DuSSS: Dual Semantic Similarity-Supervised Vision-Language Model for Semi-Supervised Medical Image Segmentation

Qingtao Pan[1,2], Wenhao Qiao[1,2], Jingjiao Lou[1,2], Bing Ji[1,2*], Shuo Li[3]

[1]School of Control Science and Engineering, Shandong University, Jinan, China
[2]Key Laboratory of Machine Intelligence and System Control, Ministry of Education, China
[3]Department of Computer and Data Science and Department of Biomedical Engineering, Case Western Reserve University, USA
{qingtaopan33, wenhaoo.qiao, jingjiaolou, slishuo}@gmail.com, b.ji@sdu.edu.cn

Abstract

Semi-supervised medical image segmentation (SSMIS) uses consistency learning to regularize model training, which alleviates the burden of pixel-wise manual annotations. However, it often suffers from error supervision from low-quality pseudo labels. Vision-Language Model (VLM) has great potential to enhance pseudo labels by introducing text prompt guided multimodal supervision information. It nevertheless faces the cross-modal problem: the obtained messages tend to correspond to multiple targets. To address aforementioned problems, we propose a Dual Semantic Similarity-Supervised VLM (DuSSS) for SSMIS. Specifically, **1)** a Dual Contrastive Learning (DCL) is designed to improve cross-modal semantic consistency by capturing intrinsic representations within each modality and semantic correlations across modalities. **2)** To encourage the learning of multiple semantic correspondences, a Semantic Similarity-Supervision strategy (SSS) is proposed and injected into each contrastive learning process in DCL, supervising semantic similarity via the distribution-based uncertainty levels. Furthermore, a novel VLM-based SSMIS network is designed to compensate for the quality deficiencies of pseudo-labels. It utilizes the pretrained VLM to generate text prompt guided supervision information, refining the pseudo label for better consistency regularization. Experimental results demonstrate that our DuSSS achieves outstanding performance with Dice of 82.52%, 74.61% and 78.03% on three public datasets (QaTa-COV19, BM-Seg and MoNuSeg).

Code — https://github.com/QingtaoPan/DuSSS/

Introduction

Medical image segmentation aims to divide medical images into specific regions with unique attributes. It contributes to detecting abnormal areas and providing clinical guidance (Chen et al. 2021a). In clinical practice, achieving precise segmentation results necessitates manual implementation and there is an urgent need for automatic medical image segmentation to aid in diagnosis and treatment. Fully-supervised medical image segmentation methods, such as U-Net (Ronneberger, Fischer, and Brox 2015a) and its variants (Zhou et al. 2018; Özgün Çiçek et al. 2016; Fang et al. 2019), have been developed for medical image segmentation through an encoder and a decoder in a U-shaped architecture, connected by skip connections. However, fully-supervised methods need a large amount of pixel-level annotated data for model training and labeling such pixel-level annotations is laborious and requires expert knowledge especially in medical images, resulting in that labeled data are expensive or simply unavailable (Shen et al. 2023). Semi-supervised medical image segmentation (SSMIS) is a method that utilizes a small amount of labeled data and a large amount of unlabeled data to learn segmentation models (Qin, Wang, and Zhang 2024).

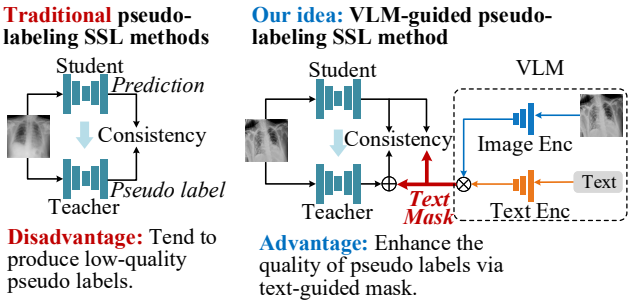

Figure 1: The VLM has potential to enhance pseudo labels via text-guided mask, improving consistency learning.

SSMIS methods, such as (Chen et al. 2021b; **?**; Wang et al. 2021; Bai et al. 2023), have shown promising performance for semi-supervised segmentation by combining consistency regularization and pseudo labeling via cross supervision between the sub-networks. However, one key disadvantage of these approaches is that they may generate low-quality pseudo labels (Ronneberger, Fischer, and Brox 2015b), which misleads model training and causes the model to learn incorrect features, thus failing to effectively utilize unlabeled data and disrupting consistency learning on unlabeled data. Therefore, the question that comes to mind is: how to effectively improve the quality of pseudo labels for SSMIS.

Vision-Language Model (VLM) has great potential to enhance the quality of pseudo-labels. It produces text-guided supervision information by leveraging textual prompts to describe visual content, thus enhancing pseudo labels for var-

*Corresponding Author

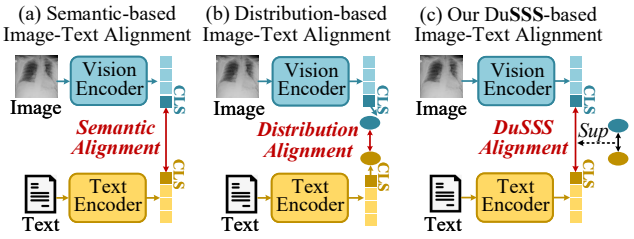

Figure 2: The limitations of current VLM methods and the solution of our DuSSS.

ious tasks (Yi et al. 2023), as shown in Fig. 1. Once successful, this approach will significantly improve the performance of SSMIS by guiding the segmentation model to locate target segmentation regions with textual prompts. Although promising performance of current VLMs, such as CLIP (Radford et al. 2021), MedCLIP (Wang et al. 2022b), MGCA (Wang et al. 2022a), etc, the cross-modal uncertainty remains a significant problem since they merely conduct one-to-one alignment between image and text (Fig. 2 (a)). Specifically, multiple images/texts may correspond to one text/image, which manifests cross-modal uncertainty.

