PMT: Progressive Mean Teacher via Exploring Temporal Consistency for Semi-Supervised Medical Image Segmentation

Ning Gao, Sanping Zhou*, Le Wang, and Nanning Zheng

National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Applications, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

Abstract. Semi-supervised learning has emerged as a widely adopted technique in the field of medical image segmentation. The existing works either focuses on the construction of consistency constraints or the generation of pseudo labels to provide high-quality supervisory signals, whose main challenge mainly comes from how to keep the continuous improvement of model capabilities. In this paper, we propose a simple yet effective semi-supervised learning framework, termed **P**rogressive **M**ean **T**eachers (PMT), for medical image segmentation, whose goal is to generate high-fidelity pseudo labels by learning robust and diverse features in the training process. Specifically, our PMT employs a standard mean teacher to penalize the consistency of the current state and utilizes two sets of MT architectures for co-training. The two sets of MT architectures are individually updated for prolonged periods to maintain stable model diversity established through performance gaps generated by iteration differences. Additionally, a difference-driven alignment regularizer is employed to expedite the alignment of lagging models with the representation capabilities of leading models. Furthermore, a simple yet effective pseudo-label filtering algorithm is employed for facile evaluation of models and selection of high-fidelity pseudo-labels outputted when models are operating at high performance for co-training purposes. Experimental results on two datasets with different modalities, i.e., CT and MRI, demonstrate that our method outperforms the state-of-the-art medical image segmentation approaches across various dimensions. The code is available at https://github.com/Axi404/PMT.

Keywords: Semi-supervised learning · Medical image segmentation · Temporal consistency regularization

1 Introduction

Semi-supervised learning is an important field in deep learning, which offers an effective way to tackle problems with limited labeled data [16, 25, 27, 32, 33]. With the continuous emergence of large-scale data, semi-supervised learning has

* Corresponding author.

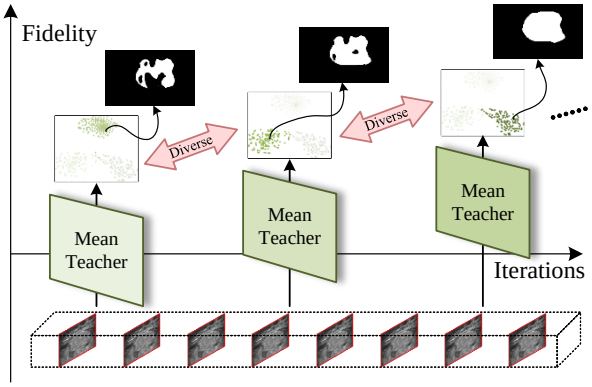


Fig. 1: Motivation of our PMT. In particular, the standard MT helps to learn robust features by keeping the consistency between teacher and student networks at the current iteration, while our PMT further helps to learn diverse features by maintaining the difference between student networks at different iterations. As a result, more and more high-fidelity pseudo labels will be generated for semi-supervised medical image segmentation.

become a hot topic in both machine learning and computer vision communities. It has gained particular attention in the medical image segmentation [21], due to the need for expert annotation in determining the accurate boundaries of targets. As a result, the labeled data is very lacking in medical image segmentation, which makes the semi-supervised learning very popular in this domain.

Most of the semi-supervised learning methods involve seeking a guidance mechanism using limited annotated data, so as to maximize the usage of unlabeled data to train the models on different downstream tasks. In general, two technologies, *i.e.*, consistency regularization [9, 10] and pseudo label generation [10, 19], are commonly used in semi-supervised medical segmentation. The former ones aim at enhancing the representation capability of networks by exploring the consistency based on different prior assumptions [5], while the latter ones focus on improving the performance on downstream tasks by generating the pseudo labels based on current model capability. For example, the well-known Mean Teacher (MT) [19] utilizes an ingenious Exponential Moving Average (EMA) that is equivalent to data augmentation and enforces consistency regularization between the outputs of teacher and student models. Besides, the representative Noisy Student [26] designs a teacher model to generate pseudo labels, so as to train a larger student model.

Even though the above methods have achieved significant improvements in semi-supervised medical image segmentation, we still argue that they are weak in exploring high-quality supervisory signals to consistently enhance the model's capability. To address this issue, an iterative unified optimization framework has been introduced to semi-supervised medical image segmentation [20, 23, 24], in which the consistency regularization strategy is taken to enhance the model's

representation capability and the pseudo label generation algorithm is applied to generate high-fidelity pseudo labels. For example, the recent MCF [23] takes VNet and 3D ResNet for representation learning, and heterogeneous networks for dynamic pseudo label generation. The challenges to these methods lies on how to continuously generate diverse pseudo labels in the forward propagation, and enhance model's capability in the backward propagation.

