Learning Dynamic Collaborative Network for Semi-supervised 3D Vessel Segmentation

Jiao Xu[1]*, Xin Chen[2]*, Lihe Zhang[1]†
[1]Dalian University of Technology [2]City University of Hong Kong
xjmmcome@mail.dlut.edu.cn, xche32@cityu.edu.hk, zhanglihe@dlut.edu.cn

Abstract

*In this paper, we present a new **d**ynam**i**c **co**llaborative network for semi-supervised 3D vessel segmentation, termed DiCo. Conventional mean teacher (MT) methods typically employ a static approach, where the roles of the teacher and student models are fixed. However, due to the complexity of 3D vessel data, the teacher model may not always outperform the student model, leading to cognitive biases that can limit performance. To address this issue, we propose a dynamic collaborative network that allows the two models to dynamically switch their teacher-student roles. Additionally, we introduce a multi-view integration module to capture various perspectives of the inputs, mirroring the way doctors conduct medical analysis. We also incorporate adversarial supervision to constrain the shape of the segmented vessels in unlabeled data. In this process, the 3D volume is projected into 2D views to mitigate the impact of label inconsistencies. Experiments demonstrate that our DiCo method sets new state-of-the-art performance on three 3D vessel segmentation benchmarks. The code repository address is https://github.com/xujiaommcome/DiCo.*

1. Introduction

In clinical applications, high-quality vascular imaging is crucial for radiologists to accurately detect and diagnose lesions, which are often subtle and challenging to identify. Consequently, improving the accuracy of 3D vessel perception is a pressing need. However, segmenting 3D vascular structures presents two challenges: i) **Scarcity of labeled data**: Labeling vessels demands extensive expertise and is labor-intensive, resulting in limited availability of high-quality annotated data. ii) **Complex appearance**: Vessels exhibit significant variability in their topological structures and diameters. Their continuous, long, and thin nature complicates the task of maintaining segmentation continuity, in contrast to other organs with more stable shapes.

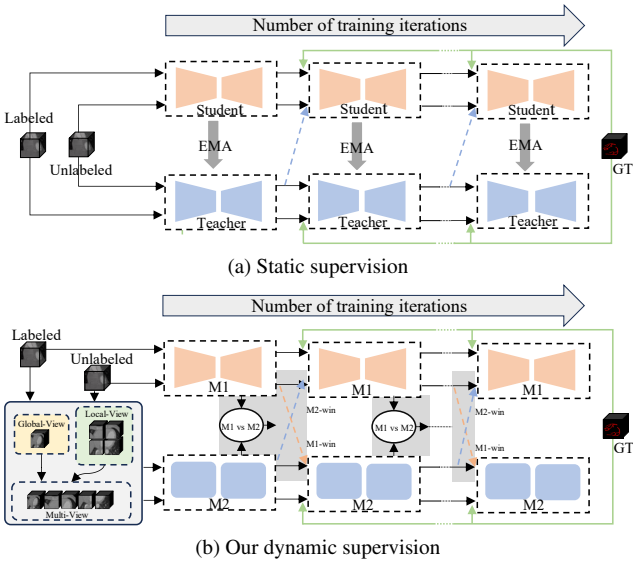

Figure 1. Semi-supervised segmentation frameworks. The dotted arrows indicate the supervision information flow of unlabeled data.

Many efforts have been made to address these challenges. To tackle the **scarcity of labeled data**, semi-supervised learning (SSL) has gained increasing attention for its ability to leverage large amounts of unlabeled data. Among SSL methods, mean teacher (MT) approaches [2, 11, 27, 40] based on consistency regularization are representative, as shown in Fig. 1(a). Although these methods have achieved significant success in segmenting common organs and tumors, they do not generalize well to vessel segmentation. This is primarily due to cognitive biases introduced by static supervision, compounded by the complexity of vessel scenes. More specifically: i) The pseudo labels provided by the static teacher model are not always superior to those of the student model, potentially misleading the student model's learning process. ii) If the student

*Equal contribution. The majority of this work was completed while Xin Chen was at Dalian University of Technology.
†Corresponding author

network learns incorrect information during early training stage, this misinformation can be propagated to the teacher model through exponential moving average (EMA). Thus, as iterations progress, errors are propagated and amplified between the MT networks, ultimately causing segmentation results to diverge from the ground truth.

To address these issues, we propose a dynamic collaborative network (DiCo) for the semi-supervised 3D vessel segmentation in this paper, as illustrated in Fig. 1(b). By dynamically switching the teacher-student roles of two models according to their current performance, DiCo alleviates the error propagation and amplification caused by static supervision. The underlying intuition is that the better-performing model at any given time should guide another model, rather than maintaining a fixed teacher-student relationship. At each training step, we compute the supervised loss based on the labeled data, designating the model with the lower loss as the teacher, while the other becomes the student. The rationale behind this approach is that, in each iteration, the labeled and unlabeled data come from the same source. Therefore, the sub-network that performs well on the labeled data generally performs better on the unlabeled data as well.

Representing the **complex appearance** of 3D vessels requires robust feature modeling capabilities. For this purpose, some efforts [4, 5, 10, 29, 37] have developed specialized layers to accommodate the tubular structures of vessels, enhancing the focus on key features. These methods compute features globally and lack attention to essential local information. In contrast, we design a straightforward multi-view integration module that captures both local vascular details and global image context. Specifically, we reorganize the original input volume to create an enhanced volume that considers both local and global views for model input. To achieve well-shaped vessel segmentation, we employ an adversarial training to align the distribution of the predicted unlabeled mask with that of the real labeled mask. Since these masks do not correspond to the same image, our goal is to ensure similar shape styles rather than pixel-level alignment. To this end, we project the 3D masks into 2D space to avoid pixel-wise correspondence supervision. Concretely, the labeled 3D volume and its ground-truth mask, as well as the unlabeled image and its predicted mask, are projected into 2D along the z axis through maximum-intensity projection (MIP). Then, the projected 2D images and masks are integrated into a discriminator for adversarial supervision, ensuring that the distribution of the unlabeled predicted masks more closely matches that of the labeled ground-truth masks, thereby preserving the continuity of the vessel structure.

