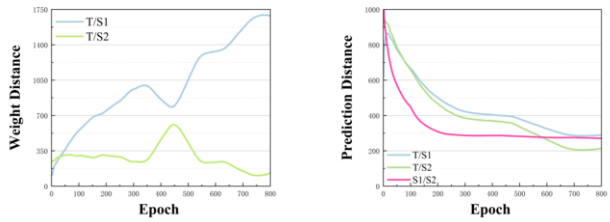
3 Methodology

3.1 The overview framework

During semi-supervised learning, it is assumed that the training dataset contains N labeled data and M unlabeled data, where $M \gg N$. For convenience, we denote the entire training set as $\mathcal{D} = \{\mathcal{D}_L, \mathcal{D}_U\}$, with a small portion of labeled data represented as $\mathcal{D}_L = \{(x_i^L, y_i^L)\}_{i=1}^N$, and the unlabeled counterpart as $\mathcal{D}_U = \{x_i^U\}_{i=1}^M$. Here, x_i denotes the training image, and y_i is the label (if available). Yet with a limited number of labeled samples x_i^L, the objective of semi-supervised learning is to achieve promising results with the aid of the extra unlabeled data x_i^U.

As discussed, the MT technique is the dominant framework for most contemporary SSL approaches [30, 19, 1, 42, 39]. However, the Exponential Moving Average (EMA) mechanism leads to excessive coupling between teachers and students, resulting in performance bottlenecks. Subsequently, although the Dual Student framework [13, 46] has addressed the coupling issue, it requires the development of a stable sample and a training method integrating entra constraints to facilitate the exchange of correct knowledge between the two students and prevent model collapse due to erroneous knowledge exchange.

Motivated by the aforementioned issues, we introduce the Decouple Competitive Framework (DCF), which employs a unique competition mechanism to select students currently with superior performance to effectively mitigate the problem of excessive coupling between a single teacher and a student. Moreover, we further devise a mentoring mechanism, which empowers teachers to grant additional learning privileges to underperforming students, narrowing the cognitive bias between the two students. In addition, we advocate mutual learning and assistance between the two students, facilitated by the teacher. Thus, a straightforward consistency loss between students is sufficient to ensure alignment with learning objectives (detailed elucidation provided in Subsection 3.2).



(a) Weight Distance (b) Prediction Distance

Figure 3. (a) denotes the weight distance between teacher and student. (b) represents the prediction distance among three networks. For simplicity, the simple Euclidean distance is used for weight and prediction distances.

The overall framework of the proposed DCF is shown in Figure 2, which mainly comprises three networks: a teacher network $f(\cdot;\theta_t)$ and two student networks $f(\cdot;\theta_{s1})$ and $f(\cdot;\theta_{s2})$, all initialized randomly. For each training data X that encompasses both labeled and unlabeled samples, we introduce random augmentation ξ and $\xi^{'}$ to generate perturbed instances. Subsequently, these samples independently traverse through the three networks, producing their respective predictions: $Y_t = f(X;\theta_t)$, $Y_{s1} = f(X+\xi;\theta_{s1})$, and $Y_{s2} = f(X+\xi^{'};\theta_{s2})$.

3.2 Decoupled Competitive mechanism

As analyzed previously, the student model may inadvertently overshadow the teacher model's capacity to assimilate information during the training phase. In the dual student architecture, the learning capabilities of the two students may not be uniform, thus introducing a cognitive bias that can result in suboptimal performance. To address these issues, we propose a straightforward yet effective decoupled competitive mechanism.

The training process of DCF is shown in Algorithm 1. Referring to previous work [42], for labeled data, cross-entropy loss L_{ce} and dice loss L_{dice} are utilized for supervised training:

$$L_{seg} = L_{ce}(f(x_i^L;\theta), y_i) + L_{dice}(f(x_i^L;\theta), y_i) \tag{1}$$

where y_i is the label of x_i^L.

With the supervised loss L_{seg} calculated using ground truth, a competition function is employed to determine which backbone performs more competitively in the current state. This competition function may encompass various metrics such as Dice coefficient, Cross-entropy (CE), and 95% Hausdorff distance (95HD), which can be achieved on-the-fly during the training process. Following this line, two advantages can be guaranteed. 1) it obviates the need for modifying the network structure, thus avoiding an increase in parameter count, and 2) as some indicators involve supervised learning processes and must be computed anyway, no additional computational overhead is introduced. Upon determining the winner for the current iteration, we utilize its parameters to update the teacher network $\theta_t^{'}$ at training step t based on EMA, *i.e.*, $\theta_t^{'} = \alpha\theta_{t-1}^{'} + (1-\alpha)\theta_t$, where α is the EMA decay that controls the updating rate. Since the winner changes dynamically, the parameters of teacher network remain uncoupled from any particular students, facilitating the assimilation of more effective information during training. Moreover, from the perspective of the teacher network, the status of both students is the same, hence the decoupling of tight dependency can be accomplished, which further allows for a more exclusive focus on the acquisition of desirable knowledge.

For unlabeled data, the latent knowledge they harbor merits comprehensive exploration. Hence, we introduce L_{unsup} to fully exploit the relationship between labeled and unlabeled data, especially in domains such as medical imaging, where scenes often manifest consistent semantic information across the dataset.