In this paper, we design a novel semi-supervised learning framework, termed Progressive Mean Teachers (PMT), for medical image segmentation, whose main idea focuses on how to obtain a diverse set of accurate pseudo labels. Inspired by the positive relationship between iteration and performance, we try to maintain the diversity of networks by exploring their states at different training epochs. Specifically, the standard MT is first taken as basic architecture to learn the parameters of network at each iteration, which can enhance the network's representation capability by using the EMA data augmentation. Our PMT further explores network diversity during training, alternating between two homogeneous MT architectures trained on the same dataset, a process we term progressive design. These models exhibit significant iteration leads due to alternating continuous individual updates, establishing performance gaps between networks at different epochs, as shown in **Fig. 1**. Consequently, the student network can acquire robust yet diverse features for medical image segmentation. Then, the Discrepancy Driven Alignment (DDA) regularizer is further designed to examine disparities between predictions obtained by the student network at different training epochs, facilitating rapid alignment to high-fidelity generated images. Finally, we design a simple Pseudo Label Filtering (PLF) algorithm to refine the basic interaction process, enabling the retention of high-fidelity pseudo-labels for training by comparing student network performance across different training epochs. As a result, more and more high-fidelity pseudo labels can be fed to train the other student networks, which will in turn to generate more accurate predictions for pseudo label generation.

In summary, the main contributions of this work are as follows: (1) We design a novel Progressive Mean Teacher framework for semi-supervised medical image segmentation. (2) We design a novel Discrepancy Driven Alignment regularizer to rapidly align the representational capacity gap between lagging and leading networks. (3) We design a simple yet effective Pseudo Label Filtering algorithm to select high-fidelity pseudo labels. Extensive experiments in both Left Atrial [28] and Pancreas-NIH [17] datasets show that our PMT can achieve the state-of-the-art results in semi-supervised medical image segmentation.

2 Related Work

2.1 Consistency Regularization

Consistency regularization is often employed in semi-supervised learning, so as to enhance the stable representational capacity of model. The behind idea is to preserve the invariance of predictions made by the same model when facing

perturbations applied in different regions, by constraining the consistency between output results under different perturbations. For example, Π model [9] introduces image-level regularization by adding perturbations to images. The well-known MT [19] introduces parameter-level regularization by using EMA to constrain the outputs between teacher and student models. Thanks to its simplicity and effectiveness, more and more works are paying attention to improve the generalization ability by formulating different consistency regularizers. For example, SASSnet [11] focuses on the regularity of geometric shapes for target object classes within consistency regularization. Besides, CPCL [29] establishes regularization between supervised and unsupervised training within a cyclic framework. Furthermore, some later works have made progress in regularization at different task and model levels. For example, DTC [12] introduces task-level regularization, presenting a novel dual-task consistency semi-supervised framework. Besides, MCF [23] introduces model-level regularization by using heterogeneous models to constrain output consistency. Compared to previous work, our PMT seeks the cross-temporal regularization between different training periods, which can help learn diverse yet robust features for subsequent pseudo label generation.

2.2 Pseudo Label Generation

Pseudo label generation is often employed in semi-supervised learning, so as to enhance the discriminative ability of model. The behind idea is to train a prior model with the labeled data and then apply it to generate the pseudo labels for unlabeled data. It is widely accepted to categorize pseudo-label generation into two methods [7], direct generation focuses on selecting pseudo labels with higher confidence, while indirect generation explores methods to generate high-fidelity pseudo labels. For direct generation, work [10] uses a fixed threshold to choose high-confidence pseudo labels. SsaNet [22] employs a trust module to reevaluate pseudo labels. UA-MT [30] introduces uncertainty estimation to filter out unreliable pseudo labels. Co-BioNet [15] introduces a feedback network to measure the uncertainty and choose high-confidence predictions of different models. For indirect generation, Tri-Net [6] is proposed to use two subnetworks to generate pseudo labels for a third subnetwork. Work [1] improves pseudo label generation using Simple Linear Iterative Clustering (SLIC) algorithm. MCF [23] dynamically generates pseudo labels using a heterogeneous network and effectively addresses cognitive bias, while DeSCO [3] focuses on the spatial correlation of medical images and generates pseudo labels using orthogonal slices. Compared to previous work, our PMT emphasizes the enhancement of pseudo label quality in the temporal domain and generates diverse and robust pseudo labels across different training periods, significantly improving performance.

2.3 Multi-Model Framework

The multi-model framework is often employed in semi-supervised learning, so as to enhance the model representation by acquiring multiple views or diversity. Its development aligns closely with consistency regularization. Here, we

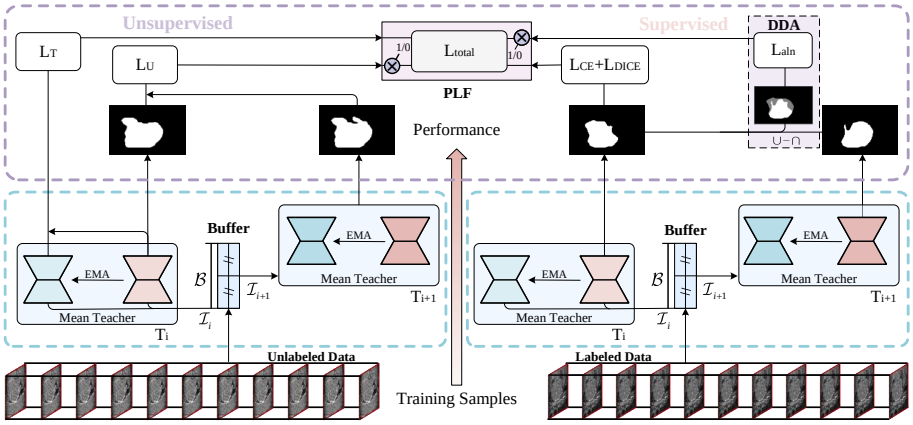


Fig. 2: An overview of PMT. We employed a progressive design and utilized the architecture of MT. The PMT framework maintains a data buffer of length $\mathcal{B}$ for cross-temporal training. The total loss function L_{total} for each network includes supervised losses $L_{CE}, L_{DICE}, L_{aln}$, and unsupervised loss L_U, L_T.