Utilizing distribution to represent semantic embeddings is a prominent approach for uncertainty awareness (Lei et al. 2023; Chun et al. 2021; Sun et al. 2020). This approach commonly transfers semantic embeddings as distribution representations for image-text alignment through Gaussian modeling (Yang et al. 2021; Yu et al. 2019), to learn the uncertainty (Fig. 2 (b)). The distribution's variance is computed for uncertainty judgment. Although the variance reflects distribution discrepancies, relying solely on distribution representations lose original semantic attributes, leading to poor semantic associations between image and text. Therefore, we argue that it is necessary to address the uncertainty problem while detaining semantic attributes.

In this paper, we propose a novel VLM paradigm with Dual Semantic Similarity-Supervision (DuSSS) to learn multiple semantic correspondences (Fig. 2 (c)). Specifically, **1)** a Dual Contrastive Learning (DCL) is constructed for cross- and intra-modal contrastive learning. It reinforces semantic correlations across modalities and captures implicit representation relationships within each modality, thereby prompting image-text alignment. **2)** A Semantic Similarity-Supervision strategy (SSS) is proposed to supervise semantic similarity based on corresponding uncertainty levels by incorporated it into each contrastive learning process in DCL, learning potential uncertainty correspondences across modalities and within modalities. **3)** For the first time, a text-guided segmentation network is constructed to compensate for quality deficiencies of pseudo-labels for SSMIS.

Our contributions are summarized as follows:

- This is the first VLM-based method for pseudo labeling SSMIS. It compensates for the quality deficiencies of pseudo-labels by utilizing the advantages of text prompts to locate segmentation regions.
- A new SSS supervises semantic similarity via uncertainty levels. It promotes model's understanding of data pairs that are similar yet semantically ambiguous, mitigating impacts of uncertain correspondences.
- A novel DCL boosts semantic associations across modalities and captures intrinsic representation relationships within each modality, thereby boosting cross-modal semantic consistency.
- Extensive experiments are conducted on three public medical image segmentation datasets. Comprehensive results demonstrate the effectiveness of each component of our method and the superiority of our DuSSS over the state-of-the-art methods.

Related Works

Semi-Supervise Medical Image Segmentation

In the field of SSMIS, the main methods includes self-training methods (Zhu et al. 2021; Zou et al. 2018), adversarial training methods (Hung et al. 2018; Li, Zhang, and He 2020), co-training methods (Wang et al. 2021; Xia et al. 2020), and consistency regularization methods (Wang et al. 2020; Zhang et al. 2023). Consistency regularization focuses on maintaining consistent model predictions under different perturbations. The state-of-the-art technique is Mean Teacher (MT) (Tarvainen and Valpola 2017). In MT, the teacher model is employed to generate pseudo-labels for unlabeled data while maintaining prediction consistency between the teacher and student models through various regularization methods. Afterward, the teacher model is the exponential moving average (EMA) of the student model's weights. This method enables the teacher model to continually aggregate historical prediction information from unlabeled data. Subsequent improvements use different consistency regularization strategies to improve the prediction quality of unlabeled data (Chen et al. 2021b; Ouali, Hudelot, and Tami 2020). However, these methods are still based on the single-modal approaches, leading to poor pseudo labels. In this paper, we introducing VLM into SSMIS to enhancing the quality of pseudo labels by generating text prompt guided multimodal supervision information.

Vision-Language Model

Although existing VLMs learns generic visual-textual representations by aligning image-text, they are limited by uncertainty awareness of cross-modal and intra-modal. CLIP (Radford et al. 2021) is a representative work of VLM, using the contrastive loss to calculate similarity scores between images and texts. ViLT (Kim, Son, and Kim 2021) is a more efficient architecture that deals with visual feature using interaction layers. Accordingly, numerous works use text information to improve the image segmentation capabilities (Yang et al. 2022; Ding et al. 2022; Xu et al. 2022). For instance, Yang et al. (Yang et al. 2022) conducted early fusion of image-text features in intermediate layers of a transformer network, achieving significantly cross-modal alignment. Ding et al. (Ding et al. 2022) built an encoder decoder attention network with transformer and multi-head attention to provide the language expression "queries" for the given image. Inspired by VLM in natural images, a few works

have started utilizing text information for medical image analysis (Müller et al. 2022; Tomar et al. 2022). Li et al. (Li et al. 2024b) proposed a Language meets Vision Transformer model (LViT) to incorporated text annotations with images in down-sampling and up-sampling processes, compensating for the quality deficiencies in image data. Boecking et al. (Boecking et al. 2022) used the attention weights learned during local alignment to conduct medical semantic segmentation. In this work, we introduce cross-modal and intra-modal self-supervision with uncertainty awareness for better image-text alignment.

Methodology

Our DuSSS (Fig. 3) proposes uncertainty-aware VLM for SSMIS. Specifically, it comprises two steps. **Step 1: VLM pre-training with DuSSS**. The DuSSS promotes cross-modal semantic associations via DCL and supervises semantic similarity based on uncertainty levels during contrastive learning process for uncertainty learning. It addresses uncertainty problems across modalities. **Step 2: Text-guided SSMIS**. It generates multimodal supervision information (i.e., text-guided mask) to improve pseudo-label quality.

SSS for Uncertain Correspondence Learning

The SSS supervises semantic similarity based on the corresponding uncertainty levels to perceive uncertain correspondences. It comprehends intricate semantic relationships between data pairs that are similar yet semantically ambiguous, improving semantic consistency representations.

