SGTC: Semantic-Guided Triplet Co-training for Sparsely Annotated Semi-Supervised Medical Image Segmentation

Ke Yan[1†], Qing Cai[1†], Fan Zhang[2*], Ziyan Cao[1], Zhi Liu[3*]

[1]Faculty of Computer Science and Technology, Ocean University of China
[2]School of Automation, Northwestern Polytechnical University
[3]School of Information Science and Engineering, Shandong University
{yk6923, caoziyan}@stu.ouc.edu.cn, cq@ouc.edu.cn, zfnlxx@mail.nwpu.edu.cn, liuzhi@sdu.edu.cn

Abstract

Although semi-supervised learning has made significant advances in the field of medical image segmentation, fully annotating a volumetric sample slice by slice remains a costly and time-consuming task. Even worse, most of the existing approaches pay much attention to image-level information and ignore semantic features, resulting in the inability to perceive weak boundaries. To address these issues, we propose a novel Semantic-Guided Triplet Co-training (SGTC) framework, which achieves high-end medical image segmentation by only annotating three orthogonal slices of a few volumetric samples, significantly alleviating the burden of radiologists. Our method consist of two main components. Specifically, to enable semantic-aware, fine-granular segmentation and enhance the quality of pseudo-labels, a novel semantic-guided auxiliary learning mechanism is proposed based on the pretrained CLIP. In addition, focusing on a more challenging but clinically realistic scenario, a new triple-view disparity training strategy is proposed, which uses sparse annotations (i.e., only three labeled slices of a few volumes) to perform co-training between three sub-networks, significantly improving the robustness. Extensive experiments on three public medical datasets demonstrate that our method outperforms most state-of-the-art semi-supervised counterparts under sparse annotation settings. The source code is available at https://github.com/xmeimeimei/SGTC.

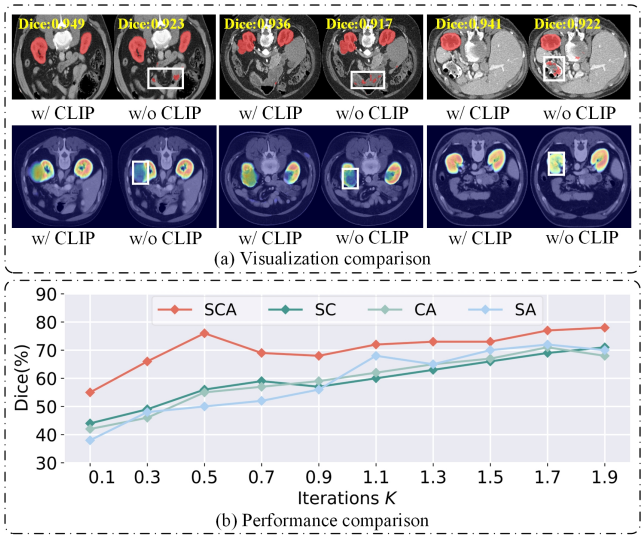

Figure 1: (a) The semantic guidance in visualization comparison. (b) The performance under different sparse annotation strategy. The S, C, and A indicate sagittal, coronal, and axial annotated slices.

Introduction

Segmentation of anatomical structures and pathology within medical images holds paramount importance for clinical diagnosis (Zhou et al. 2019), treatment planning (Li et al. 2023b; Zhang et al. 2024a), and disease research (Zhang et al. 2022b). While significant progress has been achieved through deep learning-based segmentation techniques, many approaches encounter substantial bottlenecks when lacking sufficient well-annotated datasets (Zhang et al. 2024c, 2022a). Consequently, there is a critical need to develop more effective yet precise segmentation methods to decrease the dependence on large-scale pixel-wise annotated data.

Considering that unlabeled data are easy to obtain, semi-supervised medical image segmentation has emerged as the predominant strategy in medical domain, utilizing a limited number of labeled data and lots of unlabeled data (Wang et al. 2024; Wang and Li 2024; Zhang et al. 2022c; Cai et al. 2018; Zhou et al. 2023). Current semi-supervised medical segmentation methods can be classified into two categories. The first one is pseudo-label-based methods (Bai et al. 2017; Yu et al. 2019; Bai et al. 2023), which estimate the pseudo labels based on a few labeled samples and then attach them to unlabeled ones followed by a fine-tuning stage on the newly enlisted training set. The other mainstream approaches are consistency regularization (Chen et al. 2023; Shen et al. 2023; Li et al. 2020), which enforce the output consistency for the inputs under different image or feature perturbations.

Despite the progress, these methods need dense annotations, requiring to fully annotate the entire volume of limited labeled data slice by slice, it remains a costly and time-consuming task due to the considerable number of slices per volume. Generally, radiologists only annotate a few slices from the volumetric medical data and leave others unla-

[†]These authors contributed equally.
[*]Corresponding authors.

beled (Zhang et al. 2024b; Cai et al. 2019, 2021a; Yun et al. 2023; Wu et al. 2023). Under this condition, the semi-supervised methods based on sparse annotations have been proposed, for example, PLN (Li et al. 2022) annotates only one slice per volume and propagates the pseudo labels for other slices using the parasitic-like network. Similarly, Desco (Cai et al. 2023) annotates two orthogonal slices per volume and utilizes the SyNRA method from the Ants (Avants et al. 2011) library for registration, constructing consistency between pseudo labels generated from two perspectives. Nonetheless, these methods have limitations in semantic understanding, which hinders the network's ability to accurately recognize anatomical structures and lesions. As shown in Figure 1(a), without semantic information providing contextual clues, resulting in imprecise segmentation results. Furthermore, slice-by-slice propagation or registration is a time-consuming process, especially when the volume comprises a substantial number of slices. Worse still, annotating only one or two slices misses complementary information from three different views, resulting in an incomplete representation of the volume's data distribution, consequently leading to a suboptimal segmentation (See Figure 1(b)).

To overcome these challenges, this paper proposes a novel semantic-guided triplet co-training framework, dubbed SGTC, which achieves semantic-aware and fine-granular semi-supervised medical image segmentation by merely annotating three orthogonal slices of a few volumetric samples. Specifically, to enable the use of text representations to connect semantic-aware features, a novel semantic-guided auxiliary learning mechanism is proposed. It enhances pseudo-label quality for abundant unlabeled medical data by refining the intricate structures and weak boundaries. Besides, to better align with the spatial information distribution of volumetric data using sparse annotations, a novel triple-view disparity training strategy is proposed, which better maintains the disparity of sub-networks during training, allowing the sub-networks to learn complementary knowledge from each other. More importantly, it focuses on a more challenging but clinically realistic scenario, where radiologists just need to annotate three orthogonal slices of a few volumetric samples. Extensive experiments on LA2018, KiTS19, and LiTS datasets under sparse annotation settings show that our SGTC achieves superior performance against most state-of-the-art semi-supervised learning methods.