focus on the evolution of model structures. For example, MT [19] introduces a Teacher model with significantly improved representation capabilities at a lower cost using EMA, establishing the MT architecture. Besides, CPC [8] utilizes the confidence vector of multi-model outputs as pseudo-labels for co-training, while CPS [4] employs one-hot labels, both considered as starting points for co-training. In recent years, efforts have been made to combine these approaches. For example, UCMT [18] employs two student models for co-training, simultaneously updating the same Mean Teacher using EMA, while Dual Teacher [14] alternates updates between two Teacher models using EMA with a single student model, thereby fostering diversity among Teacher models. Compared to previous work, our PMT utilizes two sets of Mean Teachers in a progressive training framework, and employ two student models to update teacher models independently. This enables our model to rapidly establish robust performance disparities across iteration gap and maintain stable diversity among models.

3 Method

3.1 The Overall Process of PMT

The training process of PMT is illustrated in **Fig. 2**. As previously mentioned, PMT can consist of multiple networks. For ease of explanation, the number of networks is assumed to be two by default in our work. In our approach, these networks share the same structure and can be denoted as $f_i(\cdot) \in \mathcal{F}$, where $\mathcal{F}$ stands for the function space within which these networks reside. In a semi-supervised process, the training data includes a small number of labeled data denoted as $\mathbf{D}_L = \{(x_i^L, y_i^L)\}_{i=1}^N$, and a large amount of unlabeled data denoted

as $\mathbf{D}_U = \{(x_i^U)\}_{i=N+1}^{N+M}$, where $N \ll M$, $x_i \in \mathbb{R}^{H\times W\times D}$ represents medical volumes, and $y_i \in \{0,1\}^{H\times W\times D}$ represents ground truth labels. Batches of input data $\mathbf{X}$ consist of an equal proportion of labeled data $(\mathbf{X}^L, \mathbf{Y}^L)$ and unlabeled data $\mathbf{X}^U$. These volumes are fed into $f_i(\cdot)$ and $f_{i+1}(\cdot)$:

$$\hat{\mathbf{Y}}_i = f_i(\mathbf{X}), \quad \hat{\mathbf{Y}}_{i+1} = f_{i+1}(\mathbf{X}). \tag{1}$$

The output consists of volume predictions for both labeled and unlabeled data: $\hat{\mathbf{Y}} = \hat{\mathbf{Y}}^L \cup \hat{\mathbf{Y}}^U$. For the sake of simplicity, the network index subscripts are omitted here.

We have ingeniously reintroduced MT architecture into the framework of semi-supervised learning. The introduction serves the purpose that, while providing high-fidelity pseudo labels to the model, the model's architecture needs to support the stable improvement of its representational capacity, aiming for better performance, while maintaining stable diversity. The teacher network is structurally identical to the student network but does not actively participate in the training process. Instead, it updates all of its parameters through EMA and applies consistency regularization to the student model through its output. The parameter is updated as follows:

$$\theta_{\text{teacher}} = \alpha\theta_{\text{teacher}} + (1-\alpha)\theta_{\text{student}}, \tag{2}$$

where θ_{teacher} represents the parameters of teacher network, θ_{student} represents the parameters of student network, and α is the EMA decay rate.

In practice, we propose a cross-temporal training approach, which involves introducing a phase shift among different models during iterations and training the most lagging model in terms of iterations. To ensure that each model can train across temporal, the number of iterations for the model recorded in the sequence of iterations is denoted as $\mathcal{I}_i$. Our model maintains a static iteration gap through the data buffer length $\mathcal{B}$:

$$\forall i \in [1, n-1], \text{Lar}(\mathcal{I}, i) - \text{Lar}(\mathcal{I}, i+1) = \frac{\mathcal{B}}{(n-1)}, \tag{3}$$

where, $\text{Lar}(\mathcal{I}, i)$ represents the i-th largest number in $\mathcal{I}$.

During each training iteration, the model with the least advanced iteration progress, referred to as the Current Progressive Model (CPM), sequentially utilizes data from the buffer for training until it has advanced ahead of the model with the most advanced iteration progress $\mathcal{B}/n$ times. Throughout the training process, the PLF screens out pseudo-labels generated by models whose performance is inferior to that of the CPM by measuring the performance gap between the CPM and other models, and learns from the remaining pseudo-labels. Simultaneously, the model computes the DDA with respect to a given input $\mathbf{X}$ and, along with all other models, examines regions of inconsistent predictions among different networks. When the performance of the CPM lags behind that of other models, the DDA corrects the CPM, enabling it to quickly align its performance with models that outperform it. It is worth noting that if the CPM is the best-performing model for that iteration, it will not guide other models in reverse. It

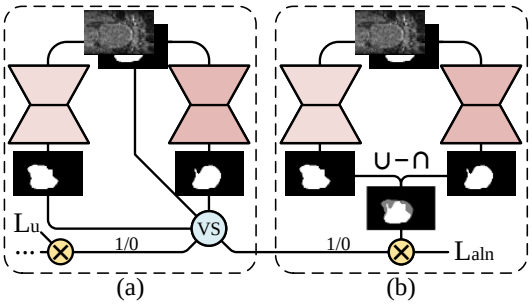


Fig. 3: Illustration of PLF and DDA. PLF is utilized for comparing the representational capacity of models, while DDA aligns model outputs by examining differences.

receives guidance during the training process and only guides others once they have completed training.