Throughout the overall training process, to effectively promote the efficacy of underperforming students (us), we utilize the teacher network to correctly steer student models toward correct optimization, thereby preventing them from converging in erroneous directions. This can be likened to a tutoring process, for which the mentoring loss is defined as:

$$L_{ms} = \frac{1}{M}\sum_{j=1}^{M} L_{seg}(f(x_j^U;\theta_{us}), \hat{Y}_j^U) \tag{2}$$

where $\hat{Y}_j^U$ is the pseudo label of $f(x_j^U;\theta_t)$.

As widely acknowledged, maintaining consistency in model prediction results is of paramount importance in semi-supervised learning methods. However, applying conformance constraints directly can cause models to collapse with each other due to the exchange of incorrect knowledge [13]. Therefore, Dual Student-based methods often incorporate additional techniques to ensure an accurate exchange of information between models, thereby preventing the collapse issue. Nevertheless, due to the presence of these mechanisms, we can achieve satisfactory results by simply adding a straightforward consistency constraint between the two students, significantly reducing performance overhead. The specific cross-pseudo-supervision loss function is as follows:

$$L_{cps} = \frac{1}{M}\sum_{j=1}^{M}(L_{seg}(f(x_j^U;\theta_{s1}), \hat{Y}_j^U) + L_{seg}(f(x_j^U;\theta_{s2}), \hat{Y}_j^U)) \tag{3}$$

where $\hat{Y}j^U$ represents the pseudo label of $f(x_j^U;\theta_{s2})$ if the result compared with him comes from Student1, and vise versa. The loss function for unsupervised data L_{unsup} can then be formulated as follows:

$$L_{unsup} = L_{cps} + L_{ms} \tag{4}$$

Note that L_{ms} is exclusively assigned to currently underperforming students.

With all the sub-loss assembled, the overall loss is given as follows. It's noteworthy that during the early stages of training, the network's uncertainty tends to be relatively high. Therefore, in line with previous practices [45, 20], we introduce a parameter within the L_{unsup} to stabilize the model training.

$$L_{total} = L_{seg} + \lambda L_{unsup} \tag{5}$$

where λ is the concerned weight for balance control of L_{unsup}.

3.3 Discussions

In Figure 3(a), we provide a graphical representation of the weight distance between the teacher network (t) and both student networks (s1 and s2), throughout the training process. To visualize this, we show the curves for more epochs of training. Notably, we observe an inverse relationship between the weight distance of t and s1, and that of t and s2, *i.e.*, as the weight distance between t and s1 decreases, there is a corresponding increase in the weight distance between t and s2, demonstrating a clear antagonistic trend. We attribute this phenomenon to the fact that when s1 performs EMA update on the parameters of t, the weight between them will become similar, resulting in a decrease in the weight distance between s1 and t. This can be interpreted as indicating that the one who updates the teacher with EMA will become closer in weight distance, while the others will show a tendency to move farther away. In this scenario, the weight of the teacher is not overly tethered to a single student, but alternates between two students. As these two students progress in tandem, the teacher can glean effective information from their interaction, thus circumventing the bottleneck of poor performance induced by excessive parameter coupling.

Algorithm 1 Training of DCF for SSL

Require:

- The set of samples: X
- The random augmentation: ξ, $\xi^{'}$
- The teacher network: $f_t(\theta_t)$
- The student networks: $f_{s1}(\theta_{s1})$, $f_{s2}(\theta_{s2})$

Procedure:

1: **for** each iteration **do**
2: Get $f(X+\xi;\theta_{s1})$, $f(X+\xi^{'};\theta_{s2})$, $f(X;\theta_t)$
3: Calculate supervised Loss on labeled samples
4: Calculate cross pseudo supervision loss on unlabeled samples between $f_{s1}(\theta_{s1})$ and $f_{s2}(\theta_{s2})$
5: Compare $f_{s1}(\theta_{s1})$, $f_{s2}(\theta_{s2})$ and get the winner $f_w(\theta_s)$ and the loser $f_l(\theta_s)$
6: $f_t(\theta_t)$ assists $f_l(\theta_s)$
7: $f_w(\theta_s)$ updates $f_t(\theta_t)$
8: **end for**

At the same time, we also plot the Prediction Distance between the three networks during the training process, as illustrated in Figure 3(b). It is evident that in the initial stages of training, the prediction distance between the two students steadily diminishes, indicating the influence of the consistency constraint among the students. Consequently, the predicted outcomes of the two students become increasingly similar. Subsequently, the prediction distance between them reaches a threshold and remains relatively constant. Moving forward, the prediction distance between the teacher and the two students exhibits a similar trend, underscoring the model's efficacy and the effective consistency observed among the three networks.

4 Experiments and Results

4.1 Datasets and Metrics

ISIC Dataset. ISIC [6] was released by the International Skin Imaging Collaboration (ISIC), which comprises 2594 dermoscopic 2D images along with the corresponding annotations. Following [33, 27], we use 1815 images for training and 779 images for validation. In the training set, 5% (91) and 10% (181) of the images are labeled for different semi-supervised experimental settings.