The primary contributions of this paper include:

- A novel semantic-guided auxiliary learning mechanism is proposed, which not only enables semantic-aware and fine-granular semi-supervised medical image segmentation, but enhances the quality of pseudo labels.
- A novel triple-view disparity training strategy is proposed, which uses only three labeled slices of a few volumes to encourage the disparity of sub-networks, significantly improving the robustness.
- Extensive experimental results on three challenging semi-supervised segmentation benchmarks, including LA2018, KiTS19, and LiTS, across different modalities (i.e., MR and CT), verify the superiority of our SGTC in comparison with recent state-of-the-art methods.

Related Work

Semi-supervised medical image segmentation methods: Recently, learning from a constrained pool of labeled data alongside copious amounts of unlabeled data becomes a pragmatic approach in medical image analysis domain (Bai et al. 2023; Cai et al. 2021b). Existing semi-supervised medical image segmentation methods can be classified into two groups: pseudo-label-based methods (Wang et al. 2022; Thompson, Di Caterina, and Voisey 2022; Yu et al. 2019; Tarvainen and Valpola 2017) and consistency-based methods (Chen et al. 2021; Li, Zhang, and He 2020; Luo et al. 2021; Zhang et al. 2024c; Huang et al. 2024). The pseudo-label-based methods, like MT (Tarvainen and Valpola 2017) and UA-MT (Yu et al. 2019), estimate the pseudo labels based on a few labeled samples. Further, BCP (Bai et al. 2023) utilizes a bidirectional copy-paste method to reduce the distribution gap between labeled and unlabeled data. The consistency-based methods, like SASSNet (Li, Zhang, and He 2020), CPS (Chen et al. 2021), and DTC (Luo et al. 2021), which employ consistency regularization among different sub-networks. Nevertheless, the above methods still need to fully annotate a volumetric sample, thus limiting their applications in clinical practice. To solve this, methods using sparse annotations, such as PLN (Li et al. 2022) and Desco (Cai et al. 2023) utilize the registration methods to propagate few labeled slices to others. Unfortunately, these methods have limitations in semantic understanding, which leads to less accurate boundary segmentation (Xu et al. 2023). Besides, due to the absence of information from certain planes, the above methods fails to fully utilize the different planes in the 3D space, thus making it ineffective in modeling the complex distribution of the entire volume.

Text-guided methods: Contrastive Language-Image Pre-training (CLIP (Radford et al. 2021)) is gaining popularity and has achieved impressive results in various downstream tasks (Guo et al. 2023; Wang et al. 2023b; Yu et al. 2022; Wang et al. 2023a; Li et al. 2024). Particularly, in the medical image segmentation domain, accurately extracting intricate structures and delineating weak boundaries typically hinges on understanding the semantic nuances within the images. Nowadays, some researchers have started to investigate cross-modal networks in the medical imaging community (Chen, Li, and Wan 2022; Cong et al. 2022; Yuan et al. 2023). For instance, Huang *et al.* (Huang et al. 2021) learns global and local representations of images by comparing subregions of images with words from medical reports. Li *et al.* (Li et al. 2023a) establish a multi-modal dataset containing X-ray and CT images, supplemented with medical text annotations to address quality issues in manually annotated images. Moreover, Liu *et al.* (Liu et al. 2023) adopt the text embedding extracted by CLIP as parameters on the image features. Exploring these foundation models for data-efficient medical image segmentation is still limited, but is highly necessary. To this end, this paper makes one of the first attempts to propose a novel semantic-guided triple co-training framework, which leverages text representations to enhance semi-supervised learning to harness more discriminative semantic information, achieving semantic-aware and fine-granular semi-supervised medical image segmentation.

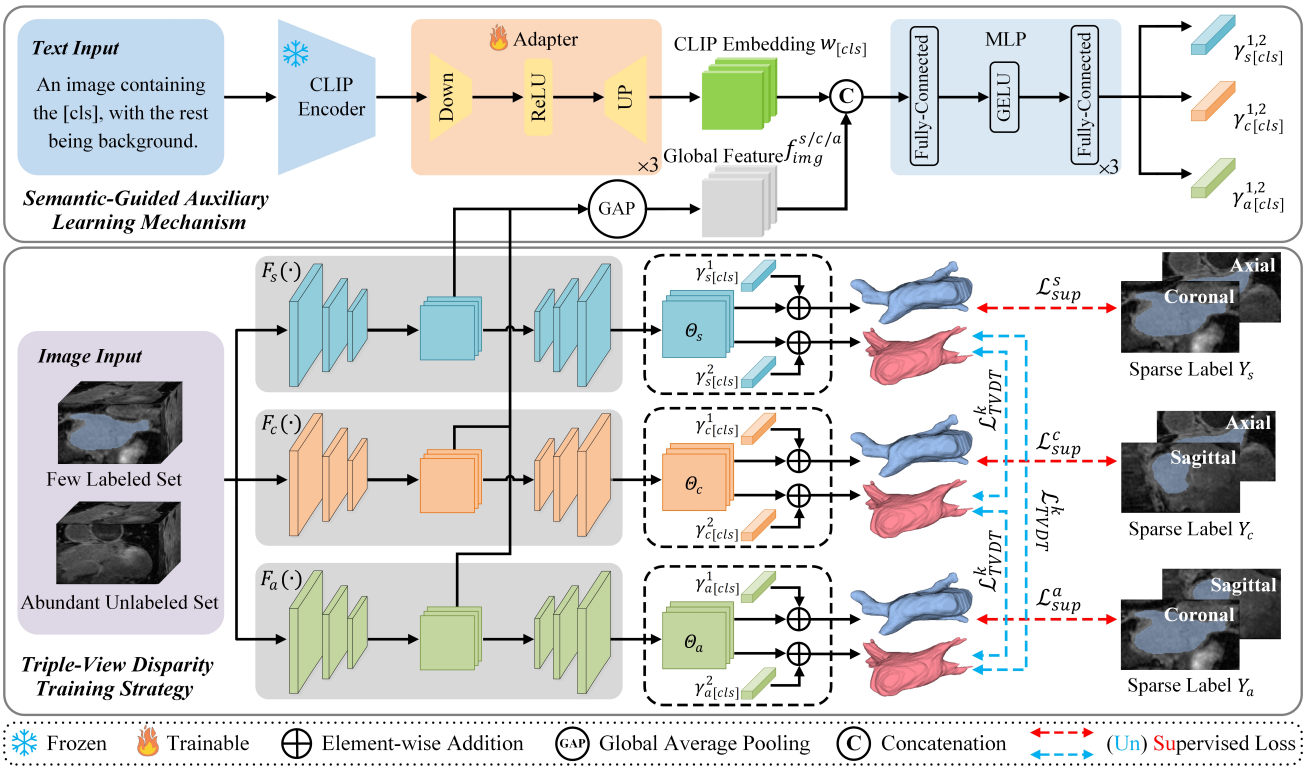

Figure 2: Architecture of the proposed SGTC framework. For volumes with sparse orthogonal labels, each volume has three corresponding labels. For model $F_s(\cdot)$, the supervision signals are selected from the Coronal and Axial plane, for model $F_c(\cdot)$ from the Sagittal and Axial plane, and for model $F_a(\cdot)$ from the Coronal and Sagittal planes. For volumes without labels, each segmentation result of $F_s(\cdot)$, $F_c(\cdot)$, and $F_a(\cdot)$ act as cross-supervision signals for other sub-networks.