3.2 Pseudo Label Filtering

To guide the learning of the current progressive model effectively, we introduce PLF, as shown in **Fig. 3** (a). PLF involves a naive approximation where we consider the representational capacity of models during the supervised learning phase as a criterion for judging the representational capacity during the unsupervised learning phase to some extent. During the supervised learning phase, PLF entails inference by all models, and through combining labels, calculates the representational capacity of each model in the current iteration. The metrics for evaluating representational capacity vary with tasks; in this task, the Dice loss is regarded as a statistical measure of model performance. We utilize the representational capacity of the CPM as a threshold and filter out pseudo-labels generated by models with representational capacities below that of the CPM. The pseudo label's passage through the PLF is recorded as $\mathcal{P}(f_i, f_{i+1})$, where a value of 1 denotes passage and 0 denotes exclusion. This aids in the model's better learning of pseudo-labels with high fidelity. Furthermore, to ensure that the pseudo-labels are accurate and well-defined, we apply a sharpening function to high-fidelity pseudo-labels:

$$\mathbf{Y}^U = \frac{{\mathbf{P_s}}^{1/T}}{{\mathbf{P_s}}^{1/T} + (1 - \mathbf{P_s})^{1/T}}, \tag{4}$$

where $\mathbf{P_s}$ represents the output of the good student, and T is a hyperparameter.

The resulting loss function $\mathrm{L_U}$ is computed as MSE between the pseudo labels $\mathbf{Y}^U$ and the model's predictions of CPM $\hat{\mathbf{Y}}_c^U$:

$$\mathrm{L_U} = \mathrm{MSE}(\hat{\mathbf{Y}}_c^U, \mathbf{Y}^U). \tag{5}$$

3.3 Discrepancy Driven Alignment

We propose DDA to compare the outputs of multiple models and achieve rapid alignment between models, as shown in **Fig. 3** (b). During the alternating process of progressive Learning, the performance of two groups of models alternately improves, providing stable pseudo-labels to each other. However, concerns often arise due to the disappearance of the performance gap between models caused by random factors. While the expectation is for alternating performance leadership between models, in practice, it may turn into one model consistently leading. Therefore, a fast and high-confidence alignment module is needed. In traditional consistency regularization, contrastive algorithms enforce consistency among the outputs of different models. However, this approach often raises concerns because we are uncertain whether this consistency regularization guides the models in the right direction. DDA focuses on the different parts of predictions among models, which often reflect the differences in representational capabilities between models. We aim to enhance the consistency regularization specifically on these parts to achieve alignment of representational capabilities between models. It is worth noting that we expect the models aligned by DDA to have representation capabilities surpassing CPM. Therefore, DDA will only take effect when PLF allows other models to provide pseudo-labels. The mask of different parts of predictions among different models can be obtained by taking the difference between the union and intersection of the binary outputs of softmax outputs $\hat{\mathbf{Y}}^L$ of multiple networks:

$$\mathcal{M}_{\text{diff}} = \bigcup_{i=1}^{n} \text{BINA}(\hat{\mathbf{Y}}_i^L) - \bigcap_{i=1}^{N} \text{BINA}(\hat{\mathbf{Y}}_i^L). \tag{6}$$

Subsequently, MSE of the model's predictions is computedand denoted as $\mathcal{M}_{\text{dist}}$:

$$\mathcal{M}_{\text{dist}} = \text{MSE}\left(\hat{\mathbf{Y}}_i^L, \mathbf{Y}^L\right). \tag{7}$$

Next, the total count of elements in $\mathcal{M}_{\text{diff}}$ is calculated, and $\mathcal{M}_{\text{diff}}$ is used to mask $\mathcal{M}_{\text{dist}}$. The ratio of the sum of the masked $\mathcal{M}_{\text{dist}}$ to the total count of elements in $\mathcal{M}_{\text{diff}}$ is denoted as L_{aln}, which characterizes the gap to the ground truth within the prediction differences between two models:

$$\text{L}_{\text{aln}} = \frac{\sum(\mathcal{M}_{\text{diff}} \cdot \mathcal{M}_{\text{dist}})}{\sum(\mathcal{M}_{\text{diff}})}. \tag{8}$$

3.4 Loss Function

In general, during an iteration cycle, CPM conducts the process of backward propagation concerning its loss function, which is defined as follow:

$$\text{L}_{\text{total}} = \text{L}_{\text{s}} + \lambda_1 \mathcal{P}(f_i, f_{i+1})\text{L}_{\text{U}} + \lambda_2 \text{L}_{\text{T}}, \tag{9}$$

where λ_1 and λ_2 increase as the number of iterations grows, up to a point where they stop growing after a fixed iterations.