Extensive experiments demonstrate the effectiveness of our DiCo method. For instance, it achieves an 86.05% DSC score on the recent vessel benchmark CAS2023 [26], surpassing the previous best method, MagicNet [2], by 2.28%, and achieves results comparable to fully supervised DSC-Net [23] using only 5% of the labeled data.On the large-scale ImageCAS [41] dataset, our method, DiCo, achieves state-of-the-art performance across three key metrics. In summary, the contributions of this work are as follows:

- We introduce a dynamic collaborative network for semi-supervised 3D vessel segmentation. This approach enables the two models in the conventional MT framework to dynamically switch the teacher-student roles, providing a new view to alleviate the cognitive bias problem.
- We propose a straightforward multi-view integration module that captures both local vascular details and global image context, leading to robust appearance modeling that supports precise vessel segmentation.
- We propose an MIP-based adversarial supervision that aligns the shape styles of predicted vessel masks with that of real mask labels, thereby enhancing the quality of predicted vessel masks.

2. Related Work

2.1. Semi-Supervised Medical Segmentation

Due to the scarcity of labeled data, semi-supervised learning has become increasingly prevalent in medical image segmentation, including the contrastive learning methods [34, 38, 43], pseudo-label methods [7, 12, 25, 32], and consistency regularization methods [13, 16, 17, 33, 36, 40]. The first two methods require intricate techniques to construct positive and negative samples and sophisticated strategies to refine pseudo-labels, respectively. The third method, consistency regularization, is gaining popularity due to its simplicity and effectiveness.

Consistency regularization methods are roughly categorized into single model paradigm and mean teacher (MT) paradigm. The former optimize models by enforcing consistency between original and perturbed unlabeled samples [17, 19, 24, 28]. Since a single model can only derive supervision from its own learning and training process, it struggles to overcome its inherent cognitive limitations, leading to significant cumulative errors. In contrast, the latter imposes consistency constraints between the student and teacher models [2, 14, 27, 31]. This paradigm introduces additional sources of supervision, alleviating the limitations of single model.

Despite their advantages, MT methods have drawbacks due to fixed teacher-student roles. Given the complexity of 3D vessel data, the teacher may not always outperform the student, leading to cognitive biases. To address this issue, this paper proposes a dynamic collaborative network that allows the two models to dynamically switch their teacher-student roles based on their timely performance.

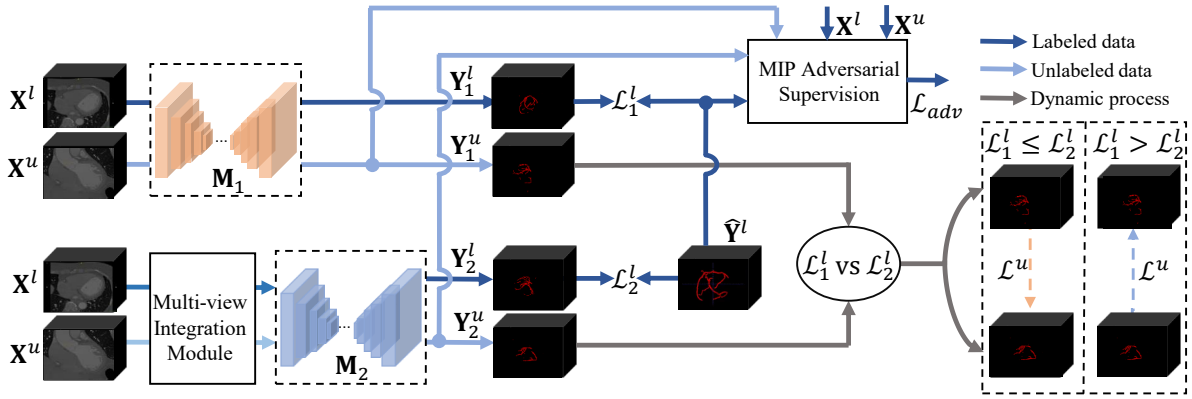

Figure 2. Architecture of the proposed DiCo method. It consists of three fundamental components: the dynamic collaborative network, the multi-view integration module, and the MIP adversarial supervision module.

2.2. Vessel Segmentation

Due to the complex nature of vessel appearance, numerous efforts have been made to achieve robust feature modeling. For example, DUNet [9] incorporates deformable convolution [3] into a U-shaped network to adaptively adjust the receptive field according to the scale and shape of blood vessels. DSCnet [22] extends deformable convolution by proposing a dynamic snake convolution method based on topological geometric constraints, which accurately captures the characteristics of tubular structures. MICNet [30] employs dense hybrid dilated convolution [39] in the connection layer, capturing richer contextual information while preserving feature resolution.

These methods typically design specialized layers tailored to the intricate tubular structure of blood vessels, resulting in enhanced adaptability to vascular images. However, such customized designs are often highly complex and lack generalizability. Additionally, these methods compute features globally and do not adequately emphasize essential local information. In contrast, this work introduces a straightforward multi-view integration module that effectively captures both the local vascular structure details and the global image context. This innovative approach requires only simple preprocessing of the input images, yet it achieves compelling vascular segmentation performance.

3. Method

In this section, we introduce the DiCo in detail. It includes the dynamic collaboration network, multi-view integration module, and MIP adversarial supervision.

3.1. Overview

The overall framework is illustrated in Fig. 2. Its core is the dynamic collaborative network, which comprises a convolutional sub-network $\mathbf{M}_1$ and a transformer sub-network $\mathbf{M}_2$. In each training iteration, we use the better-performing sub-network as the teacher and the other as the student, establishing a dynamic supervision. Building on the dynamic collaborative network, we incorporate the multi-view integration module and the MIP projection adversarial module at the input and supervision stages, respectively, to improve the model's ability to handle complex vessel appearance.

3.2. Dynamic Collaborative Network

To overcome the cognitive bias of the static MT network, the dynamic collaborative network is proposed. In this network, we utilize two sub-networks, $\mathbf{M}_1$ and $\mathbf{M}_2$, to dynamically supervise and enhance each other's performance. Specifically, in each training iteration, the labeled data $\mathbf{X}^l$ and unlabeled $\mathbf{X}^u$ are both fed into sub-networks $\mathbf{M}_1$ and $\mathbf{M}_2$ to generate the predictions, as summarized below:

$$\mathbf{Y}_1^l, \mathbf{Y}_1^u = \mathbf{M}_1\left(\mathbf{X}^l\right), \mathbf{M}_1\left(\mathbf{X}^u\right), \tag{1}$$

$$\mathbf{Y}_2^l, \mathbf{Y}_2^u = \mathbf{M}_2\left(\mathbf{X}^l\right), \mathbf{M}_2\left(\mathbf{X}^u\right). \tag{2}$$

