VCLIPSeg: Voxel-wise CLIP-Enhanced model for Semi-Supervised Medical Image Segmentation

Lei Li[1[0000−0002−1500−7446]], Sheng Lian[3,4], Zhiming Luo[2(✉)], Beizhan Wang[1], and Shaozi Li[2]

[1] The Department of Software Engineering, Xiamen University, Fujian, China
[2] The Department of Artificial Intelligence, Xiamen University, Fujian, China
[3] The College of Computer and Data Science, Fuzhou University, Fujian, China
[4] The Fujian Key Laboratory of Network Computing and Intelligent Information Processing (Fuzhou University), Fujian, China
✉Correspondences: zhiming.luo@xmu.edu.cn

Abstract. Semi-supervised learning has emerged as a critical approach for addressing medical image segmentation with limited annotation, and pseudo labeling-based methods made significant progress for this task. However, the varying quality of pseudo labels poses a challenge to model generalization. In this paper, we propose a **V**oxel-wise **CLIP**-enhanced model for semi-supervised medical image **Seg**mentation (VCLIPSeg). Our model incorporates three modules: Voxel-Wise Prompts Module (VWPM), Vision-Text Consistency Module (VTCM), and Dynamic Labeling Branch (DLB). The VWPM integrates CLIP embeddings in a voxel-wise manner, learning the semantic relationships among pixels. The VTCM constrains the image prototype features, reducing the impact of noisy data. The DLB adaptively generates pseudo-labels, effectively leveraging the unlabeled data. Experimental results on the Left Atrial (LA) dataset and Pancreas-CT dataset demonstrate the superiority of our method over state-of-the-art approaches in terms of the Dice score. For instance, it achieves a Dice score of 88.51% using only 5% labeled data from the LA dataset.

Keywords: Organ segmentation · CLIP · Semi-supervised learning.

1 Introduction

Medical image segmentation plays an essential role in disease diagnosis, treatment planning, and disease detection [17, 25]. The development of an accurate and robust segmentation model necessitates a significant amount of annotated data. However, the annotation for medical images is time-consuming and demands expert knowledge, especially for 3D volumes [22, 7, 11]. Consequently, training effective medical image segmentation models with limited annotations has become an important problem.

In recent years, semi-supervised learning (SSL) has emerged as a critical approach to tackle this challenge [1, 18, 9]. SSL aims to utilize both labeled and

unlabeled data for model training, alleviating the burden of annotation. Moreover, methods based on pseudo labeling [20, 19, 4] have made significant progress in this task. However, the varying quality of pseudo labels can greatly affect the generalization of deep learning models.

On the other aspects, Radford et al. [15] proposed the Contrastive Language-Image Pretraining (CLIP), which has achieved remarkable success. CLIP can map text and images into a shared semantic space, facilitating cross-modal understanding. Trained with a large amount of text and images from the Internet, CLIP has made generalized applications across diverse domains feasible [6, 5, 12]. Despite its proven potential in computer vision, CLIP's adoption in medical image analysis, particularly in medical image segmentation, remains relatively rare. The CLIP-Driven Universal model [10] represents a pioneering effort to integrate CLIP into voxel-level semantic understanding tasks in the medical domain. By capturing anatomical relationships, this framework empowers models to tackle partially labeled problems. However, it combines CLIP embeddings with image features in a global manner, disregarding spatial relationships among pixels, and relies on a wide range of organ datasets, limiting its ability to handle semi-supervised medical segmentation tasks. Thus, it raises a challenge: **How can we effectively utilize CLIP for 3D single-organ SSL tasks?**

In this paper, we propose a voxel-wise CLIP-enhanced model for semi-supervised medical image segmentation (VCLIPSeg), aiming to explore how to enhance semi-supervised image segmentation through CLIP. Our method designs a Voxel-Wise Prompts Module (VWPM) and a Vision-Text Consistency Module (VTCM). The VWPM integrates CLIP embeddings into image features at the voxel level, learning the semantic relationships among different pixels. The VWPM enhances image features, making the prediction results of the semi-supervised model more accurate. The VTCM calculates the similarity between the image prototype features and CLIP embeddings to constrain different categories of image features, reducing the impact of noisy data. Moreover, due to the noisy and inaccurate predictions for unlabeled training images, employing hard labels [2] or sharpened pseudo-labels [20] boosts the model's confidence in these incorrect predictions, potentially leading to overfitting the noise. We develop a dynamic labeling branch that adaptively generates pseudo-labels in real time, thereby enhancing the leverage of unlabeled data. We evaluate the proposed VCLIPSeg in two public datasets: the Left Atrial Segmentation Challenge (LA) and Pancreas-CT datasets. Extensive experiments show our approach achieves competitive results. Specifically, with 5% labeled data on the LA dataset, our approach achieves an 88.51% Dice score, 0.71% higher than the SOTA.