The distribution distance is regarded as the uncertainty level, which supervises semantic similarity measured by Euclidean distance. For a paired data (x_1, x_2), their semantic similarity is defined as $D_s(x_1, x_2) = \|s_1 - s_2\|_2$. ($s_1$, s_2) are the semantic embeddings of (x_1, x_2). To measure uncertainty level, semantic embeddings are transformed to multivariate Gaussian distributions. The mean vector μ and variance vector σ^2 denote the center position and the distribution scope, respectively. The 2-Wasserstein (Kantorovich 1960; Kim, Son, and Kim 2021) is utilized to measure the difference (i.e. uncertainty level) of multivariate Gaussian distributions. The following defines the 2-Wasserstein of two Gaussian distributions $\mathcal{N}(\mu_1, \sigma_1)$ and $\mathcal{N}(\mu_2, \sigma_2)$.

$$D_{2W} = ||\mu_1 - \mu_2||_2^2 + ||\sigma_1 - \sigma_2||_2^2 \tag{1}$$

where μ represents the mean vector and σ is the standard deviation vector. The global information of [CLS] is utilized to model the uncertainty distribution. The uncertainty level between the paired data (x_1, x_2) is given by:

$$D_u(x_1, x_2) = a \cdot D_{2W}(x_{1[\texttt{CLS}]}, x_{2[\texttt{CLS}]}) + b \tag{2}$$

where a is a positive scale that controls the degree of uncertainty and b is a deviation value. When comparing two data, both the semantic similarity $D_s(\cdot)$ and the uncertainty level $D_u(\cdot)$ need to be considered. The semantic similarity is constrained when large uncertainty exists. Here, the ratio of the uncertainty level to the semantic similarity is used to define the relative uncertainty $\widehat{D}_u$. The SSS is defined as:

$$\widehat{D}_u(x_1, x_2) = \frac{D_u(x_1, x_2)}{D_s(x_1, x_2)} \tag{3}$$

$$D_{SSS}(\mathbf{x}_1, \mathbf{x}_2) = e^{-\lambda \widehat{D}_u(\mathbf{x}_1, \mathbf{x}_2)} \tag{4}$$

where λ is a parameter for controlling constraint degree.

Summarized Advantages: The SSS provides a novel approach to handle uncertainty for robust image-text alignment. It supervises semantic similarity via the uncertainty levels measured by distribution representations. Therefore, it promotes the ability of similarity judgments for ambiguous sample pairs. It addresses the problem in existing distribution-based VLM methods where original semantic attributes are losing.

VLM Pre-training with DuSSS

The DuSSS learns multiple semantic correspondences across modalities and within each modality for uncertainty learning. Specifically, the DCL is conducted to advance the semantic correlation between images and texts by driving cross- and intra-modal representation learning. The SSS is integrated into cross-modal and intra-modal contrastive learning in DCL to supervise semantic similarity in each contrastive learning process via uncertainty levels, solving semantic uncertainty.

1) SSS-based Cross-Modal Contrastive Learning (CMC) aims to learn multiple semantic correspondences between image and text embeddings. That is, pull the matched image-text embeddings together and push the unmatched image-text embeddings away. To avoid potential uncertainties, the cosine similarity $sim(I, T)$ between image and text is constrained via $D_{SSS}(I, T)$, thus obtaining the uncertainty cosine similarity $\widehat{sim}(I, T)$. The InfoNCE loss (van den Oord, Li, and Vinyals 2018) is used to optimize $\widehat{sim}(I, T)$, thus enabling the positive image-text pairs similar and the negative pairs dissimilar. For N data pairs in a batch, N positive image-text pairs are matched and there are $N(N-1)$ negative image-text pairs. The InfoNCE loss for the SSS-based image-to-text is denoted as:

$$\widehat{sim}(I, T) = 1 - (1 - sim(I, T)) \cdot D_{SSS}(I, T) \tag{5}$$

$$\mathcal{L}_{nce}^{I2T} = -\mathbb{E}_{(I,T)} \left[log \frac{exp(\widehat{sim}((I_1, T_+)/\tau)}{\sum_{n=1}^{N} exp(\widehat{sim}(I_1, \widehat{T}_n)/\tau)} \right] \tag{6}$$

where τ is a learned temperature hyper-parameter. T_+ denotes the positive text matched with I_1 and $\widehat{T} = \{\widehat{T}_1, \ldots, \widehat{T}_N\}$ are the negative texts that do not match I_1. Similarly, the InfoNCE loss for text-to-image is given by:

$$\mathcal{L}_{nce}^{T2I} = -\mathbb{E}_{(T,I)} \left[log \frac{exp(\widehat{sim}(T_1, I_+)/\tau)}{\sum_{n=1}^{N} exp(\widehat{sim}(T_1, \widehat{I}_n)/\tau)} \right] \tag{7}$$

where I_+ is the positive image that matches to T_1 and $\widehat{I} = \{\widehat{I}_1, \ldots, \widehat{I}_N\}$ are the negative images that do not match T_1. Overall, we define the cross-modal loss of image-text as:

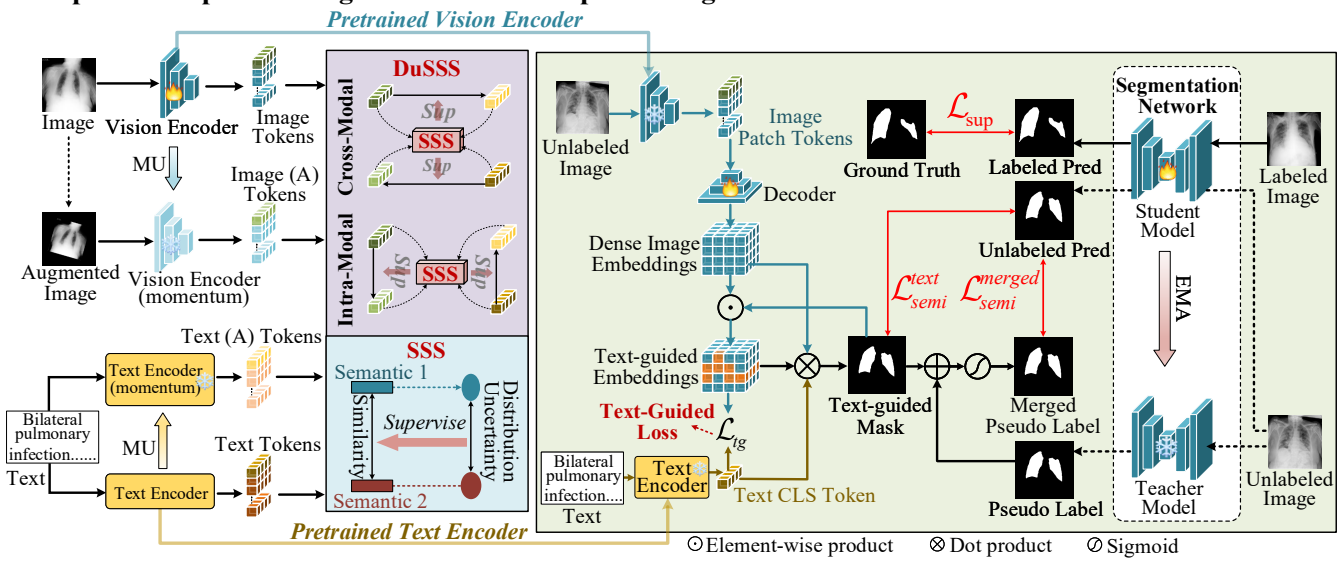

Figure 3: The framework of our DuSSS driven VLM for SSMIS. **Step 1:** Our DuSSS improves the ability of uncertainty understanding in VLM pre-training, thus enhancing the model's robustness for image-text alignment. **Step 2:** The text-guided SSMIS improves the quality of pseudo-labels for reliable semi-supervised consistency learning.

$$\mathcal{L}_{cmc} = \frac{1}{2}\left[\mathcal{L}_{nce}^{I2T}(I_1, T_+, \widehat{T}) + \mathcal{L}_{nce}^{T2I}(T_1, I_+, \widehat{I})\right] \quad (8)$$

The CMC aligns images and texts through minimizing $\mathcal{L}_{cmc}$. However, cross-modal alignment fails to build semantic associations within each modality. To solve this, SSS-based intra-modal contrastive learning is introduced.

2) SSS-based Intra-Modal Contrastive Learning (IMC) captures the underlying associations among different samples within each modality. Meanwhile, the SSS is injected into IMC for weakening semantic uncertainty within each modality. For the raw image-text (I_1, T_1) and the randomly augmented image-text (I_2, T_2), similar to cross-modal loss, intra-modal contrastive loss is defined as:

$$\mathcal{L}_{imc} = \frac{1}{2}\left[\mathcal{L}_{nce}^{I2I}(I_1, I_2, \widehat{I}) + \mathcal{L}_{nce}^{T2T}(T_1, T_2, \widehat{T})\right] \quad (9)$$

where I_2 is the image to be matched with I_1 and $\widehat{I} = \{\widehat{I_1}, \ldots, \widehat{I_N}\}$ are the negative images. Similarly, T_2 is the text to be matched with T_1 and $\widehat{T} = \{\widehat{T_1}, \ldots, \widehat{T_N}\}$ are the negative texts.

Summarized Advantages: The DuSSS provides a new scheme to tackle uncertainty problems across modalities. It incorporates SSS into each contrastive learning process in DCL to supervise semantic similarity via the corresponding uncertainty levels, addressing cross-modal uncertainty.

Text-Guided SSMIS for Pseudo-Label Quality Enhancement

The text-guided SSMIS enhances the quality of pseudo-labels via the benefit of text descriptions. The goal of SSMIS is to train a model by utilizing a labeled training dataset $X_l = \{(x_{li}, y_{li})\}_i^{N_l}$ and an unlabeled training dataset $X_u = \{(x_{uj})\}_i^{N_u}$. Our text-guided SSMIS consists of two paths: 1) Text-guided pseudo-label generation, and 2) Teacher-Student Network-guided pseudo-label generation. These two paths are detailed as follows.

1) Text-Guided Pseudo-Label Generation. The text-guided pseudo-label leverages the advantage of text to localize target segmentation regions. The patch-level image features v_f^{patch} and text features t_f from the pre-trained VLM are used to synthetize the text-guided mask. Specifically, the patch-level image features v_f are decoded into pixel-level features through a grounding decoder $f_g(\cdot)$. The text-guided mask y_u^{text} is calculated by dot product between v_f and t_f, and the process is as follows.

$$v_f = f_g(v_f^{patch}), \quad y_u^{text} = \sigma(t_f^\top v_f) \quad (10)$$

where σ is Sigmoid function, and $\top$ is transposition.

Text-Guided Loss. To suppress the irrelevant text guided mask generation, a text-guided loss $\mathcal{L}_{tg}$ is designed. It helps the model better understand both the global structure and local details. The text-guided loss is denoted as:

$$\mathcal{L}_{tg}(v_f^t, t_f) = -\frac{1}{2}\mathbb{E}_{(f,t)}\left[log\frac{exp(sim(v_f^t, t_f))/\tau)}{\sum_{k=1}^{K} exp(sim(v_f^t, \widehat{t}_{f_k})/\tau)}\right] - \frac{1}{2}\mathbb{E}_{(t,f)}\left[log\frac{exp(sim(t_f, v_f^t))/\tau)}{\sum_{k=1}^{K} exp(sim(t_f, \widehat{v}_{f_k}^t)/\tau)}\right] \quad (11)$$

2) Teacher Student Network-Guided Pseudo-Label Generation. The teacher and student segmentation net-

works have identical architecture. The student network f_{θ_s} is parameterized by θ_s, while the teacher network f_{θ_t} is updated by the Exponential Moving Average (EMA) of the student. The updating process is as follows:

$$\theta_t = \alpha\theta_t + (1-\alpha)\theta_s \tag{12}$$

where $\alpha \in$[0,1] controls the updating pace. The student network predicts the labeled image x_l and the unlabeled image x_u. The teacher network generates the pseudo-label y_u^t of the unlabeled image. The calculation process is:

$$y_l = f_{\theta_s}(x_l), \quad y_u^s = f_{\theta_s}(x_u), \quad y_u^t = f_{\theta_t}(x_u) \tag{13}$$

where y_l and y_u^s are labeled and unlabeled prediction.