Then, the predictions $\mathbf{Y}_1^l$ and $\mathbf{Y}_2^l$ for the labeled data are compared with the ground truth $\hat{\mathbf{Y}}^l$ to calculate the segmentation losses $\mathcal{L}_1^l$ and $\mathcal{L}_2^l$, respectively. By comparing $\mathcal{L}_1^l$ and $\mathcal{L}_2^l$, the teacher and student roles of $\mathbf{M}_1$ and $\mathbf{M}_2$ are determined. Concretely, if $\mathcal{L}_1^l \leq \mathcal{L}_2^l$, $\mathbf{M}_1$ is considered superior to $\mathbf{M}_2$ for the current input data type, so we set $\mathbf{M}_1$ as the teacher model and $\mathbf{M}_2$ as the student model. In this case, the pseudo-label $\hat{\mathbf{Y}}^u = \mathbf{Y}_1^u$ and the supervised prediction $\mathbf{Y}_o^u = \mathbf{Y}_2^u$. Otherwise, the roles are reversed. It is formulated as follows:

$$\hat{\mathbf{Y}}^u = \begin{cases} \mathbf{Y}_1^u & \text{if } \mathcal{L}_1^l \leq \mathcal{L}_2^l, \\ \mathbf{Y}_2^u & \text{otherwise.} \end{cases} \tag{3}$$

$$\mathbf{Y}_o^u = \begin{cases} \mathbf{Y}_2^u & \text{if } \mathcal{L}_1^l \leq \mathcal{L}_2^l, \\ \mathbf{Y}_1^u & \text{otherwise.} \end{cases} \tag{4}$$

The unsupervised loss $\mathcal{L}^u$ is then calculated by comparing $\mathbf{Y}_o^u$ and $\hat{\mathbf{Y}}^u$. This loss is used to update the current student

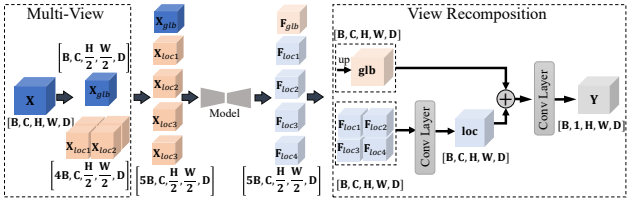

Figure 3. Multi-view integration module.

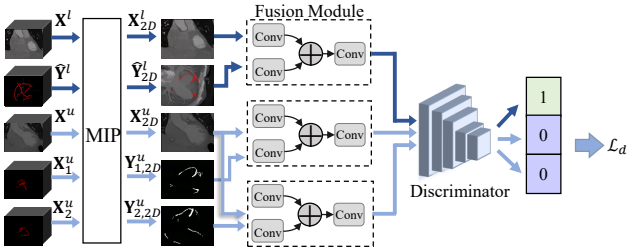

Figure 4. MIP adversarial supervision module.

model via gradient descent. In this way, our method enables dynamic teacher-student supervision between the two sub-networks, mitigating the cognitive bias problem inherent in conventional MT models.

To better leverage the advantages of this dynamic collaboration, we use different architectures for $\mathbf{M}_1$ and $\mathbf{M}_2$: $\mathbf{M}_1$ employs a convolutional architecture, while $\mathbf{M}_2$ uses a transformer architecture. This diversity allows the models to complement each other's strengths and weaknesses. $\mathbf{M}_1$'s convolutional architecture is adept at capturing fine-grained local features, while $\mathbf{M}_2$'s transformer architecture excels in understanding broader contextual relationships. By supervising each other, $\mathbf{M}_1$ and $\mathbf{M}_2$ can dynamically refine their capabilities, leading to improved overall performance and a more nuanced understanding of the complex attributes of vessel data.

3.3. Multi-view Integration Module

We introduce a multi-view integration module that captures local perspectives of the input data, as illustrated in Fig. 3. This module highlights and integrates multiple local views of the original input to improve feature representation. Due to the Transformer's superior ability to handle unstructured information, we apply this module to the input of the Transformer sub-network $\mathbf{M}_2$.

Multi-view input. The original image $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W\times D}$ is first transformed into $\mathbf{X}_{loc} \in \mathbb{R}^{n_1 n_2 n_3\times B\times C\times H/n_1\times W/n_2\times D/n_3}$, where n_1=2, n_2=2, and n_3=1 in our implementation. Subsequently, the original image is resized to $\mathbf{X}_{glb} \in \mathbb{R}^{B\times C\times H/n_1\times W/n_2\times D/n_3}$. Next, $\mathbf{X}_{glb}$ and $\mathbf{X}_{loc}$ are concatenated along the B dimension to form the input $\mathbf{X}_{input} \in \mathbb{R}^{5B\times C\times H/2\times W/2\times D}$, incorporating both global and multiple local views. This input is then fed into the $\mathbf{M}_2$ model to generate a new feature with the same dimensions. This multi-view mechanism enhances both global and local perspectives, facilitating a more comprehensive capture of the complex vessel appearance.

View recomposition. After obtaining the output from network $\mathbf{M}_2$, the features from multiple local views need to be recomposed into a cohesive feature map to support the subsequent segmentation prediction. However, due to the feature extraction process, the spatial structure of the features may not be strictly preserved, potentially causing boundary misalignment. To address this issue, we propose a view recomposition module to accurately reassemble the features. Specifically, the output features from model $\mathbf{M}_1$ are split along the batch dimension into five components: one global feature $\mathbf{F}_{\text{glb}}$ and four local features $\mathbf{F}_{\text{loc1}}, \mathbf{F}_{\text{loc2}}, \mathbf{F}_{\text{loc3}}$, and $\mathbf{F}_{\text{loc4}}$, each with dimensions $\mathbb{R}^{B\times C\times H/2\times W/2\times D}$. The global feature $\mathbf{F}_{\text{glb}}$ is upsampled to the original dimension of $\mathbb{R}^{B\times C\times H\times W\times D}$. For the local features, as illustrated in Fig. 3, we first recombine them into a feature map of size $\mathbb{R}^{B\times C\times H\times W\times D}$. Then, we apply two convolutional layers to smooth the boundaries of the recombined features. Finally, the extracted features are combined with the global features, then passed through two convolutional layers to produce a refined mixed feature map that facilitates subsequent segmentation predictions.

3.4. MIP Adversarial Supervision

3D vessels typically exhibit complex, elongated tubular structures. To enforce constraints based on vessel shape priors, we propose a maximum-intensity-projection (MIP) based adversarial supervision, as illustrated in Fig. 4. In this approach, a discriminator is trained to evaluate whether the unlabeled predicted masks capture a similar shape style to the labeled ground-truth masks. We then train networks M1 and M2 to generate masks that deceive the discriminator, thereby improving their ability to produce well-shaped vessel masks. However, using precise 3D data can pose a risk of overfitting. To mitigate this, we project the 3D data into 2D before applying the adversarial supervision.