Left Atrial (LA) Dataset. LA [40] is a benchmark dataset from the 2018 Atrial Segmentation Challenge, consisting of 100 3D gadolinium-enhanced MR imaging volumes. Each volume has an isotropic resolution of $0.625 \times 0.625 \times 0.625mm^3$, whose ground truth labels are all given. According to previous work [42], we utilize 80 scans for training purposes and reserve 20 scans for evaluation. In the training set, 10% (8), and 20% (16) of the images are labeled for different semi-supervised experimental settings.

Pancreas-CT Dataset. Pancreas-CT is also a well-known dataset [26], which is publicly accessible from the National Institutes of Health Clinical Center. For ease of research and analysis, the scans

Table 1. Performance comparison with state-of-the-art methods on LA Dataset. Taking V-Net as the baseline, the green triangle ▼ denotes the reduction degree, while upturned red triangle ▲ represents the rising rate.

Competing Methods	Volumes used		Metrics			
	Labeled	Unlabeled	Dice(%)↑	Jaccard(%)↑	95HD(voxel)↓	ASD(voxel)↓
V-Net	8(10%)	72	78.96	67.82	20.83	5.74
V-Net	16(20%)	64	86.87	77.19	11.93	3.29
UA-MT (MICCAI 2019)	8(10%)	72	84.25▲6.70%	73.48▲8.34%	13.84▼33.5%	3.36▼41.4%
SASSNet (MICCAI 2020)	8(10%)	72	87.32▲10.6%	77.72▲14.6%	9.62▼53.8%	2.55▼55.6%
DTC (AAAI 2021)	8(10%)	72	87.43▲10.7%	78.06▲15.1%	8.37▼59.8%	2.40▼58.2%
MC-Net+ (MIA 2022)	8(10%)	72	88.96▲12.7%	80.25▲18.3%	7.93▼61.9%	1.86▼67.6%
FUSSNet (MICCAI 2022)	8(10%)	72	89.12▲12.9%	80.79▲19.1%	7.13▼65.8%	1.81▼68.5%
CAML (MICCAI 2023)	8(10%)	72	89.44▲13.3%	81.01▲19.4%	10.10▼51.5%	2.09▼63.6%
UCMT (IJCAI 2023)	8(10%)	72	88.13▲11.6%	79.18▲16.7%	9.14▼56.1%	3.06▼46.7%
VSRC (JBHI 2023)	8(10%)	72	88.42▲12.0%	79.57▲17.3%	8.52▼59.1%	2.37▼58.7%
BCP (CVPR 2023)	8(10%)	72	89.62▲13.5%	81.31▲19.9%	6.81▼67.3%	**1.76**▼69.3%
DCF (ours)	8(10%)	72	**89.94**▲13.9%	**81.78**▲20.6%	**6.38**▼69.4%	1.80▼68.6%
UA-MT (MICCAI 2019)	16(20%)	64	88.88▲2.31%	80.21▲3.91%	7.32▼38.6%	2.26▼31.3%
SASSNet (MICCAI 2020)	16(20%)	64	89.54▲3.07%	81.24▲5.25%	8.24▼30.9%	2.20▼33.1%
DTC (AAAI 2021)	16(20%)	64	89.42▲2.93%	80.98▲4.91%	7.32▼38.6%	2.10▼36.2%
MC-Net+ (MIA 2022)	16(20%)	64	91.07▲4.83%	83.67▲8.39%	5.84▼51.0%	1.67▼49.2%
FUSSNet (MICCAI 2022)	16(20%)	64	91.13▲4.90%	83.79▲8.55%	**5.10**▼57.2%	1.56▼52.6%
CAML (MICCAI 2023)	16(20%)	64	90.71▲4.42%	83.08▲7.63%	6.08▼49.0%	1.59▼51.7%
UCMT (IJCAI 2023)	16(20%)	64	90.41▲4.07%	82.54▲6.93%	6.31▼47.1%	1.70▼48.3%
VSRC (JBHI 2023)	16(20%)	64	90.59▲4.28%	82.60▲7.01%	5.60▼53.1%	1.72▼47.7%
BCP (CVPR 2023)	16(20%)	64	90.74▲4.45%	83.17▲7.75%	6.40▼46.3%	1.65▼49.8%
DCF (ours)	16(20%)	64	**91.44**▲5.26%	**84.28**▲9.19%	5.24▼56.1%	**1.55**▼52.9%

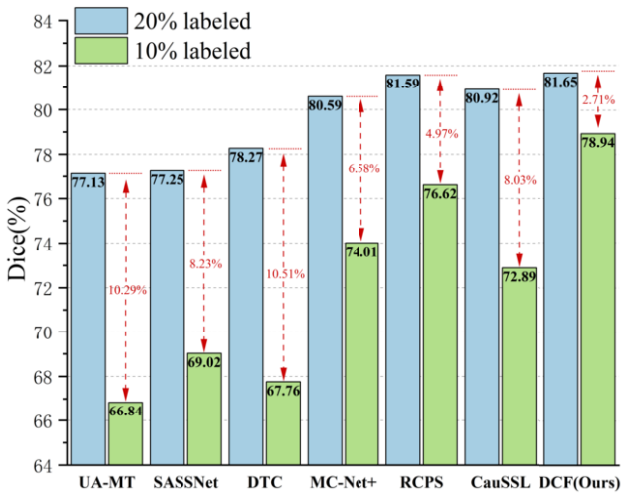


Figure 4. Dice scores for 10% and 20% labeled data across various models on the Pancreas-CT dataset. Our method exhibits a significantly smaller performance gap between these two cases.