We employed two independent Gaussian warm-up function to control the loss function weights, λ_1 and λ_2, using different parameters:

$$\begin{aligned}\lambda_1(t) &= \begin{cases}\hat{\lambda}_1 \cdot e^{-5\left(1-\frac{2t}{t_{\max}}\right)^2}, & t < \frac{t_{\max}}{2} \\ \hat{\lambda}_1, & t \geq \frac{t_{\max}}{2}\end{cases} \\ \lambda_2(t) &= \begin{cases}\hat{\lambda}_2 \cdot e^{-5\left(1-\frac{2t}{t_{\max}}\right)^2}, & t < \frac{t_{\max}}{2} \\ \hat{\lambda}_2, & t \geq \frac{t_{\max}}{2}\end{cases}\end{aligned} \tag{10}$$

where t represents the current iteration number, and $t_{\max}$ represents the total number of training iterations. The hyperparameter $\hat{\lambda}_1$ and $\hat{\lambda}_2$ were empirically set to 20.0 and 10.0.

Loss function $\mathrm{L_s}$ represents the loss during supervised learning and consists of the following components:

$$\mathrm{L_s} = \mathrm{CE} + \mathrm{DICE} + \beta \mathcal{P}(f_i, f_{i+1}) \mathrm{L_{aln}}(\mathcal{M}_{\mathrm{diff}}, \mathcal{M}_{dist}), \tag{11}$$

where β, a constant set to 0.5, is used to balance alignment loss and other losses.

Loss function $\mathrm{L_T}$ is the consistency loss generated by the MT, which is derived by the teacher through the sharpen function to convert the model outputs into pseudo-labels, and subsequently generated by computing the MSE with the student model:

$$\mathrm{L_T} = \mathrm{MSE}\left(f_i\left(\mathbf{X}^U\right), \frac{{\mathbf{P_t}}^{1/T}}{{\mathbf{P_t}}^{1/T} + (1-\mathbf{P_t})^{1/T}}\right). \tag{12}$$

4 Experiments

4.1 Implementation Details

We selected the VNet [13] model as a baseline network, which performs well in conditions with limited data and is essentially a 3D convolutional version of UNet. During inference, we use the average of the outputs from two networks as the final prediction. Specifically, the SGD optimizer was used to update the network parameters with weight decay of 0.0001 and a momentum of 0.9. The initial learning rate was set to 0.01, divided by 10 every 2500 iterations, for a total of 6000 iterations.

Following the practice in comparative literature [11, 12, 23, 30], our methods are trained for a fixed number of 6,000 iterations to obtain the final model. Additionally, our models all use a batch size of 4, with a labeled data quantity of 2. We tested the performance of the models, and all experiments were conducted on NVIDIA® GeForce A40 48GB running Ubuntu 20.04 and PyTorch 1.11.0.

4.2 Datasets and Metrics

In the experiment, we selected two datasets with different modalities and utilized four distinct metrics to assess the performance of the model. For each dataset, 80% of the data was used as the training set and 20% as the test set. The proportion of supervised data was determined based on the training set.

LA Dataset. It [28] includes 100 3D gadolinium-enhanced MR imaging volumes of left atrial with an isotropic resolution of $0.625 \times 0.625 \times 0.625 mm^3$ and the corresponding ground truth labels. For pre-processing, we first normalize all volumes to zero mean and unit variance, then crop each 3D MRI volume with enlarged margins according to the targets. During training, the training volumes are randomly cropped to $112 \times 112 \times 80$ as the model input. During inference, a sliding window of the same size is used to obtain segmentation results with a stride of $18 \times 18 \times 4$.

Pancreas-NIH Dataset. It [17] provides 82 contrast-enhanced abdominal 3D CT volumes of pancreas with manual annotation. The size of each CT volume is $512 \times 512 \times \mathrm{D}$, where $\mathrm{D} \in [181, 466]$. In pre-processing, we use the soft tissue CT window of $[-120, 240]$ HU, and we crop the CT scans centering at the pancreas region, and enlarge margins with 25 voxels. The training volumes are randomly cropped to $96 \times 96 \times 96$ as the model input. During inference, a sliding window of the same size is used to obtain segmentation results with a stride of $16 \times 16 \times 16$.

Metrics. Following [2, 11, 12, 23, 29, 30], we use four metrics to evaluate model performance, including regional sensitive metrics: Dice similarity coefficient (Dice) [30], Jaccard similarity coefficient (Jaccard) [12], and edge sensitive metrics: 95% Hausdorff Distance (95HD) [29] and Average Surface Distance (ASD) [2].

4.3 Ablation Study

For simplicity, we conducted ablation experiments on LA dataset to evaluate our design choices for each component. For a reliable assessment, all other parts were kept consistent except for the component under investigation.

Analysis of PMT Framework. The PMT architecture under progressive design is at the core of our work, wherein progressive design ensures the continuous generation of diverse pseudo-labels during the forward propagation process. Within the PMT architecture, PLF and DDA respectively denote the pseudo-label filtering and model alignment methods proposed in our paper, while MT ensures stable enhancement of model capability. We employ a non-progressive design co-training framework consisting of two VNet components as the baseline model. When adding PLF and DDA methods to the baseline model, the progressive design is simultaneously incorporated into the baseline model. Finally, we evaluate the impact of the MT architecture on model representation capability, with results presented in **Table 1**. Overall, compared to the baseline model, which already possesses good representation capability, the PMT framework greatly enhances the model's representation capability.