Specifically, we use MIP to project the labeled image $\mathbf{X}^l$ with its corresponding label $\hat{\mathbf{Y}}^l$, as well as unlabeled image $\mathbf{X}^u$ with its predictions $\mathbf{Y}_1^u$ and $\mathbf{Y}_2^u$, into 2D along the depth dimension. Mathematically, for the 3D images $\mathbf{X}^l$ and $\mathbf{X}^u$, the MIP along the depth dimension z is computed as follows:

$$\mathbf{X}_{2D}^l(x,y) = \max_z \mathbf{X}^l(x,y,z), \tag{5}$$

$$\mathbf{X}_{2D}^u(x,y) = \max_z \mathbf{X}^u(x,y,z), \tag{6}$$

where $\mathbf{X}_{2D}^l$ and $\mathbf{X}_{2D}^u$ denote the projected labeled and unlabeled images, respectively. Here, (x, y) represent the spatial coordinates in the 2D projection, and z denotes the depth dimension. Similarly, for the label $\hat{Y}$ and the predictions $\mathbf{Y}_1^u$

Table 1. Comparison on ImageCAS dataset segmentation.

Method		Source	Scans Used		Metrics		
			L Volumes	U Volumes	DSC↑ (%)	NSD ↑(%)	ASD ↓(voxel)
Semi-supervised	MT [27]	ICLR'17	45	855	71.05	56.32	24.06
	UA-MT [40]	MICCAI'19	45	855	70.11	55.83	23.11
	SASSNet [13]	MICCAI'20	45	855	72.73	58.54	21.96
	SLCNet [15]	MICCAI'22	45	855	72.78	57.36	20.44
	MagicNet [2]	CVPR'23	45	855	71.88	57.50	22.10
	CAML [6]	MICCAI'23	45	855	71.66	58.03	23.73
	CauSSL [18]	ICCV'23	45	855	69.14	53.25	24.79
	GuidedNet [42]	ACMMM'24	45	855	65.24	50.63	30.21
	DiCo	**Ours**	45	855	**73.79**	**58.59**	**20.00**
Full-supervised	VNet [1]	ICCV'16	900	0	71.19	55.36	23.88
	CTNet [21]	BIBM'22	900	0	79.71	69.31	11.97
	ERNet [35]	Med Image Anal '22	900	0	76.25	64.56	18.92
	DSCNet [23]	ICCV'23	900	0	73.49	58.06	22.78

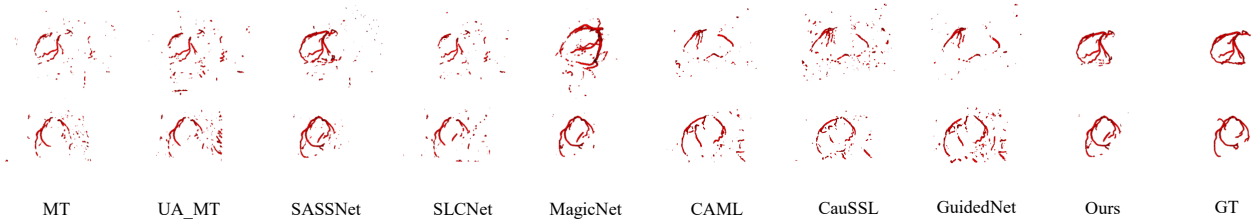

Figure 5. Visual segmentation examples from ImageCAS dataset.

and $\mathbf{Y}_2^u$, the 2D projection using MIP is given by:

$$\hat{\mathbf{Y}}_{2D}^l(x,y) = \max_z \hat{\mathbf{Y}}^l(x,y,z), \tag{7}$$

$$\mathbf{Y}_{1,2D}^u(x,y) = \max_z \mathbf{Y}_1^u(x,y,z), \tag{8}$$

$$\mathbf{Y}_{2,2D}^u(x,y) = \max_z \mathbf{Y}_2^u(x,y,z). \tag{9}$$

Here, $\mathbf{Y}_{1,2D}^u$ and $\mathbf{Y}_{2,2D}^u$ denotes the projected predictions for the unlabeled image, while $\hat{\mathbf{Y}}_{2D}^l$ represents the projected label of the labeled image. We then fuse these labels or predictions with their corresponding images in a structured manner, as summarized:

$$\hat{\mathbf{O}}_{2D}^l = \mathbf{F}(\hat{\mathbf{Y}}_{2D}^l, \mathbf{X}_{2D}^l), \tag{10}$$

$$\mathbf{O}_{1,2D}^u = \mathbf{F}(\mathbf{Y}_{1,2D}^u, \mathbf{X}_{2D}^u), \tag{11}$$

$$\mathbf{O}_{2,2D}^u = \mathbf{F}(\mathbf{Y}_{2,2D}^u, \mathbf{X}_{2D}^u). \tag{12}$$

Here, $\mathbf{F}(\cdot)$ denotes a fusion module, as illustrated by the dotted box in Fig. 4. The fused outputs are represented as $\hat{\mathbf{O}}_{2D}^l$, $\mathbf{O}_{1,2D}^u$, and $\mathbf{O}_{2,2D}^u$. We then feed the three fused images into a discriminator $\mathbf{D}(\cdot)$, training it to distinguish between images generated using the ground-truth mask and those generated using predicted masks. The loss function is the binary cross-entropy (BCE) loss. In this setup, the labels for $\mathbf{O}_{1,2D}^u$ and $\mathbf{O}_{2,2D}^u$ are set to 0, while the label for $\hat{\mathbf{O}}_{2D}^l$ is set to 1. This is summarized as:

$$\begin{aligned}\mathcal{L}_d =& \mathcal{L}_{bce}(\mathbf{D}(\hat{O}_{2D}^l), 1) + \mathcal{L}_{bce}(\mathbf{D}(\mathbf{O}_{1,2D}^u), 0) \\ &+ \mathcal{L}_{bce}(\mathbf{D}(\mathbf{O}_{2,2D}^u), 0).\end{aligned} \tag{13}$$

Then, we train $\mathbf{M}_1$ and $\mathbf{M}_2$ to deceive the discriminator $\mathbf{D}$ by minimizing an adversarial loss $\mathcal{L}_{adv}$:

$$\mathcal{L}_{adv} = \mathcal{L}_{bce}(\mathbf{D}(\mathbf{O}_{1,2D}^u), 1) + \mathcal{L}_{bce}(\mathbf{D}(\mathbf{O}_{2,2D}^u), 1). \tag{14}$$