were preprocessed, involving adjustments of Hounsfield Units (HU) to ranges of $[-125, 275]$ or $[-120, 240]$, as per the specific study requirements, and resampled to an isotropic resolution of $1.0\ mm \times 1.0\ mm \times 1.0\ mm$. Consistent with previous protocols [37, 22], the dataset is divided into 62 training samples and 20 samples for performance evaluation.

Evaluation Metrics: For 3D datasets, four typical metrics with different criteria are employed, including Dice Similarity Coefficient (Dice), Jaccard Similarity Coefficient (Jac), 95% Hausdorff Distance (95HD), and Average Surface Distance (ASD). Among them, Dice and Jaccard are regional sensitivity metrics assessing the overlap between predictions and ground truth. Both 95HD and ASD are edge-sensitive metrics. The former determines the maximum surface-to-surface distance at the 95th percentile between predicted and actual regions, while the latter computes the average distance between concerned points on both surfaces. As for 2D datasets, the primary evaluation metric for segmentation performance is the widely used Dice coefficient.

4.2 Implementation Details

DCF is implemented with PyTorch and executed on an NVIDIA 3090 GPU. In the following, we provide distinct processing methods employed for various datasets.

Left Atrial Dataset: For the LA dataset, we utilize Vnet as the baseline and trained the network for 500 epochs. The batch size is set to four, comprising two labeled and two unlabeled images. During the training phase, random volume cropping is conducted, resulting in input dimensions of $112 \times 112 \times 80$ for model updates. During the inference phase, segmentation results are generated using a sliding window with the same dimensions and a stride of $18 \times 18 \times 4$.. AdamW is employed as the optimizer, with a fixed learning rate of 1e-4.

Pancreas-CT Dataset: Throughout the training process, all volumes undergo random cropping to attain dimensions of $96 \times 96 \times 96$. While during inference, a stride of $16 \times 16 \times 16$ is implemented. We trained the network for 600 epochs for the PA dataset. Other configurations mirror those of the LA dataset.

ISIC Dataset: DeepLabv3+ augmented with ResNet50 serves as the baseline architecture for the ISIC dataset. The batch size is set to 8, including 4 labeled samples and 4 unlabeled samples. All images are resized to 256×256 during inference, with outputs reverted to their original dimensions for evaluation. Again, AdamW serves as the optimizer, with a fixed learning rate set at 1e-4. We train the network for 30 epochs for the ISIC dataset.

4.3 Results on Left Atrial Dataset

The evaluation results on LA are summarized in Table 1, where we compare our proposed DCF with several other SSMIS methods, including UA-MT [42], SASSNet [17], DTC [20], MC-Net+ [37], FUSSNet [39], CAML [8], UCMT [27], VSRC [46], and BCP [1]. Additionally, the classic V-net is used as a fully supervised benchmark model, presenting its performance for reference purposes. To ensure a fair comparison, we implement these models using their official codes and maintain consistency with their respective parameter settings. Furthermore, to comprehensively assess the performance

Table 2. Performance comparison with state-of-the-art methods on Pancreas-CT Dataset. Taking V-Net as the baseline, the green triangle ▼ denotes the reduction degree, while upturned red triangle ▲ represents the rising rate.

Competing Methods	Volumes used		Metrics			
	Labeled	Unlabeled	Dice(%)↑	Jaccard(%)↑	95HD(voxel)↓	ASD(voxel)↓
V-Net	6(10%)	56	55.10	41.02	33.72	12.79
V-Net	12(20%)	50	72.24	58.22	19.39	5.39
UA-MT (MICCAI 2019)	6(10%)	56	66.84▲21.3%	51.73▲26.1%	21.32▼36.8%	6.12▼52.2%
SASSNet (MICCAI 2020)	6(10%)	56	69.02▲25.3%	53.21▲29.7%	18.77▼44.3%	3.09▼75.8%
DTC (AAAI 2021)	6(10%)	56	67.76▲23.0%	52.14▲27.1%	15.98▼52.6%	4.21▼67.1%
MC-Net+ (MIA 2022)	6(10%)	56	74.01▲34.3%	60.02▲46.3%	12.59▼62.7%	3.34▼73.9%
RCPS (JBHI 2023)	6(10%)	56	76.62▲39.1%	62.96▲53.5%	16.32▼51.6%	3.01▼76.5%
CauSSL (ICCV 2023)	6(10%)	56	72.89▲32.3%	58.06▲41.5%	14.19▼57.9%	4.37▼65.8%
DCF (ours)	6(10%)	56	**78.94**▲43.3%	**66.05**▲61.0%	**11.69**▼65.3%	**1.38**▼89.2%
UA-MT (MICCAI 2019)	12(20%)	50	77.13▲6.77%	63.28▲8.70%	10.52▼45.7%	2.39▼55.6%
SASSNet (MICCAI 2020)	12(20%)	50	77.25▲6.94%	63.59▲9.22%	11.98▼38.2%	3.12▼42.1%
DTC (AAAI 2021)	12(20%)	50	78.27▲8.35%	64.75▲11.2%	8.36▼56.9%	2.25▼58.2%
MC-Net+ (MIA 2022)	12(20%)	50	80.59▲11.6%	68.08▲16.9%	**6.47**▼66.3%	1.74▼67.7%
RCPS (JBHI 2023)	12(20%)	50	81.59▲12.9%	69.04▲18.6%	7.50▼61.3%	2.03▼62.3%
CauSSL (ICCV 2023)	12(20%)	50	80.92▲12.0%	68.26▲17.2%	8.11▼58.2%	1.53▼71.6%
DCF (ours)	12(20%)	50	**81.65**▲13.0%	**69.48**▲19.3%	6.77▼65.1%	**1.21**▼77.6%