Methodology

Preliminaries

We first define the preliminaries of this work. Specifically, given a training dataset D, the labeled set containing M labeled cases, represented as $D_L = \{(X_i^l, Y_i)\}_{i=1}^M$, where X_i^l denotes the input images and y_i denotes the corresponding ground truth. Additionally, the unlabeled set D_U containing N unlabeled cases, represented as $D_U = \{X_i^u\}_{i=1}^N$, where $N \gg M$. Our method annotates three slices from three orthogonal planes, including X_{is}^{lp} (s means sagittal), X_{ic}^{lq}(c means coronal), and X_{ia}^{lr}(a means axial). There are p^{th} slice in plane s, q^{th} slice in plane c, and r^{th} slice in plane a. Their annotations are Y_{is}^p, Y_{ic}^q, and Y_{ia}^r, respectively.

Model Architecture

Figure 2 illustrates our proposed Semantic-Guided Triplet Co-Training (SGTC) framework, which consists of two main components: semantic-guided auxiliary learning mechanism and triple-view disparity training strategy. Benefiting from these, our SGTC introduces text representations to enhance semi-supervised learning to exploit more discriminative semantic information. Then, we will elaborate on the technical details of each component step by step.

Semantic-Guided Auxiliary Learning Mechanism (SGAL): While some methods have incorporated textual semantics to guide segmentation, they remain fully supervised and require large amounts of labeled data, limiting their clinical applicability (Li et al. 2023a; Liu et al. 2023; Shin et al. 2022). To address this, we propose a novel semantic-guided auxiliary learning mechanism, utilizing the text representations from pre-trained CLIP to enable semantic-aware and fine-granular semi-supervised medical image segmentation and enhance the quality of pseudo labels.

Specifically, the designed medical prompts are processed through the pre-trained text encoder of CLIP to obtain the text embedding w. Since the CLIP is pre-trained on natural images (i.e., the domain gap between natural images and medical images (Ye et al. 2022)), it may prevent the CLIP text encoder from fully capturing the clinical semantics contained in medical prompts. Therefore, we freeze the pre-trained CLIP and fine-tune the Adapter module followed by the frozen CLIP encoder. The Adapter module consists of a dimension reduction projection layer followed by an activation function layer, and an up projection layer. The calculation process of w can be written as:

$$w = \text{Adapter}(\text{CLIP}_{Enc}(\text{Text Prompt})). \quad (1)$$

In this paper, the text prompts are defined as "*An image containing the [CLS], with the rest being background*", where [CLS] is a concrete class name. It is well known that the template of medical prompts is crucial, and therefore the

effectiveness of different prompt templates will be verified in the following studies. After obtaining the text embedding w, we concatenate w with the global image feature f_{img} extracted through the encoder path of the segmentation network to better align the image-text modality. Subsequently, this concatenated representation is directly fed into a multilayer perception (MLP) to obtain the cross-modal parameters $\gamma^{1,2}_{s/c/a}$, which can be formulated as follows:

$$\gamma^{1,2}_{s/c/a} = \text{MLP}(w \copyright f^{s/c/a}_{img}), \tag{2}$$

where © represents concatenation. Then, we perform element-wise addition of the cross-modal parameters γ and the features extracted before the final classification layer $\Theta_{s/c/a}$ of each sub-networks, and pass it through a convolution layer to obtain the predictions $P_{s/c/a}$ of each branch.

$$P_{s/c/a} = \text{Conv3D}(\Theta_{s/c/a} \oplus \gamma^{1,2}_{s/c/a}), \tag{3}$$

where $\oplus$ represents element-wise addition, and Conv3D denotes the 3D convolution layer.

Triple-View Disparity Training Strategy (TVDT): The previous sparse annotation approach, which annotated slices on only one or two planes, failed to fully preserve spatial information, leading to a suboptimal segmentation results. To address this issue, we proposed a novel triple-view disparity training strategy, which significantly improves the robustness by maintaining the disparity of sub-networks during training as well as allowing the sub-networks to learn complementary knowledge from each other. More importantly, it just needs the clinician to annotate three orthogonal slices of a few volumetric samples.

Specifically, for a volume X^l, we first employ three distinct sparse labels Y_s, Y_c, and Y_a to supervise the three sub-networks respectively. This ensures better consistency training while maintaining the disagreement among different sub-networks. For each sub-network, we select two orthogonal annotated slices from the three orthogonal planes as supervision signals. For the sparse label Y_s, we choose Y^q_c slice and Y^r_a slice, which can be formulated as follows:

$$Y_s = Y \otimes W_s, \tag{4}$$

$$W^i_s = \begin{cases} 1, & \text{if voxel } i \text{ is on the selected slices,} \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where $\otimes$ denotes element-wise multiplication. W_s means the weight matrix. Similarly, for Y_c, we choose Y^p_s and Y^r_a slice, and for Y_a, Y^p_s and Y^q_c slice are allocate.

Our SGTC comprises three 3D segmentation networks (i.e., Vnet (Milletari, Navab, and Ahmadi 2016)), denoted as $F_s(\cdot)$, $F_c(\cdot)$, and $F_a(\cdot)$. As mentioned above, every volume X^l_i ($i \leq M$) merely has three orthogonal labels Y_s ,Y_c and Y_a. Sub-network $F_s(\cdot)$ is trained with Y_s, $F_c(\cdot)$ is trained with Y_c, and $F_a(\cdot)$ is trained with Y_a, respectively. Under this setting, all three sub-networks can learn knowledge from different planes of the others while maintaining key spatial information of 3D medical volumes.