3.5. Loss Functions

To supervise the segmentation of models M1 and M2, we employ a combination of Dice loss and cross-entropy loss:

$$\mathcal{L}_{seg} = \alpha\mathcal{L}_{dice} + \beta\mathcal{L}_{ce}, \tag{15}$$

$$\mathcal{L}_1^l = \mathcal{L}_{seg}(\mathbf{Y}_1^l, \hat{\mathbf{Y}}^l), \tag{16}$$

$$\mathcal{L}_2^l = \mathcal{L}_{seg}(\mathbf{Y}_2^l, \hat{\mathbf{Y}}^l), \tag{17}$$

where α and β are regularization parameters. For the dynamic collaborative network, the unsupervised loss function is as follows:

$$\mathcal{L}^u = \mathcal{L}_{seg}(\mathbf{Y}_o^u, \hat{\mathbf{Y}}^u). \tag{18}$$

The total loss function is given by:

$$\mathcal{L} = \mathcal{L}_1^l + \mathcal{L}_2^l + \mathcal{L}^u + \mathcal{L}_{adv}. \tag{19}$$

4. Experiments

4.1. Datasets and Metrics

We use three vessel datasets for training and evaluation, 5% of the training data are selected as labeled volumes, while the remaining data are treated as unlabeled volumes during the training process:

Table 2. Comparison on CAS2023 dataset segmentation.

Method		Source	Scans Used		Metrics		
			L Volumes	U Volumes	DSC↑ (%)	NSD ↑(%)	ASD ↓(voxel)
Semi-supervised	MT [27]	ICLR'17	5	85	73.94	58.33	15.21
	UA-MT [40]	MICCAI'19	5	85	77.31	59.22	5.89
	SASSNet [13]	MICCAI'20	5	85	75.43	58.82	7.67
	SLCNet [15]	MICCAI'22	5	85	77.21	61.65	3.65
	MagicNet [2]	CVPR'23	5	85	84.13	72.73	1.92
	CAML [6]	MICCAI'23	5	85	79.64	67.43	2.19
	CauSSL [18]	ICCV'23	5	85	76.72	62.48	2.22
	GuidedNet [42]	ACMMM'24	5	85	81.60	69.00	2.01
	DiCo	**Ours**	5	85	**86.05**	**74.35**	**1.49**
Full-supervised	VNet [1]	ICCV'16	90	0	67.25	43.38	21.36
	CTNet [21]	BIBM'22	90	0	77.51	63.24	8.76
	ERNet [35]	Med Image Anal '22	90	0	74.62	59.78	12.43
	DSCNet [23]	ICCV'23	90	0	83.14	70.09	2.69

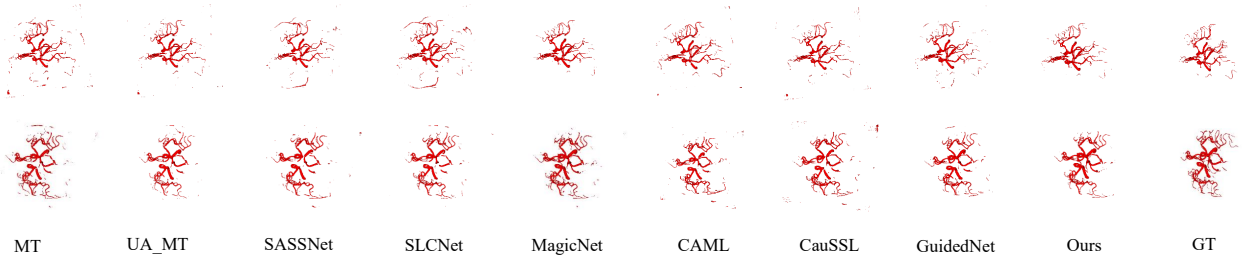

Figure 6. Visual segmentation examples from CAS2023 dataset.

- **ImageCAS** [41] comprises 1,000 3D coronary vessel computed tomography angiography (CTA) images, with voxel dimensions ranging from $512 \times 512 \times (206-275)$. Of these, 900 images are allocated for training, and 100 for evaluation.
- **CAS2023** [26] is a cerebral artery vessel segmentation dataset from the MICCAI 2023 challenge, containing 100 public magnetic resonance angiography (MRA) samples. This dataset includes 90 images for training and 10 for evaluation.
- **Parse2022** [20] provides computed tomography (CT) images for pulmonary artery vessel segmentation from the MICCAI 2022 challenge.This dataset includes 90 images for training and 10 for evaluation.

The performance is assessed by using three metrics: the Dice similarity coefficient (DSC) for region sensitivity, the normalized surface Dice coefficient (NSD) for evaluating surface overlap precision, and the average surface distance (ASD) for edge sensitivity. The DSC is generally regarded as a primary metric for medical image segmentation.

4.2. Implementation Details

All experiments in this study are conducted using Python 3.9 and PyTorch 2.2, with training and evaluation carried out on an NVIDIA 3090 GPU.

For the model $\mathbf{M1}$, we employ VNet [1], a widely recognized convolutional network for medical image segmentation. For $\mathbf{M2}$, we select UNETR [8], a well-established vision transformer network. Details are available in *supplementary materials*. During inference, the final prediction is obtained from the VNet output, consistent with the approach used in other methods that utilize VNet.

Our model is trained using the AdamW optimizer for 40,000 iterations, with an initial learning rate lr_{base} of $1e{-}2$. A learning rate decay strategy is employed, defined by:

$$\text{lr} = \text{lr}_{\text{base}} \times \left(1 - \frac{t}{T}\right)^{\gamma}. \tag{20}$$

Here, t represents the current training iteration, T denotes the total number of iterations, γ is the exponent used to adjust the rate of decay. The batch size is set to 2. We employ a center-crop with size $96{\times}96{\times}96$ for the input volumes. During the inference phase, a sliding window approach is employed to generate the final result for the entire volume.

4.3. Comparisons with State-of-the-arts

We compare our DiCo method with state-of-the-art (SOTA) self-supervised medical image segmentation methods, including MT [27], UA-MT [40], SASSNet [13], SLCNet [15], MagicNet [2], CAML [6], CauSSL [18], and GuidedNet [42], as well as fully-supervised methods designed for vessel data, such as CTNet [21], DSCNet [23], and ERNet [35]. We also incorporate the baseline method VNet [1] in the comparison, which is a fully-supervised medical image segmentation method.

Table 3. Comparison on Parse2022 dataset segmentation.