The supervised loss $\mathcal{L}_{sup}$ and semi-supervised loss $\mathcal{L}_{semi}$ are utilized to jointly optimize the segmentation network.

For labeled images, $\mathcal{L}_{sup}$ is computed between the ground truth y_{gt} and the labeled prediction y_l. For unlabeled images x_u, $\mathcal{L}_{semi}$ is obtained by $\mathcal{L}_{semi}^{merged}$ and $\mathcal{L}_{semi}^{text}$. $\mathcal{L}_{semi}^{merged}$ is calculated between the unlabeled prediction y_u^s and the merged pseudo-label, where the merged pseudo-label is the combination of both the teacher network-generated pseudo-label and the text-guided mask. $\mathcal{L}_{semi}^{text}$ is calculated between the unlabeled prediction y_u^s and the text-guided mask.

$$\mathcal{L}_{sup} = -\frac{1}{N_l}\frac{1}{HW}\sum_{i=1}^{N_l}\sum_{j=1}^{HW}\ell_{ce}(y_{l_{i,j}}, y_{gt_{i,j}}) \tag{14}$$

$$\mathcal{L}_{semi}^{merged} = -\frac{1}{N_u}\frac{1}{HW}\sum_{i=1}^{N_u}\sum_{j=1}^{HW}\ell_{ce}(y_{u_{i,j}}^s, \sigma(y_{u_{i,j}}^t + y_{u_{i,j}}^{text}) \tag{15}$$

$$\mathcal{L}_{semi}^{text} = -\frac{1}{N_u}\frac{1}{HW}\sum_{i=1}^{N_u}\sum_{j=1}^{HW}\ell_{ce}(y_{u_{i,j}}^s, y_{u_{i,j}}^{text}) \tag{16}$$

$$\mathcal{L}_{semi} = (\mathcal{L}_{semi}^{merged} + \mathcal{L}_{semi}^{text})/2 \tag{17}$$

where $\ell(\cdot)$ is the cross-entropy loss. (i,j) represents the j-th pixel in i-th mask. N_l and N_u are the batch size of labeled images and unlabeled images. W and H represent the width and height of an image.

Summarized Advantages: To the best of our knowledge, this is the first attempt to integrate text-guided mask into SSMIS. It compensates quality deficiencies of pseudo-labels, achieving reliable consistency regularization in SSMIS.

Theoretical Analysis

Our theoretical proofs have proven that SSS can address the alignment uncertainty between image and text.

Proof.

$$\begin{aligned}\widehat{sim}(I,T) &= 1-(1-sim(I,T))\cdot D_{SSS}(I,T)\\ &= 1-(1-sim(I,T))\cdot e^{-\lambda\frac{D_u(x_1,x_2)}{D_s(x_1,x_2)}}\\ &\propto D_u(x_1,x_2).\end{aligned} \tag{18}$$

For the image-text pair with high uncertainty $D_u(x_1,x_2)$, i.e. high relative uncertainty $\frac{D_u(x_1,x_2)}{D_s(x_1,x_2)}$, the proposed uncertainty cosine similarity make such image-text pair more similar, addressing the alignment uncertainty.

Experiments

Datasets

Extensive experiments are conducted on three public medical image segmentation datasets: chest infection area segmentation, bone metastases segmentation, and nuclei instances segmentation.

QaTa-COV19 (Degerli et al. 2022) contains 9258 COVID-19 chest X-ray radiographs. (Li et al. 2024b) provided text annotations and split 7145 samples for training and 2113 samples for testing.

BM-Seg (Afnouch et al. 2023) consists of 23 CT-scans from 23 patients, totaling 1517 slices from different skeletal views. 270 slices from a single view are selected, allocating 200 for training and 70 for testing.

MoNuSeg (Kumar et al. 2017) includes 44 images, and the image size is 1000×1000. The training dataset contains 30 images and the testing dataset contains 14 images.

Implementation Details

Our method is implemented using Pytorch. The operating system is Ubuntu 20.04.4 LTS with 24GB V100 GPU. The learning rate is set to 3e-4 for both the QaTa-COV19 and BM-Seg datasets, and 1e-3 for the MoNuSeg dataset. Early stopping is implemented if the model's performance does not improve after 20 epochs based on its current performance. The batch size is 32 for the QaTa-COV19 dataset, 16 for the BM-Seg dataset, and 4 for the MoNuSeg dataset.

Evaluation Metrics

The Dice $Dice = \sum_{i=1}^{N}\sum_{j=1}^{C}\frac{1}{NC}\cdot\frac{2|p_{ij}\cap y_{ij}|}{(|p_{ij}|+|y_{ij}|)}$ and mIoU $mIoU = \sum_{i=1}^{N}\sum_{j=1}^{C}\frac{1}{NC}\cdot\frac{|p_{ij}\cap y_{ij}|}{|p_{ij}\cup y_{ij}|}$ are used to evaluate our method and other compared methods, where C is the number of categories, and N is the number of pixels.

Comparison Study Shows Our Superiority

The DuSSS is compared with 13 state-of-the art methods, including U-Net (Ronneberger, Fischer, and Brox 2015b), CLIP (Radford et al. 2021), ViLT (Kim, Son, and Kim 2021), MT (Tarvainen and Valpola 2017), CCT (Ouali, Hudelot, and Tami 2020), BCP (Bai et al. 2023), MC-Net (Wu et al. 2022a), SS-Net (Wu et al. 2022b), UCMT (Shen et al. 2023), and LViT (Li et al. 2024b), CMITM (Chen et al. 2023), ASG (Li et al. 2024a), MGCA (Wang et al. 2022a).