For abundant unlabeled data, the triplet sub-networks guide each other in turn, where the predictions generated by model $F_s(\cdot)$ serve as pseudo labels for $F_c(\cdot)$ and $F_a(\cdot)$. Here, following (Yu et al. 2019), we compute uncertainty using Monte Carlo dropout, selecting voxels with uncertainties lower than a threshold as better pseudo labels for triplet co-training. Similarly, the predicted results of $F_c(\cdot)$ and $F_a(\cdot)$ serve as pseudo labels for the other two sub-networks. The loss is computed using weighted cross-entropy loss as:

$$\mathcal{L}^k_{TVDT} = -\frac{1}{\sum_{i=1}^{H\times W\times D} m_i} \sum_{i=1}^{H\times W\times D} m_i \hat{y}^k_i \log p_i, k = 1, 2, \tag{6}$$

where m_i denotes whether the i^{th} voxel is selected. p_i is the prediction result of the current model, and $\hat{y}^1_i$ and $\hat{y}^2_i$ are the pseudo labels generated by the other two sub-networks.

In addition, for limited labeled data, the supervised loss contains weighted cross-entropy loss and weighted dice loss, which can be defined as follows:

$$\mathcal{L}_{WCE} = -\frac{1}{\sum_{i=1}^{H\times W\times D} w_i} \sum_{i=1}^{H\times W\times D} w_i y_i \log p_i, \tag{7}$$

$$\mathcal{L}_{Dice} = 1 - \frac{2 \times \sum_{i=1}^{H\times W\times D} w_i p_i y_i}{\sum_{i=1}^{H\times W\times D} w_i (p^2_i + y^2_i)}, \tag{8}$$

where w_i is the i^{th} voxel value of the weight matrix. p_i and y_i respectively represent the predicted results and the ground truth labels of the network on voxel i. Therefore, the supervised learning loss is represented as follows:

$$\mathcal{L}_{Sup} = \frac{1}{2}\mathcal{L}_{WCE} + \frac{1}{2}\mathcal{L}_{Dice}. \tag{9}$$

Overall, the total loss during training is defined as:

$$\mathcal{L}_{SGTC} = (1-\alpha)\mathcal{L}_{Sup} + \alpha\mathcal{L}^k_{TVDT}, \tag{10}$$

where the first term $\mathcal{L}_{Sup}$ is tailored for labeled data, and the second term $\mathcal{L}^k_{TVDT}$ is employed for unlabeled data. α denotes the proposed dynamic parameter, which enables the supervised loss from the three annotated slices to dominate the training at the first few epochs, and then gradually increase the weight of the unsupervised loss to stabilize the entire semi-supervised learning.

Experimental Results

Datasets & Metrics

LA2018 Dataset (Xiong et al. 2021) contains 100 gadolinium enhanced MR imaging scans with labels. All scans have the same isotropic resolution of $0.625 \times 0.625 \times 0.625 mm^3$. Following existing works (Yu et al. 2019; Luo et al. 2021), we utilize 80 training samples and 20 testing samples for fair comparison with other methods.

KiTS19 Dataset (Heller et al. 2019) is provided by the Medical Centre of Minnesota University and consists of 300 abdominal CT scans. The slice thickness ranges from 1mm to 5mm. The dataset is divided into 190 training samples and 20 testing samples.

LiTS Dataset (Bilic et al. 2023) is a CT dataset focused on liver and liver tumour segmentation. The dataset collects 201 abdominal scans. Among these, 131 scans with segmentation masks are publicly available. We use the same split of

Method	Labeled Slices	Scans Used		Metrics			
		Labeled	Unlabeled	Dice ↑	Jaccard ↑	HD ↓	ASD ↓
MT (Tarvainen and Valpola 2017) (NIPS'17)	CA	8	72	0.661 ± 0.124	0.505 ± 0.136	38.681 ± 9.600	13.597 ± 3.730
UA-MT (Yu et al. 2019) (MICCAI'19)	CA	8	72	0.650 ± 0.096	0.489 ± 0.108	40.442 ± 8.739	14.841 ± 4.041
SASSNet (Li, Zhang, and He 2020) (MICCAI'20)	CA	8	72	0.617 ± 0.119	0.456 ± 0.121	41.913 ± 9.348	15.521 ± 4.717
CPS (Chen et al. 2021) (CVPR'21)	CA	8	72	0.661 ± 0.082	0.499 ± 0.092	38.718 ± 6.958	13.992 ± 3.158
DTC (Luo et al. 2021) (AAAI'21)	CA	8	72	0.686 ± 0.120	0.533 ± 0.133	35.624 ± 8.651	11.556 ± 4.035
BCP (Bai et al. 2023) (CVPR'23)	CA	8	72	0.784 ± 0.093	0.653 ± 0.115	22.432 ± 8.282	5.753 ± 2.667
Desco (Cai et al. 2023) (CVPR'23)	CA	8	72	0.711 ± 0.093	0.559 ± 0.111	35.671 ± 7.134	12.776 ± 3.532
SGTC (Dual)	CA	8	72	0.739 ± 0.078	0.592 ± 0.097	35.464 ± 9.110	11.503 ± 4.015
SGTC (Ours)	SCA	8	72	$\mathbf{0.847 \pm 0.044}$	$\mathbf{0.738 \pm 0.066}$	$\mathbf{17.442 \pm 11.719}$	$\mathbf{4.256 \pm 3.836}$

Table 1: Quantitative comparisons with the seven state-of-the-art methods on the LA2018 dataset under 10% labeled cases. In this paper, **bold** values denote the best-performing method. S, C, and A indicate sagittal, coronal, and axial annotated slices.

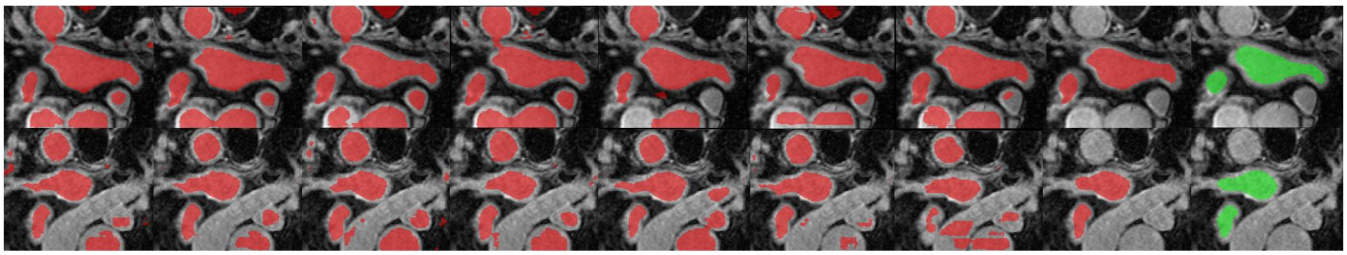

Figure 3: Qualitative comparisons on the LA2018 dataset with 10% labeled cases. From left to right: segmentation results of MT (Tarvainen and Valpola 2017), UAMT (Yu et al. 2019), SASSNet (Li, Zhang, and He 2020), CPS (Chen et al. 2021), DTC (Luo et al. 2021), BCP (Bai et al. 2023), Desco (Cai et al. 2023), our SGTC and ground truth (GT), respectively.

the dataset as done in (Cai et al. 2023), where 100 scans are used for training, and the remaining 31 scans for test.