Method		Source	Scans Used		Metrics		
			L Volumes	U Volumes	DSC↑ (%)	NSD ↑(%)	ASD ↓(voxel)
Semi-supervised	MT [27]	ICLR'17	5	85	58.36	42.58	11.37
	UA-MT [40]	MICCAI'19	5	85	62.70	45.72	10.51
	SASSNet [13]	MICCAI'20	5	85	68.33	29.19	7.57
	SLCNet [15]	MICCAI'22	5	85	66.13	48.87	8.18
	MagicNet [2]	CVPR'23	5	85	69.19	53.25	**5.53**
	CAML [6]	MICCAI'23	5	85	66.75	50.22	7.27
	CauSSL [18]	ICCV'23	5	85	66.45	48.88	7.94
	GuidedNet [42]	ACMMM'24	5	85	68.80	51.80	6.74
	DiCo	**Ours**	5	85	**70.93**	**55.26**	5.74
Full-supervised	VNet [1]	ICCV'16	90	0	65.53	48.87	8.08
	CTNet [21]	BIBM'22	90	0	73.12	58.92	5.88
	ERNet [35]	Med Image Anal '22	90	0	76.39	64.72	3.81
	DSCNet [23]	ICCV'23	90	0	75.04	59.66	4.49

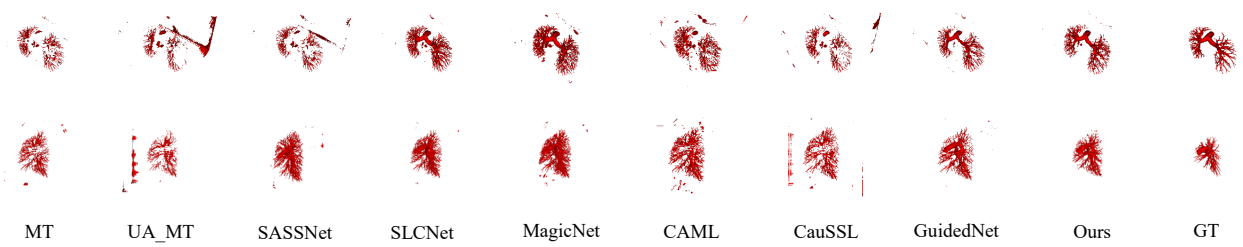

Figure 7. Visual segmentation examples from Parse2022 dataset.

ImageCAS. As shown in Tab 1, on the ImageCAS dataset, our DiCo method achieves a DSC of 73.79%, an NSD of 58.59%, and an ASD of 20.00 voxels. These results surpass other semi-supervised methods, demonstrating the superiority of the DiCo method. Moreover, DiCo achieves comparable performance to fully supervised methods while using only 5% of the labeled data, underscoring its efficiency in leveraging limited annotations. Visualization results are provided in Fig. 5.

CAS2023. As shown in Tab 2, on the CAS2023 dataset, our DiCo method achieves the top performance with 86.05% in DSC, 74.35% in NSD, and 1.49 voxels in ASD, outperforming all compared semi-supervised methods across each metric. In particular, DiCo surpasses the second-best MagicNet by 2.28% in the primary metric DSC. Furthermore, DiCo demonstrates comparable performance to fully supervised methods that rely on significantly more labeled data, underscoring DiCo's efficiency in semi-supervised learning. Examples of segmentation results are shown in Fig. 6.

Parse2022. As shown in Tab 3, on the Parse2022 dataset, our DiCo method achieves a DSC of 70.93%, an NSD of 55.26%, and an ASD of 5.74 voxels. Both the DSC and NSD are superior to other semi-supervised methods. Specifically, DiCo surpasses the recently proposed GuidedNet by 3.10% in DSC and by 6.68% in NSD. In terms of the ASD metric, DiCo also delivers a competitive result, trailing MagicNet by only 0.21 voxels while surpassing all other semi-supervised methods. Representative segmentation results are shown in Fig. 7.

Dataset	Metric	C+T	C+C	T+T	MT
ImageCAS	*DSC*↑	70.37	69.45	62.67	**71.05**
	NSD↑	**56.36**	53.42	47.54	56.32
	ASD↓	**23.87**	25.89	32.47	24.06
CAS2023	*DSC*↑	**83.59**	82.74	78.56	73.94
	NSD↑	**73.44**	70.89	67.28	58.33
	ASD↓	**1.79**	1.97	2.17	15.21
Parse2022	*DSC*↑	**63.85**	62.77	62.89	58.36
	NSD↑	**43.39**	43.21	42.45	42.58
	ASD↓	**9.02**	10.84	9.65	11.37

Table 4. Ablation study results of the DiCo architecture. MT stands for the basic mean teacher architecture. C represents the CNN network V-Net, and T denotes the ViT network UNETR. C+C, C+T, and T+T represent DiCo with different combinations of sub-networks.

4.4. Ablation Study

DiCo variants. We conducted experiments to explore various combinations of sub-networks, including CNN+ViT (our default configuration, denoted as C+T), CNN+CNN (denoted as C+C), and ViT+ViT (denoted as T+T). The results are reported in Tab. 4. Among these configurations, the C+T combination achieves the best performance. This superior performance can be attributed to the complementary strengths of the CNN and ViT: the CNN effectively cap-

Dataset	Metric	MT	Base	+MIP	+MV	All
ImageCAS	*DSC*↑	71.05	70.37	71.15	72.37	**73.79**
	NSD↑	56.32	56.36	53.64	57.36	**58.59**
	ASD↓	24.06	23.87	21.97	**19.44**	20.00
CAS2023	*DSC*↑	73.94	83.59	84.42	85.63	**86.05**
	NSD↑	58.33	73.44	73.24	73.40	**74.35**
	ASD↓	15.21	1.79	2.31	1.63	**1.49**
Parse2022	*DSC*↑	58.36	63.85	66.29	68.46	**70.93**
	NSD↑	42.58	47.39	50.15	52.14	**55.26**
	ASD↓	11.37	9.02	7.24	5.92	**5.74**

Table 5. Component-wise study. MT represents the basic mean teacher architecture, MIP denotes our MIP adversarial supervision, MV indicates our multi-view integration model, Base denotes our DiCo without MIP and MV, and All is our default DiCo model.

Datasets	Method	Metric		
		DSC↑	*NSD*↑	*ASD*↓
ImageCAS	3D	60.79	43.73	46.31
	2D	**71.15**	**53.64**	**21.97**
CAS2023	3D	83.25	71.39	2.49
	2D	**84.42**	**73.24**	**2.31**
Pares2022	3D	59.33	46.92	12.68
	2D	**66.29**	**50.15**	**7.24**

Table 6. Ablation study of the 2D projection in MIP adversarial supervision.

tures fine-grained local features, while the transformer excels at understanding broader contextual relationships. By supervising each other, the two sub-networks dynamically enhance their capabilities, leading to a more nuanced understanding of the complex characteristics of vessel data.