across varying degrees of supervision, we employ 8 (10% supervision), and 16 (20% supervision) samples from the training dataset as labeled data, while treating the remainings as unlabeled data.

In the table presented, our proposal exhibits a notable advancement in Dice score, elevating from 78.96% to 89.94% when utilizing only 10% labeled data, showcasing a distinct advantage over alternative methodologies. With a subsequent increase in labeled data to 20%, DCF further boosts the performance to 91.44%, marking a notable gap of 4.57% compared to the baseline. In particular, DCF consistently outperforms the other competing approaches in both supervision settings, underscoring its superiority. Furthermore, for a more visually comprehensible depiction of segmentation outcomes, Figure 5 illustrates the results on the LA dataset. At first glance, it can be easily observed that our approach yields clearer segmentation boundaries and finer-grained details, aligning closely with the Ground Truth.

4.4 Result on Pancreas-CT Dataset

Table 2 shows the results specific to Pancreas-CT. Note that the volumes in this dataset provide a more complex backdrop compared to LA MRIs, rendering pancreas segmentation a more challenging task. To facilitate an intuitive comparison, we again employ several state-of-the-art competitors, namely UA-MT [42], SASSNet [17], DTC [20], MC-Net+ [37], RCPS [45], and CauSSL [22]. The performance metrics reported in their respective papers are directly adopted. Similarly, we employ Vnet with varying proportions of labeled data (10%, and 20%) for comparative analysis. Similarly, we select 6 samples for 10% supervision and 12 samples for 20% supervision from the training dataset, and consider the remainder as unlabeled data.

Yet within a notably challenging task, the proposed DCF demonstrates promising performance in both scenarios. Even with only 10% of the data labeled, DCF significantly improves Dice scores from 55.10% to 78.94%, surpassing all other SSL methods. With 20% labeled data, DCF achieves a Dice score of 81.65%, outperforming again other cutting-edge competitors. It should be noted that our proposal exhibits the smallest disparity between 10% and 20% labeled scenarios, as illustrated in Figure 4. The marginal 2.71% variance in DCF's Dice scores between these proportions suggests its effective utilization of unlabeled data, thereby demonstrating robustness and generalization capabilities. We further deliver visualization of the results obtained by DCF and others, as illustrated in Figure 6. It is observed that our results closely resemble the ground truth (GT) compared to those of other methods. Moreover, our method exhibits more precise boundary positioning and provides the more detailed information.

Table 3. Performance comparison with state-of-the-art methods on ISIC Dataset. The best results are in Bold font.

Methods	Volumes used		Metrics
	Labeled	Unlabeled	Dice(%)↑
MT (NIPS 2017)	90(5%)	1725	86.43
UA-MT (MICCAI 2019)	90(5%)	1725	87.02
CCT (CVPR 2020)	90(5%)	1725	84.48
CPS (CVPR 2021)	90(5%)	1725	86.79
UCMT (IJCAI 2023)	90(5%)	1725	88.22
DCF (ours)	90(5%)	1725	**88.88**
MT (NIPS 2017)	181(10%)	1634	86.97
UA-MT (MICCAI 2019)	181(10%)	1634	87.48
CCT (CVPR 2020)	181(10%)	1634	85.72
CPS (CVPR 2021)	181(10%)	1634	87.92
UCMT (IJCAI 2023)	181(10%)	1634	88.46
DCF (ours)	181(10%)	1634	**89.23**