Table 1: Ablation results about **PMT framework** on LA dataset

Method			Labeled	Metrics			
PLF	DDA	MT		Dice↑	Jaccard↑	95HD↓	ASD↓
-	-	-	8(10%)	88.28	79.31	7.59	2.34
✓	-	-	8(10%)	89.90	81.73	6.14	1.72
-	✓	-	8(10%)	90.43	82.60	5.98	1.66
✓	✓	-	8(10%)	90.43	82.60	**5.49**	**1.49**
✓	✓	✓	8(10%)	**90.81**	**83.23**	5.61	1.50

Analysis of Regularization Strength. The hyperparameters $\hat{\lambda}_1$ and $\hat{\lambda}_2$ respectively characterize the regularization strength of the progressive methods and MT architecture on the model. We use a tenfold scale and scale $\hat{\lambda}_1$ and $\hat{\lambda}_2$ based on the hyperparameter specifications we use, as shown in **Table 2**. The results indicate that variations in the two parameters within a certain range do not significantly affect the model's performance, demonstrating a certain level of robustness of our model to parameter variations. Within a certain range, on the LA dataset with 10% labeled data, our model achieves the best performance when $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are 20.0 and 10.0 respectively.

Table 2: Ablation results about **regularization strength** on LA dataset

λ		Labeled	Metrics			
λ_{1max}	λ_{2max}		Dice↑	Jaccard↑	95HD↓	ASD↓
2.0	1.0	8(10%)	89.66	81.33	6.35	1.84
2.0	10.0	8(10%)	90.47	82.66	**5.58**	1.64
2.0	100.0	8(10%)	87.74	78.28	9.33	2.84
20.0	1.0	8(10%)	90.67	82.98	6.31	1.77
20.0	10.0	8(10%)	**90.81**	**83.23**	5.61	**1.50**
20.0	100.0	8(10%)	88.00	78.73	9.05	2.78
200.0	1.0	8(10%)	89.08	80.42	11.40	3.07
200.0	10.0	8(10%)	89.75	81.47	7.08	2.22
200.0	100.0	8(10%)	83.80	72.55	16.44	5.11

4.4 Comparison with Other Methods

We compared our approach with previous state-of-the-art methods on LA dataset and Pancreas-NIH dataset.

We chose VNet as baseline models for comparison. For the selected alternative models, we opted for UA-MT [30] with uncertainty estimation, SASSNet [11] focusing on the regularity of geometric shapes, DTC [12] with task-level regularization, BCP [2] using bidirectional CutMix [31], and MCF [23] with model-level

regularization, with BCP and MCF being state-of-the-art results. Noting that, for BCP, we follow its parameter settings for pre-training 2,000 times and self-training 15,000 times.

Comparison on LA Dataset. We conducted a cross-model comparison on the classic LA dataset. We tested with 5% and 10% of labeled data. Results of the experiments are presented in **Table 3**. To provide a more intuitive demonstration of the performance of various models on the LA dataset, we have selected some representative results for visualization, as illustrated in **Fig. 4**. Areas with inaccurate segmentation have been annotated accordingly.

Table 3: Comparison results on LA dataset with 5% and 10% labeled data

Method	Labeled	Metrics			
		Dice↑	Jaccard↑	95HD↓	ASD↓
VNet	4(5%)	52.55	39.60	47.05	9.87
UA-MT	4(5%)	82.26	70.98	13.71	3.82
SASSNet	4(5%)	81.60	69.63	16.16	3.58
DTC	4(5%)	81.25	69.33	14.90	3.99
MCF^{SOTA}	4(5%)	-	-	-	-
BCP^{SOTA}	4(5%)	88.02	78.72	7.90	2.15
PMT(Ours)	4(5%)	**89.47**	**81.04**	**6.45**	**1.86**
VNet	8(10%)	82.74	71.72	13.35	3.26
UA-MT	8(10%)	86.28	76.11	18.71	4.63
SASSNet	8(10%)	85.22	75.09	11.18	2.89
MCF^{SOTA}	8(10%)	88.71	80.41	6.32	1.90
BCP^{SOTA}	8(10%)	89.62	81.31	6.81	1.76
PMT(Ours)	8(10%)	**90.81**	**83.23**	**5.61**	**1.50**

Our model outperformed previous models on all four metrics across 5% and 10% labeled data. At 5% of the data, compared to the best results from previous work, our PMT showed a 1.45% improvement in Dice, a 2.32% improvement in Jaccard, a reduction of 1.45 in 95HD, and a reduction of 0.29 in ASD. Our PMT also showed a 1.19% improvement in Dice, a 1.92% improvement in Jaccard, a reduction of 1.20 in 95HD, and a reduction of 0.26 in ASD at 10% of the data. It is noteworthy that our model maintains good performance even with a small amount of data. In comparison with state-of-the-art results, our model achieves a leading performance in the majority of metrics using only half of their data. For instance, we surpass MCF's performance at 10% labeled data using only 5% labeled data. The above indicates that our model achieves superior results compared to existing methods on differently proportioned annotated data in the task of segmenting the left atrium.