DiCo *v.s.* MT. Tab. 4 also presents the results of the mean teacher (MT) method, which employs static supervision. Our DiCo approach demonstrates strong overall effectiveness, highlighting the power of dynamic collaboration. Notably, in the DiCo approach, our default C+T configuration surpasses the aligned MT method by an average of 7.16% in DSC across three vessel datasets.

Component-wise Study. We conduct experiments to explore the impact of each component in DiCo, with the results presented in Tab. 5. In this table, MT represents the basic mean teacher architecture, MIP denotes our MIP adversarial supervision, MV indicates our multi-view integration model, Base denotes our DiCo without MIP and MV, and All is our default DiCo model.

Compared to the MT method, even our base DiCo model demonstrates a significant improvement, with a 13.05% increase on the CAS2023 dataset. Building on this, the MIP adversarial supervision approach further enhances performance across the three datasets by 1.11%, 1.00%, and 3.82% in DSC score, respectively, showcasing its effectiveness. The MV model improves performance by 2.84%, 2.44%, and 7.22% on the three datasets, underscoring the benefits of integrating multiple data views. Ultimately, the combination of all three designs delivers the best performance across all datasets.

#	Loss	DSC↑	NSD↑	ASD↓
1	$\mathcal{L}_{ce}$	74.23	59.56	11.47
2	$\mathcal{L}_{dice}$	76.00	62.25	8.90
3	$\mathcal{L}_{seg}$	78.41	63.69	5.72
4	$\mathcal{L}_{seg} + \mathcal{L}_{u}$	85.63	73.40	1.63
5	$\mathcal{L}_{seg} + \mathcal{L}_{u} + \mathcal{L}_{adv}$	**86.05**	**74.35**	**1.49**

Table 7. The impact of the loss function is verified on the CAS2023 dataset.

3D v.s. 2D adversarial supervision. We conduct experiments to assess the impact of using MIP to project 3D data into 2D for adversarial supervision. The results presented in Tab. 6 indicate that projecting the 3D data to 2D improves the DSC scores by 17.04%, 1.41%, and 11.73% points across the three datasets, respectively. These findings demonstrate the effectiveness of our 2D projection approach for the adversarial supervision.

Impact of loss function. We investigate the impact of the loss functions used, with the results detailed in Table 7. The evaluated dataset is CAS2023. As shown in Rows #1 and #2, relying solely on cross-entropy loss $\mathcal{L}ce$ or Dice loss $\mathcal{L}dice$ can result in performance degradation. Combining these two loss functions, as demonstrated in Row #3, leads to improved performance. Row #4 demonstrates that incorporating our unsupervised loss $\mathcal{L}_u$ significantly enhances performance. Row#5 shows that our MIP adversarial supervision further improves the performance.

5. Conclusion

This work introduces DiCo, a new semi-supervised 3D vessel segmentation framework that enables dynamic collaboration between two models. By dynamically alternating teacher and student roles based on model performance, DiCo mitigates the cognitive biases inherent in conventional static semi-supervision approaches. Additionally, the integration of a multi-view module and MIP-based adversarial supervision further enhances segmentation quality. Experimental results demonstrate that DiCo is effective, achieving competitive performance compared to state-of-the-art medical segmentation methods. We hope this work provides a new perspective on learning for 3D vessel segmentation.

Acknowledgements. This work was supported by the National Natural Science Foundation of China under Grant 62431004 and 62276046, and by Dalian Science and Technology Innovation Foundation under Grant 2023JJ12GX015.

References

[1] Abolfazl Abdollahi, Biswajeet Pradhan, and Abdullah Alamri. Vnet: An end-to-end fully convolutional neural network for road extraction from high-resolution remote sensing data. *IEEE Access*, page 179424–179436, 2020. 5, 6, 7

[2] Duowen Chen, Yunhao Bai, Wei Shen, Qingli Li, Lequan Yu, and Yan Wang. Magicnet: Semi-supervised multi-organ segmentation via magic-cube partition and recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23869–23878, 2023. 1, 2, 5, 6, 7

[3] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 764–773, 2017. 3

[4] Shunjie Dong, Jinlong Zhao, Maojun Zhang, Zhengxue Shi, Jianing Deng, Yiyu Shi, Mei Tian, and Cheng Zhuo. Deunet: Deformable u-net for 3d cardiac mri video segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, pages 98–107. Springer, 2020. 2

[5] Shunjie Dong, Zixuan Pan, Yu Fu, Qianqian Yang, Yuanxue Gao, Tianbai Yu, Yiyu Shi, and Cheng Zhuo. Deu-net 2.0: Enhanced deformable u-net for 3d cardiac cine mri segmentation. *Medical Image Analysis*, page 102389, 2022. 2

[6] Shengbo Gao, Ziji Zhang, Jiechao Ma, Zihao Li, and Shu Zhang. Correlation-aware mutual learning for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 98–108. Springer, 2023. 5, 6, 7

[7] Kai Han, Lu Liu, Yuqing Song, Yi Liu, Chengjian Qiu, Yangyang Tang, Qiaoying Teng, and Zhe Liu. An effective semi-supervised approach for liver ct image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(8): 3999–4007, 2022. 2

[8] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 574–584, 2022. 6

[9] Qiangguo Jin, Zhaopeng Meng, Tuan D. Pham, Qi Chen, Leyi Wei, and Ran Su. Dunet: A deformable network for retinal vessel segmentation. *Knowledge-Based Systems*, page 149–162, 2019. 3

[10] Bin Kong, Xin Wang, Junjie Bai, Yi Lu, Feng Gao, Kunlin Cao, Jun Xia, Qi Song, and Youbing Yin. Learning tree-structured representation for 3d coronary artery segmentation. *Computerized Medical Imaging and Graphics*, page 101688, 2020. 2

[11] Tao Lei, Dong Zhang, Xiaogang Du, Xuan Wang, Yong Wan, AsokeK Nandi, and Lei Lei. Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network. *IEEE transactions on medical imaging*, 42(5):1265–1277, 2022. 1

[12] Caizi Li, Li Dong, Qi Dou, Fan Lin, Kebao Zhang, Zuxin Feng, Weixin Si, Xuesong Deng, Zhe Deng, and Pheng-Ann Heng. Self-ensembling co-training framework for semi-supervised covid-19 ct segmentation. *IEEE Journal of Biomedical and Health Informatics*, page 4140–4151, 2021. 2

[13] Shuailin Li, Chuyu Zhang, and Xuming He. 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*, pages 552–561. Springer, 2020. 2, 5, 6, 7

[14] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation consistent self-ensembling model for semi-supervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, page 523–534, 2021. 2