4.5 Result on ISIC

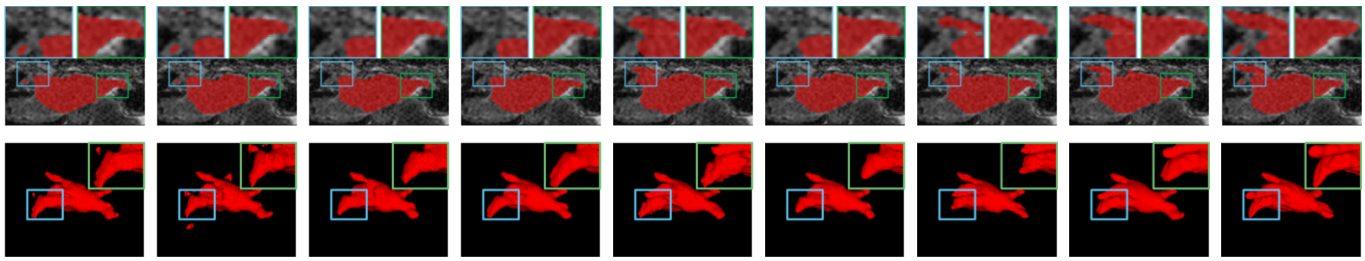
To further validate the generalizability of our proposed model, an additional verification is carried out using 2D images. Several state-of-the-art methods are re-implemented on the ISIC dataset, including MT [30], UA-MT [42], CCT [24], CPS [5], and UCMT [27]. The concerned results are presented in Table 3. Similarly, two scenarios with 5% and 10% labeled data are respectively established. As given, we observe that DCF outperforms the other methods, demonstrating its superior generalization capability

5 Ablation Study

Comprehensive ablation experiments are conducted on the final DCF architecture, validating the effectiveness of the tutoring mechanism and assessing the impact of various competitive approaches.

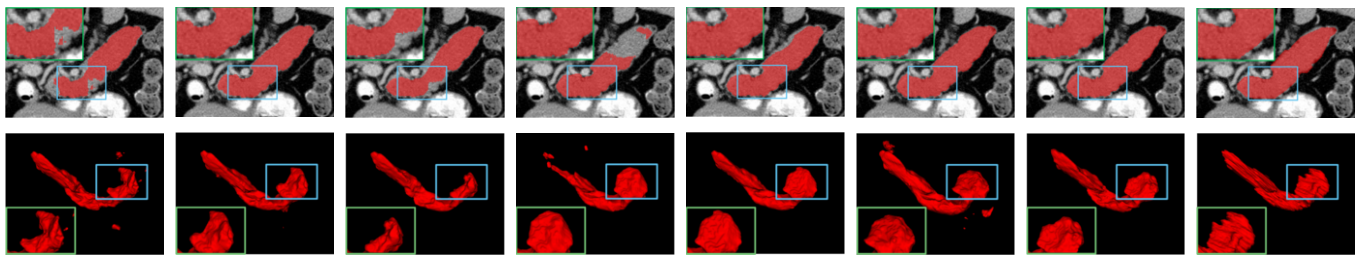
Impact of Competitive treatments: When evaluating student performance in the current iteration, we have multiple indicators to be chosen. In this experiment, we use Dice, CE, Jac, 95HD, and ASD, studying their individual and combined effects.

When 20% of labeled data is employed, Table 5 presents specific findings demonstrating that on the LA dataset, optimal outcomes are achieved with the utilization of Dice as the sole competitive metric. We believe that this is because the Dice metric outperforms other metrics in accurately evaluating model performance in the field of medical image segmentation. However, for varied tasks, it is also believed that alternative evaluation metrics should be contemplated.



(a) UAMT (b) SASSNet (c) DTC (d) MC-Net+ (e) FUSSNet (f) CAML (g) VSRC (h) Ours (i) Ground Truth

Figure 5. Comparison of visualization results on LA Dataset. The first row shows the results in 2D form, while the second row provides the visualizations in 3D form, where certain details have been enlarged for better clarity.



(a) UAMT (b) SASSNet (c) DTC (d) MC-Net+ (e) RCPS (f) CauSSL (g) Ours (h) Ground Truth

Figure 6. Comparison of visual results on Pancreas-CT Dataset. The first row shows the results in 2D form, while the second row provides the visualizations in 3D form, where certain details have been enlarged for better clarity.

Table 4. Ablations of different tutoring mechanisms on LA dataset.

Methods	Metrics			
	Dice(%)↑	Jaccard(%)↑	95HD(voxel)↓	ASD(voxel)↓
(1)	91.10	83.72	5.36	1.50
(2)	90.26	82.31	5.68	1.67
(3)	90.63	82.92	5.86	1.60
(4)	90.70	83.03	5.68	1.59
(5)	**91.44**	**84.28**	**5.24**	**1.55**

Effectiveness of Tutoring Mechanism. To assess the engineered tutoring mechanism, various scenarios are devised: 1) Teachers refraining from tutoring. 2) Teachers tutoring irrespective of performance. 3) Alternating tutoring duties among students. 4) Teachers offering extra support to high-performing students. 5) Providing tutoring to low-performing students. The experimental outcomes on the LA dataset (20% labeled data) are detailed in Table 4.

It is illustrated that the model exhibits optimal performance when the teacher administers remediation to poorly performing students, thereby validating the efficacy of our remediation treatment. However, the efficacy of the model diminishes when remediation is provided exclusively to well-performing students or when simultaneous remediation is administered to two students. This decline may be due to an exacerbated variance between students. Regarding the randomized remediation method, we contend that its indiscriminate nature undermines its ability to yield favorable outcomes. In instances where no tutoring mechanism is employed, the model neither exacerbates variance among students nor achieves optimization, potentially yielding subpar results.