2 Methodology

2.1 Problem Setting

In the semi-supervised segmentation setting, the whole dataset consists of N labeled data and M unlabeled data, where $N \ll M$. Let $D_L = \{(x_i^l, y_i^l)\}_{i=1}^N$ and $D_U = \{(x_i^u)\}_{j=1}^M$ denote the labeled set and unlabeled set, respectively. The x_i^l

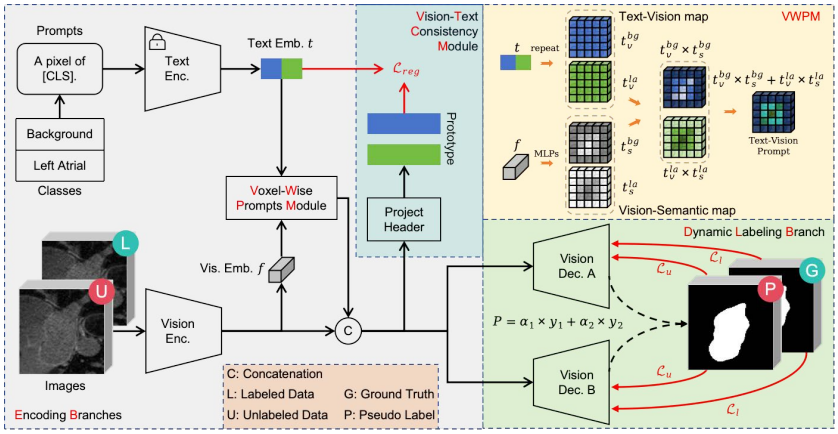

Fig. 1. Overview of our proposed VCLIPSeg. The VCLIPSeg comprises a vision encoder, a text encoder, and two vision decoders. The VWPM and the VTCM use CLIP embeddings to enhance model features. The DLB utilizes unlabeled data by adaptively generating pseudo labels in real-time.

and x_i^u correspond to the medical volumes, while the y_i^l represents the manual annotated Ground Truth (GT). The primary objective is to jointly utilize D_L and D_U to learn a segmentation model that can achieve higher segmentation accuracy. We propose a SSL framework named VCLIPSeg, shown in Fig. 1.

2.2 Encoding Branches

As illustrated in Fig. 1, the Encoding Branches consist of a text branch and a vision branch.

Text Branch: In this branch, we utilize the pre-trained language model CLIP as the Text Encoder while keeping its parameters frozen. It is used solely for extracting text embeddings once before training. To better learn pixel-level semantic relationships, we employ the medical prompt "A pixel of [CLS]" where "[CLS]" corresponds to the class names in the segmentation results. Let $t \in \mathbb{R}^{C\times 512}$ denote the text embeddings, where C represents the number of categories in the task. Combining the knowledge of visual and language modalities improves the performance and generalization ability of the model.

Vision Branch: In this branch, we adopt the V-Net encoder [14] as the Vision Encoder. Let $f \in \mathbb{R}^{D\times W\times H}$ denote the Vision Embeddings extracted from the encoder, where D, H, W corresponding to the depth, width, and height of the embeddings, respectively.

2.3 Voxel-Wise Prompts Module (VWPM)

Due to the large distribution difference between text features and image features, we design VWPM to generate text-vision prompts based on both image and text.

This module aims to explore the semantic relationships between different pixels in medical volume data for obtaining more meaningful segmentation results.

As shown in the top-right in Fig. 1, we first repeat the CLIP embeddings t to match the size of the vision embedding to obtain text-vision map t_v. For the vision embedding f, we input it into the vision-semantic controller, comprising two multi-layer perceptrons (MLPs). This controller aims to learn the vision semantic map (v_s) for the pixels in the volume data.

$$v_s = \text{softmax}\left(\text{controller}(f)\right), \tag{1}$$

Subsequently, we obtain the voxel-level text-vision prompt by multiplying and summing the text-vision map and the vision-semantic map.

$$prompt = t_v^{bg} \times t_s^{bg} + t_v^{la} \times t_s^{la}. \tag{2}$$

Finally, we concatenate f and $prompt$ as the decoder input.

$$f_p = \text{conv}\left(\text{cat}(f, prompt)\right), \tag{3}$$

where cat is a concatenation operation and conv denotes a convolution operation.