To fairly compare our method with others, we adopt four commonly used metrics: Dice similarity coefficient (Dice), Jaccard similarity coefficient (Jaccard), 95% Hausdorff Distance (95HD), and Average Surface Distance (ASD).

Implementation Details

To make sparse annotations more effective, three selected slices should contain the foreground area of the segmentation target, and we choose slices as close to the center position as possible in all three planes. The entire training is conducted in the PyTorch framework. We set the batch size to 4, with each batch containing two volumes with labels and two volumes without labels. We train for 6000 iterations using the Stochastic Gradient Descent (SGD) optimizer. The initial learning rate is set to 0.01 and gradually decays to 0.0001. The value of parameter α is initialized to 0.1 and is increased every 150 iterations.

Comparison Experiments

In this paper, we compare our SGTC with 7 recent SOTAs semi-supervise medical image segmentation methods, including MT (Tarvainen and Valpola 2017), UA-MT (Yu et al. 2019), SASSNet (Li, Zhang, and He 2020), CPS (Chen et al. 2021), DTC (Luo et al. 2021), BCP (Bai et al. 2023) and Desco (Cai et al. 2023) on three challenging benchmarks, i.e., LA2018 (Xiong et al. 2021), KiTS19 (Heller et al. 2019), and LiTS (Heller et al. 2019) datasets. In other methods and the SGTC (Dual) version, we annotate two slices per volume according to the previous settings (Cai et al. 2023). In the SGTC version, we annotate three slices per volume. The S, C, A indicates whether the annotated slices is sagittal, coronal or axial plane.

Comparison Results on LA2018 dataset: Table 1 compares our method with these state-of-the-art models on LA2018 dataset under 10% (8 samples), which shows our method achieves the best performance (i.e., surpassing the second best by 6.3% on Dice). Due to the semantic guidance, the network can comprehensively perceive subtle changes in boundary regions and voxel-level semantic information, which makes the network perform better in dealing with complex anatomy structures. While methods with two orthogonal annotations perform poorly, our SGTC with three orthogonal annotations achieves better shape-related performance by modeling the entire volume's supervision signal distribution more comprehensively. Visual comparisons are shown in Figure 3, where our SGTC accurately delineates the intricate structures and weak boundaries.

Comparison Results on KITS19 dataset: Table 2 demonstrates the comparison between our method and recent SOTAs on KITS19 dataset under 10% (19 samples) labeled data settings. We can find that our SGTC outperforms existing methods in all metrics. For example, compared with BCP (Bai et al. 2023) and DTC (Luo et al. 2021), SGTC shows 1.2% and 1.5% improvements on Dice. Besides, the qualitative comparison results are shown in Figure 4, which shows our method effectively handles complex boundaries.

Comparison Results on LITS dataset: Table 3 compares our method with others on the LITS dataset under 10% (10 samples) labeled data. It can be observed that our method's advantage becomes more pronounced with less labeled data. The main reason behind this is that our method achieves semantic-aware and fine-granular segmentation. Moreover,

Method	Labeled Slices	Scans Used		Metrics			
		Labeled	Unlabeled	Dice ↑	Jaccard ↑	HD ↓	ASD ↓
MT (Tarvainen and Valpola 2017) (NIPS'17)	CA	19	171	0.780 ± 0.084	0.647 ± 0.113	43.797 ± 13.586	16.152 ± 6.264
UA-MT (Yu et al. 2019) (MICCAI'19)	CA	19	171	0.856 ± 0.096	0.758 ± 0.137	32.081 ± 15.390	9.120 ± 4.739
SASSNet (Li, Zhang, and He 2020) (MICCAI'20)	CA	19	171	0.914 ± 0.054	0.845 ± 0.087	22.130 ± 15.827	5.856 ± 3.551
CPS (Chen et al. 2021) (CVPR'21)	CA	19	171	0.813 ± 0.106	0.697 ± 0.146	42.438 ± 9.210	13.754 ± 5.482
DTC (Luo et al. 2021) (AAAI'21)	CA	19	171	0.918 ± 0.058	0.853 ± 0.094	12.049 ± 16.383	3.174 ± 3.084
BCP (Bai et al. 2023) (CVPR'23)	CA	19	171	0.921 ± 0.046	0.864 ± 0.083	6.726 ± 11.021	2.888 ± 3.807
Desco (Cai et al. 2023) (CVPR'23)	CA	19	171	0.880 ± 0.099	0.798 ± 0.145	21.567 ± 18.906	6.257 ± 4.967
SGTC (Dual)	CA	19	171	0.927 ± 0.059	0.870 ± 0.092	7.821 ± 9.638	2.639 ± 1.996
SGTC (Ours)	SCA	19	171	**0.933 ± 0.041**	**0.877 ± 0.067**	**5.145 ± 6.983**	**2.038 ± 1.868**

Table 2: Quantitative comparisons with the seven state-of-the-art methods on the KITS19 dataset under 10% labeled cases.

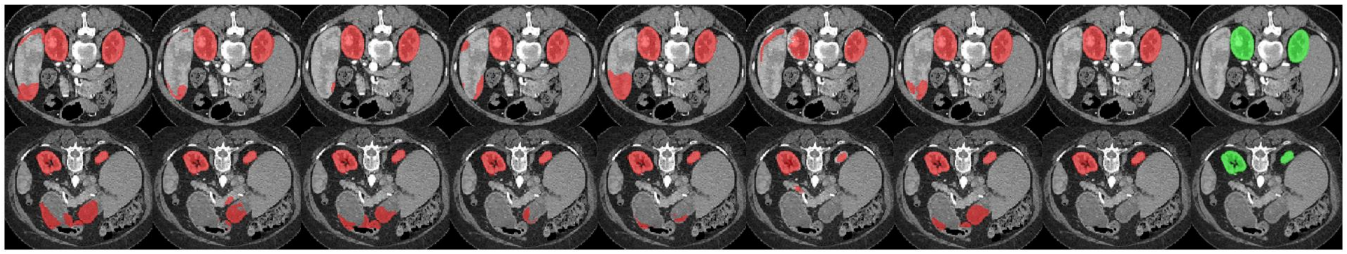

Figure 4: Qualitative comparisons on the KITS19 dataset with 10% labeled cases.