6 Conclusion

In this study, we present a novel semi-supervised framework for 3D medical image segmentation, which is specifically designed to tackle the challenge of tight coupling in the Teacher-Student structure within MT-based methods. In addition, a competitive tutoring mechanism is crafted to improve communication between models, thus mitigating the risk of model collapse resulting from the acquisition of erroneous knowledge. Besides, we employ weight distance and prediction distance to perform a detailed analysis of the state changes among the three networks throughout the training process. The effectiveness of our DCF in semi-supervised medical image segmentation is validated on three public benchmark datasets. Furthermore, we believe that our proposed DCF framework can serve as a plug-and-play solution, readily applicable across diverse SSL fields. However, our method still suffers from certain limitations, *e.g.*, the dynamic interplay between two students is mostly pronounced during the initial and middle periods, which would gradually wane as time progresses. Moving forward, we intend to further investigate strategies to enhance consistency between students.

Table 5. Variations in model performance on LA Dataset under differing evaluation schemes during student competition.

Methods	Metrics			
	Dice(%)↑	Jaccard(%)↑	95HD(voxel)↓	ASD(voxel)↓
Dice	**91.44**	**84.28**	**5.24**	**1.55**
CE	90.50	82.70	6.02	1.56
Jac	89.32	80.79	7.43	2.21
ASD	90.80	83.20	5.51	1.52
95HD	90.72	83.07	5.75	1.60
Dice+Jac	89.07	80.40	6.56	2.15
Dice+CE	90.37	82.51	6.08	1.58
CE+Jac	89.44	81.01	7.66	2.20
95HD+ASD	91.05	83.62	5.60	1.48
CE+Jac+Dice	88.45	79.43	7.60	2.39

Acknowledgements

This work was supported in part by the Key Program of Natural Science Foundation of Zhejiang Provinceunder under Grant LZ24F030012, and in part by the National Natural Science Foundation of China under Grant 62276232.

References

[1] Y. Bai, D. Chen, Q. Li, W. Shen, and Y. Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proc.CVPR*, pages 11514–11524, 2023.

[2] H. Basak and Z. Yin. Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In *Proc.CVPR*, pages 19786–19797, 2023.

[3] P. Cascante-Bonilla, F. Tan, Y. Qi, and V. Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proc.AAAI*, volume 35, pages 6912–6920, 2021.

[4] D. Chen, Y. Bai, W. Shen, Q. Li, L. Yu, and Y. Wang. Magicnet: Semi-supervised multi-organ segmentation via magic-cube partition and recovery. In *Proc.CVPR*, pages 23869–23878, 2023.

[5] X. Chen, Y. Yuan, G. Zeng, and J. Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proc.CVPR*, pages 2613–2622, 2021.

[6] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *Proc.ISBI*, pages 168–172, 2018.

[7] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging*, 39(8):2626–2637, 2020.

[8] S. Gao, Z. Zhang, J. Ma, Z. Li, and S. Zhang. Correlation-aware mutual learning for semi-supervised medical image segmentation. In *Proc.MICCAI*, pages 98–108. Springer, 2023.

[9] K. Grünberg, O. A. J. del Toro, A. Jakab, G. Langs, T. S. Fernandez, M. Winterstein, M.-A. Weber, and M. Krenn. Annotating medical image data. In *Cloud-Based Benchmarking of Medical Image Analysis*, 2017.

[10] F. Huang, Z. Yao, and W. Zhou. Dtbs: Dual-teacher bi-directional self-training for domain adaptation in nighttime semantic segmentation. In *Proc.ECAI*, 2024.

[11] W. Huang, C. Chen, Z. Xiong, Y. Zhang, X. Chen, X. Sun, and F. Wu. Semi-supervised neuron segmentation via reinforced consistency learning. *IEEE Transactions on Medical Imaging*, 41(11):3016–3028, 2022.

[12] X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu. Missformer: An effective transformer for 2d medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(5):1484–1494, 2023.

[13] Z. Ke, D. Wang, Q. Yan, J. Ren, and R. W. Lau. Dual student: Breaking the limits of the teacher in semi-supervised learning. In *Proc.ICCV*, pages 6728–6736, 2019.

[14] D. Kwon and S. Kwak. Semi-supervised semantic segmentation with error localization network. In *Proc.CVPR*, pages 9957–9967, 2022.

[15] S. Laine and T. Aila. Temporal ensembling for semi-supervised learning. In *Proc.ICLR*, 2017.

[16] D.-H. Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. 2013. URL https://api.semanticscholar.org/CorpusID:18507866.

[17] S. Li, C. Zhang, and X. He. Shape-aware semi-supervised 3d semantic segmentation for medical images. In *Proc.MICCAI*, pages 552–561, 2020.

[18] C. Liu, C. Gao, F. Liu, P. Li, D. Meng, and X. Gao. Hierarchical supervision and shuffle data augmentation for 3d semi-supervised object detection. In *Proc.CVPR*, pages 23819–23828, 2023.

[19] Y. Liu, Y. Tian, Y. Chen, F. Liu, V. Belagiannis, and G. Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proc.CVPR*, pages 4258–4267, 2022.