2.4 Vision-Text Consistency Module (VTCM)

Considering the large amount of unlabeled data in the training dataset, *prompt* learned through VWPM may introduce noise, potentially hindering the feature discrimination based on CLIP embeddings t. We further propose the VTCM to regularize f_p to ensure the feature separation between different categories.

After the VWPM, we further add a project header $proj$ that maps f_p into l_2-normalized 512-d feature vectors. Subsequently, we use the prototype of each class to perform the regularization of image features. In this step, we only consider regions with the highest confidence to compute the prototypes. Specifically, we assign a mask value of 1 to regions where both Decoder A and Decoder B in our model accurately predict the outcome. Then, we compute the prototype features by averaging the selected features across categories based on the mask.

$$f_{proto} = \text{AVG-POOL}(proj(f_p), mask). \tag{4}$$

Following, we compute the cosine similarity between prototype features and text embeddings to impose constraints on the features by maximizing their similarity. This regularization loss is defined as:

$$\mathcal{L}_{reg} = 1 - cos(f_{proto}, t), \tag{5}$$

where cos computes the cosine similarity. Note that the VTCM is exclusively applied during the training phase, which does not impose extra computational overhead during deployment.

2.5 Dynamic Labeling Branches (DLB)

Since predictions for unlabeled training datasets can be noisy and inaccurate, employing hard [2] or sharpened pseudo-labels [20] might boost confidence in incorrect predictions, ultimately raising the risk of overfitting to the noise. To tackle this issue, we design a dynamic labeling branch, based on the simple assumption that the decoder with a lower loss performs better. This module dynamically generates pseudo labels based on the loss value, enabling the effective utilization of unlabeled data. For the labeled data, we simply use the dice loss.

$$\mathcal{L}_l = \mathcal{L}_{dice}(predA, GT) + \mathcal{L}_{dice}(predB, GT), \tag{6}$$

where $\mathcal{L}_{dice}$ denotes the dice loss function. For unlabeled data, we calculate weights based on the loss of different decoders and dynamically synthesize pseudo-labels in real time,

$$w_1 = \frac{e^{1-\sigma(\mathcal{L}_{dice}(predA,GT))}}{e^{1-\sigma(\mathcal{L}_{dice}(predA,GT))} + e^{1-\sigma(\mathcal{L}_{dice}(predB,GT))}}, \tag{7}$$

$$w_2 = \frac{e^{1-\sigma(\mathcal{L}_{dice}(predB,GT))}}{e^{1-\sigma(\mathcal{L}_{dice}(predA,GT))} + e^{1-\sigma(\mathcal{L}_{dice}(predB,GT))}}, \tag{8}$$

$$P = w_1 * predA + w_2 * predB, \tag{9}$$

where σ denotes the sigmoid function, and P is the weighted pseudo labels.

Next, the generated pseudo-labels P are utilized to provide supervision during the training of unlabeled data. The unlabeled data is trained by both the dice loss and cross-entropy loss, which are represented in the following equations:

$$\mathcal{L}_u = \mathcal{L}_{dice}(predA, P) + \mathcal{L}_{ce}(predA, P) + \mathcal{L}_{dice}(predB, P) + \mathcal{L}_{ce}(predB, P), \tag{10}$$

where $\mathcal{L}_{ce}$ denotes the cross-entropy loss function. Finally, we combine the segmentation loss and the regularization loss to train our framework as the following:

$$\mathcal{L} = \mathcal{L}_l + \lambda_u \mathcal{L}_u + \lambda_r \ \mathcal{L}_{reg} \tag{11}$$

3 Experiments

3.1 Experimental Setup

Dataset. In this study, we conduct experiments on two public datasets: the Left Atrium Segmentation Challenge (LA) [21] dataset and the Pancreas-CT [3] dataset. For LA dataset, the volumes maintain a consistent isotropic resolution at $0.625^3\ mm^3$. The dataset is split into 80 volumes for training and 20 volumes for testing. For Pancreas-CT dataset, the image intensity is windowed by [-125, 275]. Following [19], 62 volumes are allocated for training and 20 for testing.

Implementation details. To ensure a fair comparison, we conduct all evaluations using Pytorch 2.0.1 and CUDA 11.8 on an NVIDIA GeForce RTX 3090

Table 1. Comparison with other methods on the LA dataset.