Method	Labeled Slices	Scans Used		Metrics			
		Labeled	Unlabeled	Dice ↑	Jaccard ↑	HD ↓	ASD ↓
MT (Tarvainen and Valpola 2017) (NIPS'17)	CA	10	90	0.829 ± 0.114	0.721 ± 0.137	46.629 ± 20.668	13.122 ± 7.901
UA-MT (Yu et al. 2019) (MICCAI'19)	CA	10	90	0.781 ± 0.212	0.677 ± 0.221	20.097 ± 15.522	5.277 ± 4.343
SASSNet (Li, Zhang, and He 2020) (MICCAI'20)	CA	10	90	0.830 ± 0.119	0.724 ± 0.143	46.115 ± 12.276	13.822 ± 5.276
CPS (Chen et al. 2021) (CVPR'21)	CA	10	90	0.827 ± 0.089	0.713 ± 0.122	47.098 ± 11.201	12.521 ± 4.162
DTC (Luo et al. 2021) (AAAI'21)	CA	10	90	0.896 ± 0.071	0.817 ± 0.107	18.926 ± 16.982	5.519 ± 4.432
BCP (Bai et al. 2023) (CVPR'23)	CA	10	90	0.922 ± 0.060	0.860 ± 0.094	11.224 ± 15.202	3.294 ± 4.121
Desco (Cai et al. 2023) (CVPR'23)	CA	10	90	0.885 ± 0.055	0.798 ± 0.083	16.095 ± 11.683	4.323 ± 3.501
SGTC (Dual)	CA	10	80	0.919 ± 0.049	0.854 ± 0.079	10.960 ± 13.698	3.166 ± 3.926
SGTC (Ours)	SCA	10	90	**0.927 ± 0.044**	**0.867 ± 0.071**	**9.302 ± 11.694**	**2.710 ± 3.104**

Table 3: Quantitative comparisons with the seven state-of-the-art methods on the LITS dataset under 10% labeled cases.

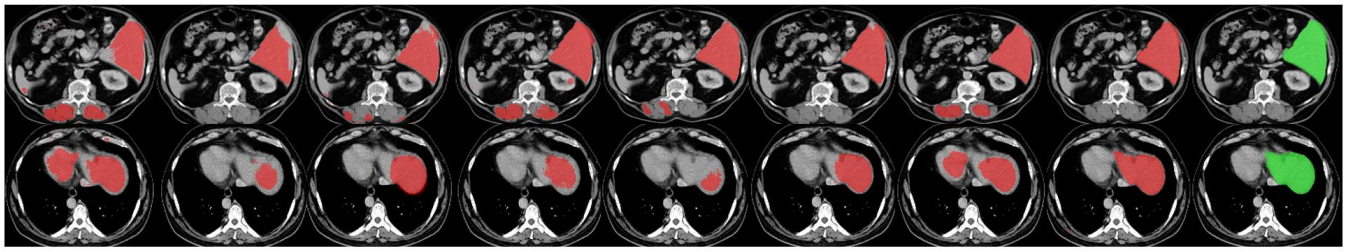

Figure 5: Qualitative comparisons on the LITS dataset with 10% labeled cases.

visual results in Figure 5 that show our SGTC effectively handles complex boundaries and fine structures.

Ablation Study and Analysis

Ablation study of each component: As shown in Table 4, ablation experiment on the KiTS19 dataset with 9 volumes to better evaluate the components of our method. By incorporating semantic cues, our method is able to focus more effectively on boundary information, improving performance.

Ablation study of triple-view disparity training strategy: As shown in Table 5, where all these methods use the same annotated strategy (i.e., CAC or SCA). The results indicate that our method outperforms others under the same settings, and SCA provides better performance improvement, demonstrating that the performance gains are due to

SGAL	TVDT	Dice ↑	Jaccard ↑	HD ↓	ASD ↓
×	×	0.780	0.655	41.630	14.110
✓	×	0.892	0.814	10.945	3.318
×	✓	0.896	0.817	11.985	3.106
✓	✓	**0.915**	**0.837**	**9.968**	**2.668**

Table 4: Ablation study of components on KiTS19 dataset.

Method	Labeled Slices	Scan Used		Dice
		Labeled	Unlabeled	
MT	CAC / SCA	9	181	0.782 / 0.809
UA-MT	CAC / SCA	9	181	0.851 / 0.871
SASSNet	CAC / SCA	9	181	0.885 / 0.897
CPS	CAC / SCA	9	181	0.826 / 0.841
DTC	CAC / SCA	9	181	0.887 / 0.896
BCP	CAC / SCA	9	181	0.909 / 0.913
Desco	CAC / SCA	9	181	0.828 / 0.855
SGTC (Ours)	CAC / SCA	9	181	**0.911 / 0.915**

Table 5: Ablation study of the proposed TVDT on KITS19 .

Text prompts	Scan Used		Dice
	Labeled	Unlabeled	
None.	6	184	0.873
A photo of a [cls].	6	184	0.882
There is a [cls] in this computerized tomography/ magnetic resonance imaging.	6	184	0.883
An image containing the [cls], with the rest being background.	6	184	**0.888**

Table 6: Ablation study of different text prompts.

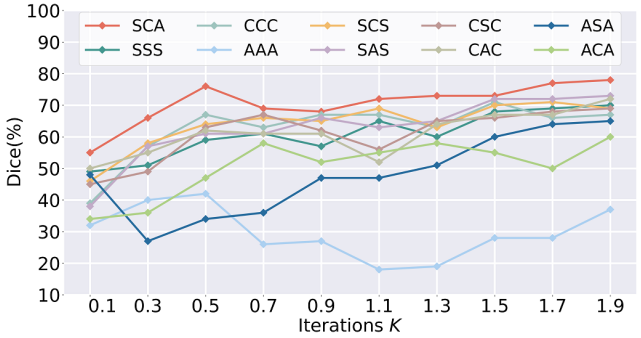

Figure 6: Performance comparisons of different annotations.

the more effective triple-view disparity training strategy.

Ablation study of text prompts in CLIP: Table 6 shows the results of different text prompts. It can be observed that, compared to solely using image features, combining visual and textual modalities brings large performance gains. Our tailored text prompt, integrating semantics into the network by describing more details, result in superior outcomes.

Ablation study of the different annotation method: Figure 6 illustrates segmentation performance using our proposed three orthogonal annotations, two orthogonal annotations (i.e., the third slice parallel to one of the first two), and three parallel annotations (i.e., slices in the same plane). It is shown that our annotation scheme achieves superior results at different iterations. Additionally, Figure 7 shows t-SNE visualization of the extracted features using different annotation methods. Specifically, we trained five networks with just one single annotated slice per volume: three orthogonal slices for training s, c, and a_1, and three parallel slices for training a_1, a_2, and a_3. Features from three orthogonal annotations are more concentrated, demonstrating the effectiveness of our triple-view disparity training strategy.