Comparison on Pancreas-NIH Dataset. We conducted a cross-model comparison on the classic Pancreas dataset. Detailed results of the experiments are presented in **Table 4**. We tested with 10% and 20% of labeled data. To

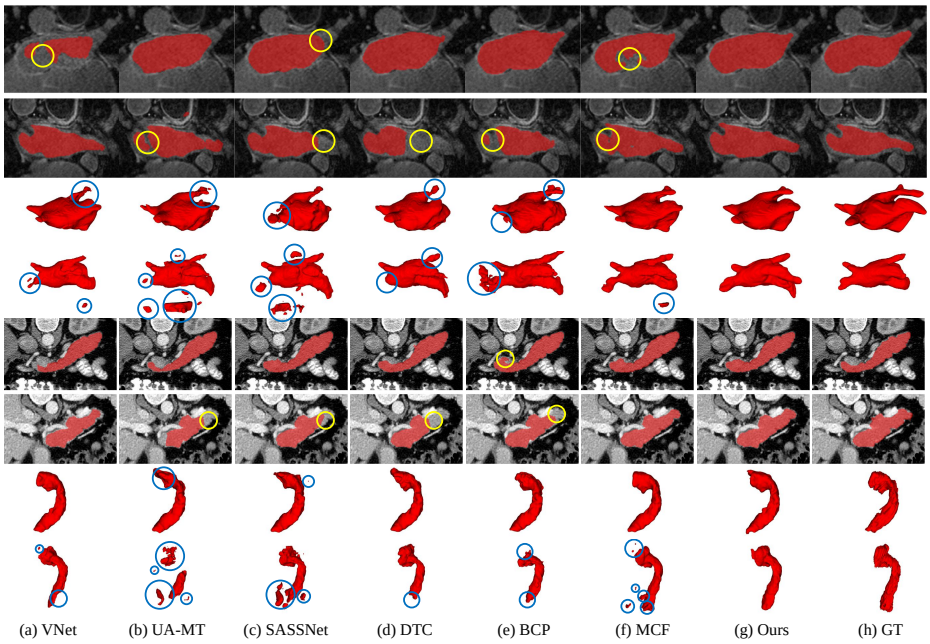


Fig. 4: 2D & 3D segmentation visualization of different semi-supervised methods under 10% labeled on LA (upper) and pancreas (bottom) dataset.

provide a more intuitive demonstration of the performance of various models on the Pancreas-NIH dataset, we have selected some representative results for visualization, as illustrated in **Fig.** 4. Areas with inaccurate segmentation have been annotated accordingly. It is worth noting that the results in the table clearly indicate that Pancreas-NIH dataset is significantly more challenging than LA dataset. Therefore, we increased $\hat{\lambda}_1$ by a factor of two, resulting in improved performance.

Our model outperformed previous models on all four metrics across 10% and 20% labeled data. At 10% of the data, compared to the best results from previous work, our PMT showed a 7.17% improvement in Dice, a 9.09% improvement in Jaccard, a reduction of 6.35 in 95HD, and a reduction of 2.10 in ASD. Our PMT also showed a 0.31% improvement in Dice, a 0.55% improvement in Jaccard, and a reduction of 0.36 in ASD at 20% of the data. The above indicates that our model achieves superior results compared to existing methods on differently proportioned annotated data in the task of segmenting the pancreas.

5 Conclusion

In this paper, we propose a semi-supervised medical image segmentation framework named Progress Mean Teacher (PMT). PMT adopts a progressive

Table 4: Comparison results on Pancreas-NIH dataset with 10% and 20% labeled data

Method	Labeled	Metrics			
		Dice↑	Jaccard↑	95HD↓	ASD↓
VNet	6(10%)	55.60	41.74	45.33	18.63
UA-MT	6(10%)	66.34	53.21	17.21	4.57
SASSNet	6(10%)	68.78	53.86	19.02	6.26
DTC	6(10%)	69.21	54.06	17.21	5.95
BCP^{SOTA}	6(10%)	73.83	59.24	12.71	3.72
MCF^{SOTA}	6(10%)	-	-	-	-
PMT(Ours)	6(10%)	**81.00**	**68.33**	**6.36**	**1.62**
VNet	12(20%)	72.38	58.26	19.35	5.89
UA-MT	12(20%)	76.10	62.62	10.84	2.43
SASSNet	12(20%)	77.66	64.08	10.93	3.05
DTC	12(20%)	78.27	64.75	8.36	2.25
BCP^{SOTA}	12(20%)	82.91	70.97	**6.43**	2.25
MCF^{SOTA}	12(20%)	75.00	61.27	11.59	3.27
PMT(Ours)	12(20%)	**83.22**	**71.52**	7.60	**1.89**

design training process, establishing temporal-level model alignment and pseudo-label filtering while leveraging a network architecture based on MT. The core idea of this model framework is to generate diverse pseudo-labels consistently during the backpropagation process by establishing representation capability differences caused by iteration gaps, thereby stabilizing and enhancing the model's representation capability. In addition to the progressive architecture, PMT employs two simple yet effective methods, Pseudo Label Filtering (PLF) and Discrepancy Driven Alignment (DDA). PLF utilizes the representation capability of the model to filter pseudo-labels, discarding low-fidelity ones detrimental to co-training, while DDA aligns the differences in model predictions, allowing lagging models to catch up with leading models rapidly. Results from ablation experiments demonstrate that each component of the PMT framework significantly enhances the model's performance. In comparative experiments with other methods, PMT achieves state-of-the-art performance in terms of accuracy , surpassing previous methods significantly, and maintains this advantage compared to other methods in situations with more limited data and more challenging tasks.