Qualitative Analysis. Fig. 4 presents the excellent segmentation performance of our method in comparison with state-of-the art methods. It can be observed that our method not only accurately localizes target regions but also generates coherent boundaries, even in small object circumstances. As shown in Fig. 4, SS-Net, MC-Net, BCP, MT and CCT all have more severe mis-segmentation than the proposed method. This indicates that the introduction of the

Method	Data used		QaTa-COV19		BM-Seg		MoNuSeg		Complexity	
	Labeled	Text	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)	Param (M)	Flops (G)
U-Net	100%	×	79.02	69.46	74.15	60.32	78.66	68.46	14.75	25.19
CLIP	100%	✓	79.81	70.66	74.49	59.56	79.79	68.27	87.00	57.60
ViLT	100%	✓	79.63	70.12	73.23	58.22	77.92	67.81	87.40	28.00
MT	25%	×	77.88	67.53	66.31	53.35	72.80	56.70	14.75	25.19
CCT		×	78.02	67.03	70.66	55.27	74.25	56.55	4.65	8.45
BCP		×	74.79	65.26	71.26	55.28	72.06	54.69	1.81	2.28
MC-Net		×	74.58	64.26	69.67	53.36	71.53	48.12	1.81	2.28
SS-Net		×	67.93	58.13	70.21	54.68	71.80	47.80	1.81	2.28
UCMT		×	76.09	64.13	71.22	55.82	75.53	54.82	1.81	2.30
LViT		✓	78.12	66.75	69.45	54.26	75.69	56.14	29.72	27.08
CMITM		✓	78.04	65.84	70.63	54.66	75.14	54.28	14.75	25.19
ASG		✓	77.92	65.09	70.26	54.33	75.69	55.92	14.75	25.19
MGCA		✓	78.17	67.03	70.19	53.67	76.14	56.22	14.75	25.19
Ours		✓	**79.00**	**68.21**	**71.48**	**55.97**	**76.51**	**58.08**	14.75	25.19
MT	50%	×	79.64	71.88	72.31	56.99	75.15	59.64	–	–
CCT		×	80.25	71.69	72.65	60.54	76.23	55.87	–	–
BCP		×	75.57	65.31	74.21	61.25	72.17	50.88	–	–
MC-Net		×	74.98	61.97	73.37	60.74	71.59	48.75	–	–
SS-Net		×	68.37	56.73	73.59	60.89	73.05	49.88	–	–
UCMT		×	77.81	68.65	74.32	60.41	76.36	59.29	–	–
LViT		✓	80.32	72.16	73.15	60.14	76.57	65.44	–	–
CMITM		✓	81.33	72.84	74.96	60.18	77.63	66.31	–	–
ASG		✓	80.47	71.84	74.22	59.47	76.79	65.91	–	–
MGCA		✓	81.24	73.56	**75.17**	**61.49**	77.06	65.23	–	–
Ours		✓	**82.52**	**75.87**	74.61	61.08	**78.03**	**66.93**	–	–

Table 1: The comparative experiments on the **QaTa-COV19**, **BM-Seg** and **MoNuSeg** datasets demonstrate that our powerful uncertainty processing and pseudo-label enhancing ability.

Method			QaTa-COV19		BM-Seg		MoNuSeg	
SSS	DCL	$\mathcal{L}_{tg}$	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)
×	×	×	79.64	71.88	72.31	56.99	75.15	59.64
✓	×	×	80.86	72.33	74.34	60.22	77.25	63.81
×	✓	×	80.13	72.31	73.11	60.14	76.90	62.73
✓	×	✓	81.49	75.25	75.58	60.79	77.53	65.10
×	✓	✓	81.05	73.80	74.27	60.65	77.10	64.01
✓	✓	×	81.10	75.08	75.24	59.11	77.74	64.62
✓	✓	✓	**82.52**	**75.87**	**76.41**	**61.08**	**78.03**	**66.93**

Table 2: Ablation studies demonstrate that significant improvements of the proposed innovations. The results are based on 50% labeled data on the QaTa-COV19, BM-Seg and MoNuSeg datasets. Baseline: Teacher-Student Network; SSS: Semantic Similarity Supervision; DCL: Dual Contrastive Learning; $\mathcal{L}_{tg}$: Text-Guided Loss.

text-guided mask in the semi-supervised learning can better guide the training of the model, and consequently generate more accurate segmentation. In addition, compared with other VLM methods, the DuSSS is more delicate in the segmentation boundary. This is attributed to the understanding of cross-modal alignment uncertainty.

Quantitative Analysis. Our DuSSS achieves highly competitive performance on three datasets against existing state-of-the art methods. Specifically, Table 1 indicates that our DuSSS outperforms all the other methods in both Dice (82.52%) and mIoU (75.87%) on 50% labeled images, and both Dice (79.00%) and mIoU (68.21%) on 25% labeled images from the QaTa-COV19 dataset. It is noteworthy that the proposed method surpasses fully supervised methods. Additionally, the proposed method achieves better performance compared to the text-equipped methods, while requiring fewer parameters and having lower computational costs. It indicates that our DuSSS driven SSMIS has the ability of precisely locating target regions with text guidance. Table 1 also demonstrates the superiority our DuSSS in the BM-Seg and MoNuSeg datasets. Overall, our DuSSS achieves the best results in different segmentation tasks, demonstrat-