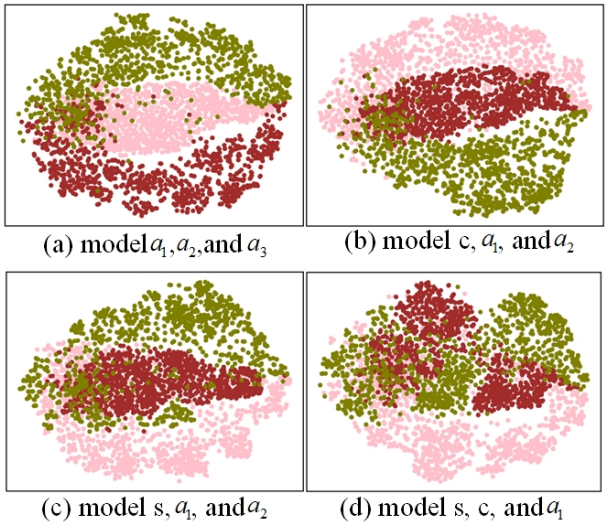

Figure 7: The t-SNE visualization of feature representations extracted from models trained on (a) three parallel slices, (b), (c) two orthogonal slices. (d) three orthogonal slices.

Hyper-Parameter	Scan Used		Dice	Jaccard
	Labeled	Unlabeled		
$\alpha = 0.1$	38	152	0.942	0.891
$\alpha = 0.15$	38	152	0.944	0.895
$\alpha = 0.2$	38	152	0.945	0.897
$\alpha = 0.25$	38	152	0.929	0.870
$\alpha = 0.3$	38	152	0.936	0.882
The proposed α	38	152	**0.947**	**0.901**

Table 7: Anaysis of hyper-parameter α.

Parameter Analysis: Table 7 shows the analysis of the hyper-parameter α in Eq. 10. Specifically, we conducted ablation experiments on the KITS19 dataset by setting α to 0.1, 0.15, 0.2, 0.25, 0.3, and the proposed dynamic coefficient. It is observed that the proposed dynamic coefficient yields the best performance, as it makes the training more stable.

Conclusion

This paper presents a novel Semantic-Guided Triplet Co-training framework (SGTC), for accurate and clinically realistic semi-supervised medical image segmentation, which consists of two major contributions. The proposed semantic-guided auxiliary learning mechanism that generates high-quality pseudo labels for semantic-aware and fine-granular segmentation. The proposed triple-view disparity training strategy that requires annotating only three orthogonal slices, enhancing sub-network diversity and robustness. Extensive experiments and ablations conducted on three challenging benchmarks demonstrate the effectiveness of our proposed SGTC, showcasing its superiority over most state-of-the-art methods. **Limitations.** Our SGTC exhibits performance degradation when the selected slices contain limited foreground information. In future work, we plan to optimize our approach to address this limitation.

Acknowledgments

This work was supported in part by the National Science Foundation of China under Grant62471448, 62102338; in part by Shandong Provincial Natural Science Foundation under Grant ZR2024YQ004; in part by TaiShan Scholars Youth Expert Program of Shandong Province under Grant No.tsqn202312109.

References

Avants, B. B.; Tustison, N. J.; Song, G.; Cook, P. A.; Klein, A.; and Gee, J. C. 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage*, 54(3): 2033–2044.

Bai, W.; Oktay, O.; Sinclair, M.; Suzuki, H.; Rajchl, M.; Tarroni, G.; Glocker, B.; King, A.; Matthews, P. M.; and Rueckert, D. 2017. Semi-supervised learning for network-based cardiac MR image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention.*, 253–260. Springer.

Bai, Y.; Chen, D.; Li, Q.; Shen, W.; and Wang, Y. 2023. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11514–11524.

Bilic, P.; Christ, P.; Li, H. B.; Vorontsov, E.; Ben-Cohen, A.; Kaissis, G.; Szeskin, A.; Jacobs, C.; Mamani, G. E. H.; Chartrand, G.; et al. 2023. The liver tumor segmentation benchmark (LiTS). *Medical Image Analysis*, 84: 102680.

Cai, H.; Li, S.; Qi, L.; Yu, Q.; Shi, Y.; and Gao, Y. 2023. Orthogonal annotation benefits barely-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3302–3311.

Cai, Q.; Liu, H.; Qian, Y.; Zhou, S.; Duan, X.; and Yang, Y.-H. 2019. Saliency-guided level set model for automatic object segmentation. *Pattern Recognition*, 93: 147–163.

Cai, Q.; Liu, H.; Qian, Y.; Zhou, S.; Wang, J.; and Yang, Y.-H. 2021a. A novel hybrid level set model for non-rigid object contour tracking. *IEEE Transactions on Image Processing*, 31: 15–29.

Cai, Q.; Liu, H.; Zhou, S.; Sun, J.; and Li, J. 2018. An adaptive-scale active contour model for inhomogeneous image segmentation and bias field estimation. *Pattern Recognition*, 82: 79–93.

Cai, Q.; Qian, Y.; Zhou, S.; Li, J.; Yang, Y.-H.; Wu, F.; and Zhang, D. 2021b. AVLSM: Adaptive variational level set model for image segmentation in the presence of severe intensity inhomogeneity and high noise. *IEEE Transactions on Image Processing*, 31: 43–57.

Chen, D.; Bai, Y.; Shen, W.; Li, Q.; Yu, L.; and Wang, Y. 2023. 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*, 23869–23878.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2613–2622.

Chen, Z.; Li, G.; and Wan, X. 2022. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5152–5161.

Cong, F.; Xu, S.; Guo, L.; and Tian, Y. 2022. Caption-aware medical VQA via semantic focusing and progressive cross-modality comprehension. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3569–3577.

Guo, S.; Cai, Q.; Qi, L.; and Dong, J. 2023. CLIP-Hand3D: Exploiting 3D Hand Pose Estimation via Context-Aware Prompting. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4896–4907.

Heller, N.; Sathianathen, N.; Kalapara, A.; Walczak, E.; Moore, K.; Kaluzniak, H.; Rosenberg, J.; Blake, P.; Rengel, Z.; Oestreich, M.; et al. 2019. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*.

Huang, H.; Huang, Y.; Xie, S.; Lin, L.; Tong, R.; Chen, Y.-W.; Li, Y.; and Zheng, Y. 2024. Combinatorial CNN-Transformer Learning with Manifold Constraints for Semi-supervised Medical Image Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2330–2338.

Huang, S.-C.; Shen, L.; Lungren, M. P.; and Yeung, S. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3942–3951.

Li, S.; Cai, H.; Qi, L.; Yu, Q.; Shi, Y.; and Gao, Y. 2022. PLN: Parasitic-like network for barely supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(3): 582–593.

Li, S.; Cai, Q.; Zhang, F.; Zhang, M.; Shu, Y.; Liu, Z.; Li, H.; and Liu, L. 2024. PP-SSL: Priority-Perception Self-Supervised Learning for Fine-Grained Recognition. *arXiv preprint arXiv:2412.00134*.