Limitation and Future Work. Despite the outstanding performance of PMT, there is still room for further exploration of semi-supervised training architectures established through temporal consistency. Investigating more advanced and stable strategies under a progressive design paradigm, and assessing whether they can further enhance performance, are topics worthy of further research in the future.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grants 62088102, 12326608 and 62106192, Natural Science Foundation of Shaanxi Province under Grant 2022JC-41, and Fundamental Research Funds for the Central Universities under Grant XTR042021005.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE transactions on pattern analysis and machine intelligence **34**(11), 2274–2282 (2012) 4
2. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11514–11524 (2023) 10, 11
3. Cai, H., Li, S., Qi, L., Yu, Q., Shi, Y., Gao, Y.: Orthogonal annotation benefits barely-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3302–3311 (2023) 4
4. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2613–2622 (2021) 5
5. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021) 2
6. Dong-DongChen, W., WeiGao, Z.: Tri-net for semi-supervised deep learning. In: Proceedings of twenty-seventh international joint conference on artificial intelligence. pp. 2014–2020 (2018) 4
7. Jiao, R., Zhang, Y., Ding, L., Cai, R., Zhang, J.: Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. arXiv preprint arXiv:2207.14191 (2022) 4
8. Ke, Z., Qiu, D., Li, K., Yan, Q., Lau, R.W.: Guided collaborative training for pixel-wise semi-supervised learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. pp. 429–445. Springer (2020) 5
9. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016) 2, 4
10. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896. Atlanta (2013) 2, 4
11. Li, S., Zhang, C., He, X.: Shape-aware semi-supervised 3d semantic segmentation for medical images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. pp. 552–561. Springer (2020) 4, 9, 10, 11
12. 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) 4, 9, 10, 11
13. 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) 9

14. Na, J., Ha, J.W., Chang, H.J., Han, D., Hwang, W.: Switching temporary teachers for semi-supervised semantic segmentation. Advances in Neural Information Processing Systems **36** (2024) 5
15. Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M.: Uncertainty-guided dual-views for semi-supervised volumetric medical image segmentation. Nature Machine Intelligence **5**(7), 724–738 (2023) 4
16. Rizve, M.N., Kardan, N., Shah, M.: Towards realistic semi-supervised learning. In: European Conference on Computer Vision. pp. 437–455. Springer (2022) 1
17. 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–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18. pp. 556–564. Springer (2015) 3, 10
18. Shen, Z., Cao, P., Yang, H., Liu, X., Yang, J., Zaiane, O.R.: Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation. arXiv preprint arXiv:2301.04465 (2023) 5
19. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017) 2, 4, 5
20. Wang, H., Li, X.: Dhc: Dual-debiased heterogeneous co-training framework for class-imbalanced semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 582–591. Springer (2023) 2
21. Wang, J., Zhou, S., Fang, C., Wang, L., Wang, J.: Meta corrupted pixels mining for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. pp. 335–345. Springer (2020) 2
22. Wang, X., Yuan, Y., Guo, D., Huang, X., Cui, Y., Xia, M., Wang, Z., Bai, C., Chen, S.: Ssa-net: Spatial self-attention network for covid-19 pneumonia infection segmentation with semi-supervised few-shot learning. Medical Image Analysis **79**, 102459 (2022) 4
23. Wang, Y., Xiao, B., Bi, X., Li, W., Gao, X.: Mcf: Mutual correction framework for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15651–15660 (2023) 2, 3, 4, 9, 10, 11
24. Wu, H., Wang, Z., Song, Y., Yang, L., Qin, J.: Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11666–11675 (2022) 2
25. Xia, K., Wang, L., Zhou, S., Hua, G., Tang, W.: Learning from noisy pseudo labels for semi-supervised temporal action localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10160–10169 (2023) 1
26. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10687–10698 (2020) 2
27. Xin, X., Wang, J., Xie, R., Zhou, S., Huang, W., Zheng, N.: Semi-supervised person re-identification using multi-view clustering. Pattern Recognition **88**, 285–297 (2019) 1
28. 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) 3, 10
29. Xu, Z., Wang, Y., Lu, D., Yu, L., Yan, J., Luo, J., Ma, K., Zheng, Y., Tong, R.K.y.: All-around real label supervision: Cyclic prototype consistency learning for semi-supervised medical image segmentation. IEEE Journal of Biomedical and Health Informatics **26**(7), 3174–3184 (2022) 4, 10
30. 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: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. pp. 605–613. Springer (2019) 4, 9, 10, 11
31. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019) 11
32. Zheng, M., You, S., Huang, L., Wang, F., Qian, C., Xu, C.: Simmatch: Semi-supervised learning with similarity matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14471–14481 (2022) 1
33. Zhou, S., Wang, J., Shu, J., Meng, D., Wang, L., Zheng, N.: Multinetwork collaborative feature learning for semisupervised person reidentification. IEEE Transactions on Neural Networks and Learning Systems **33**(9), 4826–4839 (2021) 1