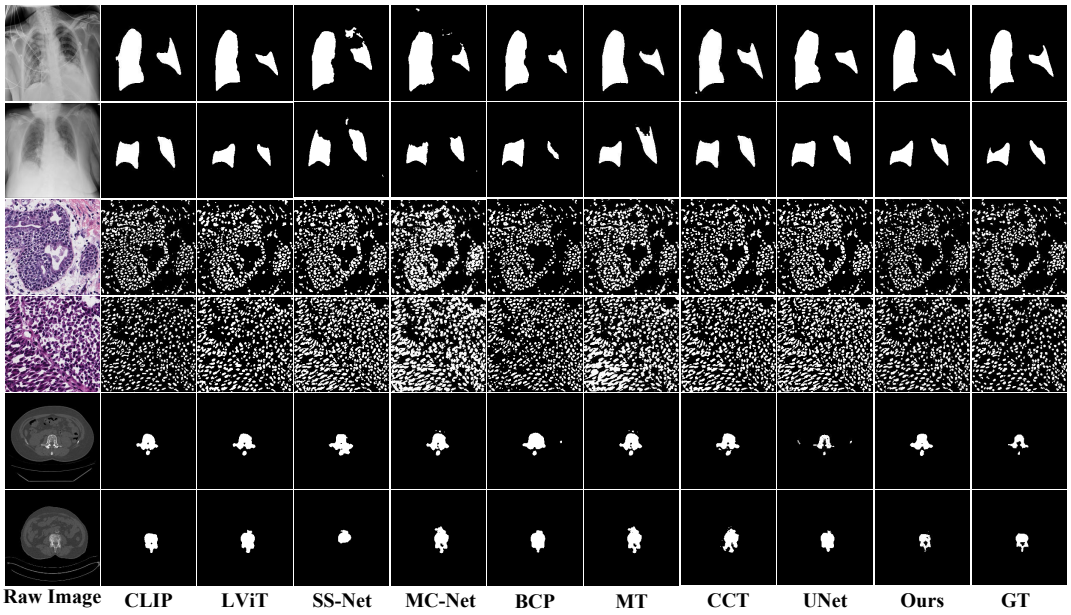

Figure 4: The visual superiority of the proposed method (DuSSS) on the QaTa-COV19, BM-Seg and MoNuSeg datasets. The proposed method shows high-quality segmentation, compared with semi-supervised and VLM-based methods.

ing its superior generalization and robustness.

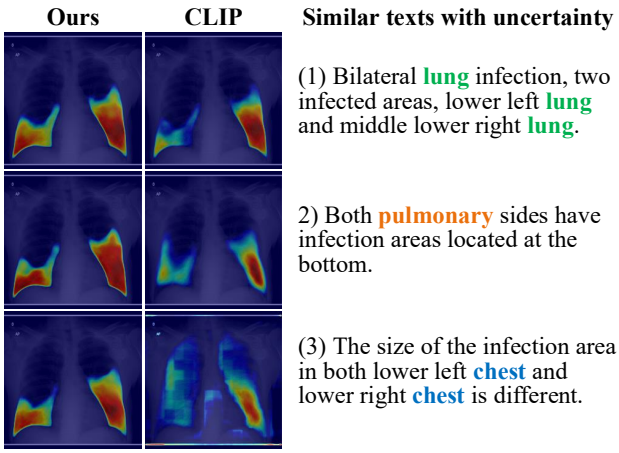

Figure 5: Our DuSSS shows great target region activation effects for multiple similarly yet lexically ambiguous texts.

Ablation Study

We perform ablation experiments to verify the contribution of novel designs, and the number of segmentation model parameters does not increase after adding novel designs.

Effectiveness of SSS. Table 2 shows that removing the SSS limits the performance, resulting in a decrease of 1.22% in Dice and 0.45% in mIoU on the QaTa-COV19 dataset. Additionally, adding SSS to DCL could further improve 0.97% in Dice and 2.77% mIoU. This indicates that the SSS contributes to supervising semantic similarity in ambiguous data via being aware of the uncertainty.

Effectiveness of DCL. Table 2 demonstrates that the DCL improves the Dice and mIoU by 0.49% and 0.43% on the QaTa-COV19 dataset respectively, compared with the cross-modal contrastive learning alone. This owes to alleviating uncertainty during image-text alignment by enhancing semantic consistent representations within each modality.

Effectiveness of Text-Guided Loss. Table 2 indicates that introducing the text-guided loss improves the performance of SSS, DCL, and their combination. It demonstrates the superiority of text-guided loss in restraining the interference of negative text regions.

Uncertainty Analysis

Fig. 5 demonstrates the strong robustness of our DuSSS against alignment uncertainty. For multiple similar texts with lexical uncertainty, our DuSSS generates robust target region activation effects. However, other VLM-based method shows unstable activation effects. Therefore, the proposed DuSSS can comprehend various semantic correspondences between image and text, aiding the model to better address alignment uncertainty between image and text.

Conclusion

In this paper, we propose a novel VLM named DuSSS for SSMIS. Our DuSSS integrates the SSS into each contrastive learning process to supervise semantic similarity based on uncertainty levels, addressing semantic uncertainty across modalities. Finally, using the pretrained VLM, a text-guide SSMIS framework is proposed to enhance the quality of pseudo labels, improving the model's consistency learning capability. Experimental results demonstrate that our DuSSS outperforms SOTA methods.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (Grant No.62173212), Taishan Scholars Program of Shandong Province(Grant No.tsqn202306017), Shandong Province"Double-Hundred Talent Plan"on 100 Foreign Experts and 100 Foreign Expert Teams (Grant No.WSR2023049).

References

Afnouch, M.; Gaddour, O.; Hentati, Y.; Bougourzi, F.; Abid, M.; Alouani, I.; and Ahmed, A. T. 2023. BM-Seg: A new bone metastases segmentation dataset and ensemble of CNN-based segmentation approach. *Expert Systems with Applications*, 228: 120376.

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 (CVPR)*, 11514–11524.