[20] X. Luo, J. Chen, T. Song, and G. Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proc.AAAI*, volume 35, pages 8801–8809, 2021.

[21] F. Lyu, M. Ye, J. F. Carlsen, K. Erleben, S. Darkner, and P. C. Yuen. Pseudo-label guided image synthesis for semi-supervised covid-19 pneumonia infection segmentation. *IEEE Transactions on Medical Imaging*, 42(3):797–809, 2023. doi: 10.1109/TMI.2022.3217501.

[22] J. Miao, C. Chen, F. Liu, H. Wei, and P.-A. Heng. Caussl: Causality-inspired semi-supervised learning for medical image segmentation. *Proc.ICCV*, pages 21369–21380, 2023.

[23] S. Mittal, M. Tatarchenko, and T. Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1369–1379, 2019.

[24] Y. Ouali, C. Hudelot, and M. Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proc.CVPR*, pages 12674–12684, 2020.

[25] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc.MICCAI*, pages 234–241. Springer, 2015.

[26] H. R. Roth, L. Lu, A. A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *Proc.MICCAI*, page 556–564, 2015. URL https://api.semanticscholar.org/CorpusID:5776545.

[27] Z. Shen, P. Cao, H. Yang, X. Liu, J. Yang, and O. R. Zaiane. Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation. In *Proc.IJCAI*, pages 4199–4207, 2023.

[28] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Proc.NeurIPS*, volume 33, pages 596–608, 2020.

[29] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, and C. Lippert. 3d self-supervised methods for medical imaging. In *Proc.NeurIPS*, volume 33, pages 18158–18172, 2020.

[30] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc.NeurIPS*, page 1195–1204, 2017.

[31] G. Wang, M. A. Zuluaga, W. Li, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, et al. Deepigeos: a deep interactive geodesic framework for medical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1559–1572, 2018.

[32] G. Wang, S. Zhai, G. Lasio, B. Zhang, B. Yi, S. Chen, T. J. Macvittie, D. Metaxas, J. Zhou, and S. Zhang. Semi-supervised segmentation of radiation-induced pulmonary fibrosis from lung ct scans with multi-scale guided dense attention. *IEEE Transactions on Medical Imaging*, 41(3):531–542, 2021.

[33] T. Wang, J. Lu, Z. Lai, J. Wen, and H. Kong. Uncertainty-guided pixel contrastive learning for semi-supervised medical image segmentation. In L. D. Raedt, editor, *Proc.IJCAI*, pages 1444–1450, 7 2022.

[34] Y. Wang, Y. Zhang, J. Tian, C. Zhong, Z. Shi, Y. Zhang, and Z. He. Double-uncertainty weighted method for semi-supervised learning. In *Proc.MICCAI*, pages 542–551, 2020.

[35] Y. Wang, H. Chen, Q. Heng, W. Hou, Y. Fan, Z. Wu, J. Wang, M. Savvides, T. Shinozaki, B. Raj, B. Schiele, and X. Xie. Freematch: Self-adaptive thresholding for semi-supervised learning. In *Proc.ICLR*, 2023.

[36] H. Wu, Z. Wang, Y. Song, L. Yang, and J. Qin. Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images. In *Proc.CVPR*, pages 11656–11665, 2022.

[37] Y. Wu, Z. Ge, D. Zhang, M. Xu, L. Zhang, Y. Xia, and J. Cai. Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81:102530, 2022.

[38] Y. Xia, D. Yang, Z. Yu, F. Liu, J. Cai, L. Yu, Z. Zhu, D. Xu, A. Yuille, and H. Roth. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. *Medical image analysis*, 65:101766, 2020.

[39] J. Xiang, P. Qiu, and Y. Yang. Fussnet: Fusing two sources of uncertainty for semi-supervised medical image segmentation. In *Proc.MICCAI*, pages 481–491, 2022.

[40] Z. Xiong, Q. Xia, Z. Hu, N. Huang, S. Vesal, N. Ravikumar, A. Maier, C. Li, Q. Tong, W. Si, et al. A global benchmark of algorithms for segmenting late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical Image Analysis*, page 101832, 2020.

[41] Z. Yang and S. Farsiu. Directional connectivity-based segmentation of medical images. In *Proc.CVPR*, pages 11525–11535, 2023.

[42] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *Proc.MICCAI*, pages 605–613, 2019.

[43] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Proc.NeurIPS*, 2021.

[44] X. Zhao, C. Fang, D.-J. Fan, X. Lin, F. Gao, and G. Li. Cross-level contrastive learning and consistency constraint for semi-supervised medical image segmentation. In *Proc.ISBI*, pages 1–5. IEEE, 2022.

[45] X. Zhao, Z. Qi, S. Wang, Q. Wang, X. Wu, Y. Mao, and L. Zhang. Rcps: Rectified contrastive pseudo supervision for semi-supervised medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2023.

[46] Y. Zhao, K. Lu, J. Xue, S. Wang, and J. Lu. Semi-supervised medical image segmentation with voxel stability and reliability constraints. *IEEE Journal of Biomedical and Health Informatics*, 27(8):3912–3923, 2023. doi: 10.1109/JBHI.2023.3273609.

[47] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 2019.