[15] Jinhua Liu, Christian Desrosiers, and Yuanfeng Zhou. Semi-supervised medical image segmentation using cross-model pseudo-supervision with shape awareness and local context constraints. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 140–150. Springer, 2022. 5, 6, 7

[16] Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianyong Chen, Guotai Wang, and Shaoting Zhang. *Efficient Semi-supervised Gross Target Volume of Nasopharyngeal Carcinoma Segmentation via Uncertainty Rectified Pyramid Consistency*, page 318–329. Springer-Verlag, 2021. 2

[17] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. *Proceedings of the AAAI Conference on Artificial Intelligence*, page 8801–8809, 2022. 2

[18] Juzheng Miao, Cheng Chen, Furui Liu, Hao Wei, and Pheng-Ann Heng. Caussl: Causality-inspired semi-supervised learning for medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21426–21437, 2023. 5, 6, 7

[19] Yassine Ouali, Celine Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12674–12684, 2020. 2

[20] Rohan Padhy, Akansh Maurya, Kunal Dasharath Patil, Kalluri Ramakrishna, and Ganapathy Krishnamurthi. Parse challenge 2022: Pulmonary arteries segmentation using swin u-net transformer(swin unetr) and u-net. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4, 2023. 6

[21] Chengwei Pan, Baolian Qi, Gangming Zhao, Jiaheng Liu, Chaowei Fang, Dingwen Zhang, and Jinpeng Li. Deep 3d vessel segmentation based on cross transformer network. In *2022 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 1115–1120. IEEE, 2022. 5, 6, 7

[22] Yaolei Qi, Yuting He, Xiaoming Qi, Yuan Zhang, and Guanyu Yang. Dynamic snake convolution based on topo-

logical geometric constraints for tubular structure segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6070–6079, 2023. 3

[23] Yaolei Qi, Yuting He, Xiaoming Qi, Yuan Zhang, and Guanyu Yang. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6070–6079, 2023. 2, 5, 6, 7

[24] Margherita Rosnati, Melanie Roschewitz, and Ben Glocker. Robust semi-supervised segmentation with timestep ensembling diffusion models, 2023. 2

[25] Yinghuan Shi, Jian Zhang, Tong Ling, Jiwen Lu, Yefeng Zheng, Qian Yu, Lei Qi, and Yang Gao. Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. *IEEE transactions on medical imaging*, 41 (3):608–620, 2021. 2

[26] Sunhaozhong. Cerebral artery segmentation challenge (cas) 2023. CodaLab Competitions, 2023. https://codalab.lisn.upsaclay.fr/competitions/9804. 2, 6

[27] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *International Conference on Learning Representations*, 30, 2017. 1, 2, 5, 6, 7

[28] Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Arno Solin, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *Neural Networks*, page 90–106, 2022. 2

[29] Dong Wang, Zhao Zhang, Ziwei Zhao, Yuhang Liu, Yihong Chen, and Liwei Wang. Pointscatter: Point set representation for tubular structure extraction. In *European Conference on Computer Vision*, pages 366–383. Springer, 2022. 2

[30] Jinke Wang, Lubiao Zhou, Zhongzheng Yuan, Haiying Wang, and Changfa Shi. Mic-net: multi-scale integrated context network for automatic retinal vessel segmentation in fundus image. *Mathematical Biosciences and Engineering*, page 6912–6931, 2023. 3

[31] Kaiping Wang, Bo Zhan, Chen Zu, Xi Wu, Jiliu Zhou, Luping Zhou, and Yan Wang. Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. *Medical Image Analysis*, 79:102447, 2022. 2

[32] Xiaoyan Wang, Yiwen Yuan, Dongyan Guo, Xiaojie Huang, Ying Cui, Ming Xia, Zhenhua Wang, Cong Bai, and Shengyong Chen. Ssa-net: Spatial self-attention network for covid-19 pneumonia infection segmentation with semi-supervised few-shot learning. *Medical image analysis*, 79:102459, 2022. 2

[33] Yixin Wang, Yao Zhang, Jiang Tian, Cheng Zhong, Zhongchao Shi, Yang Zhang, and Zhiqiang He. *Double-Uncertainty Weighted Method for Semi-supervised Learning*, page 542–551. Springer-Verlag, 2020. 2

[34] Yicheng Wu, Zhonghua Wu, Qianyi Wu, Zongyuan Ge, and Jianfei Cai. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 34–43. Springer, 2022. 2

[35] Likun Xia, Hao Zhang, Yufei Wu, Ran Song, Yuhui Ma, Lei Mou, Jiang Liu, Yixuan Xie, Ming Ma, and Yitian Zhao. 3d vessel-like structure segmentation in medical images by an edge-reinforced network. *Medical Image Analysis*, 82: 102581, 2022. 5, 6, 7

[36] Zhe Xu, Yixin Wang, Donghuan Lu, Lequan Yu, Jiangpeng Yan, Jie Luo, Kai Ma, Yefeng Zheng, and Raymond Kai-yu Tong. All-around real label supervision: Cyclic prototype consistency learning for semi-supervised medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, page 3174–3184, 2022. 2

[37] Xin Yang, Zhiqiang Li, Yingqing Guo, and Dake Zhou. Dcu-net: a deformable convolutional neural network based on cascade u-net for retinal vessel segmentation. *Multimedia Tools and Applications*, page 15593–15607, 2022. 2

[38] Chenyu You, Yuan Zhou, Ruihan Zhao, Lawrence Staib, and James S. Duncan. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, page 2228–2237, 2022. 2

[39] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 472–480, 2017. 3

[40] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. 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*, pages 605–613. Springer, 2019. 1, 2, 5, 6, 7

[41] An Zeng, Chunbiao Wu, Meiping Huang, Jian Zhuang, Shanshan Bi, Dan Pan, Najeeb Ullah, KaleemNawaz Khan, Tianchen Wang, Yiyu Shi, Xiaomeng Li, Guisen Lin, and Xiaowei Xu. Imagecas: A large-scale dataset and benchmark for coronary artery segmentation based on computed tomography angiography images. *Computerized Medical Imaging and Graphics*, 109:102287, 2023. 2, 6

[42] Haochen Zhao, Hui Meng, Deqian Yang, Xiaozheng Xie, Xiaoze Wu, Qingfeng Li, and Jianwei Niu. Guidednet: Semi-supervised multi-organ segmentation via labeled data guide unlabeled data. *arXiv preprint arXiv:2408.04914*, 2024. 5, 6, 7

[43] Xiangyu Zhao, Zengxin Qi, Sheng Wang, Qian Wang, Xuehai Wu, Ying Mao, and Lichi Zhang. Rcps: Rectified contrastive pseudo supervision for semi-supervised medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2023. 2