Li, S.; Zhang, C.; and He, X. 2020. Shape-aware semi-supervised 3D semantic segmentation for medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention.*, 552–561. Springer.

Li, X.; Yu, L.; Chen, H.; Fu, C.-W.; Xing, L.; and Heng, P.-A. 2020. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2): 523–534.

Li, Z.; Li, Y.; Li, Q.; Wang, P.; Guo, D.; Lu, L.; Jin, D.; Zhang, Y.; and Hong, Q. 2023a. Lvit: language meets vision transformer in medical image segmentation. *IEEE Transactions on Medical Imaging*.

Li, Z.; Zheng, Y.; Luo, X.; Shan, D.; and Hong, Q. 2023b. ScribbleVC: Scribble-supervised Medical Image Segmentation with Vision-Class Embedding. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3384–3393.

Liu, J.; Zhang, Y.; Chen, J.-N.; Xiao, J.; Lu, Y.; A Landman, B.; Yuan, Y.; Yuille, A.; Tang, Y.; and Zhou, Z. 2023. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21152–21164.

Luo, X.; Chen, J.; Song, T.; and Wang, G. 2021. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8801–8809.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth International Conference on 3D Vision*, 565–571. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Shen, Z.; Cao, P.; Yang, H.; Liu, X.; Yang, J.; and Zaiane, O. R. 2023. Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 4199–4207.

Shin, H.; Kim, H.; Kim, S.; Jun, Y.; Eo, T.; and Hwang, D. 2022. COSMOS: cross-modality unsupervised domain adaptation for 3D medical image segmentation based on target-aware domain translation and iterative self-training. *arXiv preprint arXiv:2203.16557*.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30.

Thompson, B. H.; Di Caterina, G.; and Voisey, J. P. 2022. Pseudo-label refinement using superpixels for semi-supervised brain tumour segmentation. In *2022 IEEE 19th International Symposium on Biomedical Imaging*, 1–5. IEEE.

Wang, A.; Chen, H.; Lin, Z.; Ding, Z.; Liu, P.; Bao, Y.; Yan, W.; and Ding, G. 2023a. Hierarchical Prompt Learning Using CLIP for Multi-label Classification with Single Positive Labels. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5594–5604.

Wang, H.; and Li, X. 2024. Towards generic semi-supervised framework for volumetric medical image segmentation. *Advances in Neural Information Processing Systems*, 36.

Wang, X.; Rigall, E.; An, X.; Li, Z.; Cai, Q.; Zhang, S.; and Dong, J. 2024. A New Benchmark and Low Computational Cost Localization Method for Cephalometric Analysis. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, X.; Yuan, Y.; Guo, D.; Huang, X.; Cui, Y.; Xia, M.; Wang, Z.; Bai, C.; and Chen, S. 2022. SSA-Net: Spatial self-attention network for COVID-19 pneumonia infection segmentation with semi-supervised few-shot learning. *Medical Image Analysis*, 79: 102459.

Wang, Y.; Huang, S.; Gao, Y.; Wang, Z.; Wang, R.; Sheng, K.; Zhang, B.; and Liu, S. 2023b. Transferring CLIP's Knowledge into Zero-Shot Point Cloud Semantic Segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3745–3754.

Wu, Y.; Chen, J.; Yan, J.; Zhu, Y.; Chen, D. Z.; and Wu, J. 2023. GCL: Gradient-Guided Contrastive Learning for Medical Image Segmentation with Multi-Perspective Meta Labels. In *Proceedings of the 31st ACM International Conference on Multimedia*, 463–471.

Xiong, Z.; Xia, Q.; Hu, Z.; Huang, N.; Bian, C.; Zheng, Y.; Vesal, S.; Ravikumar, N.; Maier, A.; Yang, X.; et al. 2021. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical Image Analysis*, 67: 101832.

Xu, X.; Xiong, T.; Ding, Z.; and Tu, Z. 2023. Masqclip for open-vocabulary universal image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 887–898.

Ye, Y.; Liu, Z.; Zhang, Y.; Li, J.; and Shen, H. 2022. Alleviating style sensitivity then adapting: Source-free domain adaptation for medical image segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1935–1944.

Yu, L.; Wang, S.; Li, X.; Fu, C.-W.; and Heng, P.-A. 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention.*, 605–613. Springer.

Yu, Y.; Zhan, F.; Wu, R.; Zhang, J.; Lu, S.; Cui, M.; Xie, X.; Hua, X.-S.; and Miao, C. 2022. Towards counterfactual image manipulation via clip. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3637–3645.

Yuan, Z.; Jin, Q.; Tan, C.; Zhao, Z.; Yuan, H.; Huang, F.; and Huang, S. 2023. Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. In *Proceedings of the 31st ACM International Conference on Multimedia*, 547–556.

Yun, B.; Xie, X.; Li, Q.; and Wang, Y. 2023. Uni-Dual: A Generic Unified Dual-Task Medical Self-Supervised Learning Framework. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3887–3896.

Zhang, F.; Liu, H.; Cai, Q.; Feng, C.-M.; Wang, B.; Wang, S.; Dong, J.; and Zhang, D. 2024a. Federated Cross-Incremental Self-Supervised Learning for Medical Image Segmentation. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, F.; Liu, H.; Cao, C.; Cai, Q.; and Zhang, D. 2022a. RVLSM: Robust variational level set method for image segmentation with intensity inhomogeneity and high noise. *Information Sciences*, 596: 439–459.

Zhang, F.; Liu, H.; Duan, X.; Wang, B.; Cai, Q.; Li, H.; Dong, J.; and Zhang, D. 2024b. DSLSM: Dual-kernel-induced statistic level set model for image segmentation. *Expert Systems with Applications*, 242: 122772.

Zhang, F.; Liu, H.; Wang, J.; Lyu, J.; Cai, Q.; Li, H.; Dong, J.; and Zhang, D. 2024c. Cross co-teaching for semi-supervised medical image segmentation. *Pattern Recognition*, 110426.

Zhang, W.; Zhang, X.; Huang, S.; Lu, Y.; and Wang, K. 2022b. Pixelseg: Pixel-by-pixel stochastic semantic segmentation for ambiguous medical images. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4742–4750.

Zhang, W.; Zhang, X.; Huang, S.; Lu, Y.; and Wang, K. 2022c. A probabilistic model for controlling diversity and accuracy of ambiguous medical image segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4751–4759.

Zhou, H.-Y.; Lu, C.; Chen, C.; Yang, S.; and Yu, Y. 2023. A unified visual information preservation framework for self-supervised pre-training in medical image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, Y.; Wang, Y.; Tang, P.; Bai, S.; Shen, W.; Fishman, E.; and Yuille, A. 2019. Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. In *2019 IEEE Winter Conference on Applications of Computer Vision*, 121–140. IEEE.