Boecking, B.; Usuyama, N.; Bannur, S.; Castro, D. C.; Schwaighofer, A.; Hyland, S.; Wetscherek, M.; Naumann, T.; Nori, A.; Alvarez-Valle, J.; Poon, H.; and Oktay, O. 2022. Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 1–21.

Chen, C.; Zhong, A.; Wu, D.; Luo, J.; and Li, Q. 2023. Contrastive masked image-text modeling for medical visual representation learning. *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 493–503.

Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021a. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv preprint arXiv:2102.04306*.

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

Chun, S.; Oh, S. J.; de Rezende, R. S.; Kalantidis, Y.; and Larlus, D. 2021. Probabilistic Embeddings for Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8411–8420.

Degerli, A.; Kiranyaz, S.; Chowdhury, M. E. H.; and Gabbouj, M. 2022. Osegnet: Operational Segmentation Network for Covid-19 Detection Using Chest X-Ray Images. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2306–2310.

Ding, H.; Liu, C.; Wang, S.; and Jiang, X. 2022. VLT: Vision-Language Transformer and Query Generation for Referring Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 1–16.

Fang, Z.; Chen, Y.; Nie, D.; Lin, W.; and Shen, D. 2019. RCA-U-Net: Residual Channel Attention U-Net for Fast Tissue Quantification in Magnetic Resonance Fingerprinting. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 101–109.

Hung, W.-C.; Tsai, Y.-H.; Liou, Y.-T.; Lin, Y.-Y.; and Yang, M.-H. 2018. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*.

Kantorovich, L. V. 1960. Mathematical Methods of Organizing and Planning Production. *Management Science*, 6: 366–422.

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 5583–5594.

Kumar, N.; Verma, R.; Sharma, S.; Bhargava, S.; Vahadane, A.; and Sethi, A. 2017. A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology. *IEEE Transactions on Medical Imaging*, 36(7): 1550–1560.

Lei, T.; Sun, R.; Wang, X.; Wang, Y.; He, X.; and Asoke, N. 2023. CiT-Net: Convolutional Neural Networks Hand in Hand with Vision Transformers for Medical Image Segmentation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 1017–1025.

Li, Q.; Yan, X.; Xu, J.; Yuan, R.; Zhang, Y.; Feng, R.; Shen, Q.; Zhang, X.; and Wang, S. 2024a. Anatomical structure-guided medical vision-language pre-training. *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 80–90.

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

Li, Z.; Li, Y.; Li, Q.; Wang, P.; Guo, D.; Lu, L.; Jin, D.; Zhang, Y.; and Hong, Q. 2024b. LViT: Language Meets Vision Transformer in Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 43(1): 96–107.

Müller, P.; Kaissis, G.; Zou, C.; and Rueckert, D. 2022. Joint learning of localized representations from medical images and reports. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 685–701.

Ouali, Y.; Hudelot, C.; and Tami, M. 2020. Semi-supervised semantic segmentation with cross-consistency training. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12674–12684.

Qin, C.; Wang, Y.; and Zhang, J. 2024. URCA: Uncertainty-based region clipping algorithm for semi-supervised medical image segmentation. *Computer Methods and Programs in Biomedicine*, 108278.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 8748–8763.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015a. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the International Conference on*

Medical Image Computing and Computer Assisted Intervention (MICCAI), 234–241.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015b. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 234–241.

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 International Joint Conference on Artificial Intelligence (IJCAI)*, 4199–4207.

Sun, J. J.; Zhao, J.; Chen, L.-C.; Schroff, F.; Adam, H.; and Liu, T. 2020. Probabilistic Embeddings for Cross-Modal Retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 53–70.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, volume 30.

Tomar, N. K.; Jha, D.; Bagci, U.; and Ali, S. 2022. TGANet: Text-guided attention for improved polyp segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 151–160.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748*.

Wang, F.; Zhou, Y.; Wang, S.; Vardhanabhuti, V.; and Yu, L. 2022a. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, 35: 33536–33549.

Wang, P.; Peng, J.; Pedersoli, M.; Zhou, Y.; Zhang, C.; and Desrosiers, C. 2021. Self-paced and self-consistent co-training for semi-supervised image segmentation. *Medical Image Analysis*, 73: 102146.

Wang, Y.; Zhang, Y.; Tian, J.; Zhong, C.; Shi, Z.; Zhang, Y.; and He, Z. 2020. Double-uncertainty weighted method for semi-supervised learning. *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 542–551.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022b. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3876–3887.

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

Wu, Y.; Wu, Z.; Wu, Q.; Ge, Z.; and Cai, J. 2022b. Exploring Smoothness and Class-Separation for Semi-supervised Medical Image Segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 34–43.

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

Xu, J.; De Mello, S.; Liu, S.; Byeon, W.; Breuel, T.; Kautz, J.; and Wang, X. 2022. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18134–18144.

Yang, G.; Zhang, J.; Zhang, Y.; Wu, B.; and Yang, Y. 2021. Probabilistic modeling of semantic ambiguity for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12522–12531.

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2022. LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18134–18144.

Yi, M.; Cui, Q.; Wu, H.; Yang, C.; Yoshie, O.; and Lu, H. 2023. A simple framework for text-supervised semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7071–7080.

Yu, T.; Li, D.; Yang, Y.; Hospedales, T. M.; and Xiang, T. 2019. Robust Person Re-Identification by Modelling Feature Uncertainty. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 552–561.

Zhang, S.; Zhang, J.; Tian, B.; Lukasiewicz, T.; and Xu, Z. 2023. Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 83: 102656.

Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; and Liang, J. 2018. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 3–11.

Zhu, Y.; Zhang, Z.; Wu, C.; Zhang, Z.; He, T.; Zhang, H.; Manmatha, R.; Li, M.; and Smola, A. 2021. Improving semantic segmentation via efficient self-training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3): 1589–1602.

Zou, Y.; Yu, Z.; Kumar, B.; and Wang, J. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 289–305.

Özgün Çiçek; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 424–432.