Method	Volumes used		Metrics			
	Labeled	Unlabeled	Dice(%)↑	Jaccard (%)↑	95 HD (%)↓	ASD (voxel)↓
V-Net [14]	4 (5%)	0	50.13	39.32	32.04	10.27
V-Net [14]	8 (10%)	0	83.33	72.86	19.03	5.71
V-Net [14]	16 (20%)	0	89.40	81.01	7.22	2.05
V-Net [14]	80 (100%)	0	92.27	85.69	4.33	1.45
MT (NeurIPS'17) [16]	4 (5%)	76 (95%)	80.83	68.54	18.34	5.62
UA-MT (MICCAI'19) [24]			79.49	66.59	24.04	6.89
MC-Net (MICCAI'21) [20]			82.48	70.79	18.71	4.49
MC-Net+ (MIA'22) [19]			83.31	72.24	14.61	3.46
CAML (MICCAI'23) [4]			87.42	77.74	9.57	2.23
†ARCO (NeurIPS'23) [23]			87.80	-	-	-
VCLIPSeg (Ours)			88.51	79.49	10.48	2.28
MT (NeurIPS'17) [16]	8 (10%)	72 (90%)	87.45	77.88	18.60	4.96
UA-MT (MICCAI'19) [24]			85.71	75.45	15.14	3.92
MC-Net (MICCAI'21) [20]			87.37	77.80	10.07	1.91
MC-Net+ (MIA'22) [19]			88.63	79.72	9.01	1.78
CAML (MICCAI'23) [4]			89.33	80.81	8.05	2.11
†ARCO (NeurIPS'23) [23]			89.90	-	-	-
VCLIPSeg (Ours)			90.59	82.87	6.22	1.61
MT (NeurIPS'17) [16]	16 (20%)	64 (80%)	90.23	82.29	6.62	2.05
UA-MT (MICCAI'19) [24]			90.19	82.26	7.64	2.14
MC-Net (MICCAI'21) [20]			91.10	83.73	5.94	1.57
MC-Net+ (MIA'22) [19]			90.91	83.40	6.58	1.57
CAML (MICCAI'23) [4]			90.70	83.06	7.90	1.94
VCLIPSeg (Ours)			91.15	83.82	5.86	1.49

† represents data from referenced paper.

GPU. We randomly extract 3D patches of size $112 \times 112 \times 80$ from the LA dataset and $96 \times 96 \times 96$ from the Pancreas-CT dataset following [19]. We use the SGD optimizer with a learning rate of 10^{-2} and train for 15,000 iterations. Each training batch comprised four patches, two of which are unlabeled. To balance the loss items, we introduce a time-dependent Gaussian warming-up function [8] for λ_u, defined as $\lambda_u(t) = e^{-5(1-t/t_{max})^2}$. Additionally, we set λ_r to 0.5. To evaluate the performance, four evaluation metrics are utilized: The Dice coefficient, the Jaccard coefficient, the 95% Hausdorff Distance (95 HD), and the average surface distance (ASD).

3.2 Comparsion with Sate-of-the-Art Methods

We conduct experiments on the LA and Pancreas-CT datasets to evaluate our method, comparing it with several state-of-the-art methods [16, 24, 20, 19, 4, 23][5]. We first compare with different methods at three annotation levels and the results are presented in Fig. 1. In summary, our method demonstrates the best overall performance at all annotation levels. Especially, with just 5% labeled data, our approach surpasses ARCO by 0.71% in terms of the Dice score. With only 10% labeled data, we achieve a Dice score of 90.59%. When the amount of

[5] The code of ARCO [23] is not open-source, we only report the results in their paper.

Table 2. Comparison with other methods on the Pancreas-CT dataset.

Method	Volumes used		Metrics			
	Labeled	Unlabeled	Dice (%)↑	Jaccard (%)↑	95 HD (%)↓	ASD (voxel)↓
V-Net [14]	12 (20%)	0	75.07	61.96	10.79	3.31
V-Net [14]	62 (100%)	0	83.48	71.98	4.38	1.25
MT (NeurIPS'17) [16]	12 (20%)	50 (80%)	77.33	64.20	9.70	3.12
UA-MT (MICCAI'19) [24]			77.36	64.25	8.40	2.89
MC-Net (MICCAI'21) [20]			79.22	66.54	7.46	1.57
MC-Net+ (MIA'22) [19]			78.19	65.30	8.59	2.80
CAML (MICCAI'23) [4]			78.28	65.88	7.97	1.16
CauSSL (ICCV'23) [13]			77.02	63.85	7.69	1.62
VCLIPSeg (Ours)			80.56	68.13	6.75	1.49

annotated data increases to 20%, the performance of our model is only slightly lower than the V-Net trained with 100% labeled data, reaching 91.15% compared to the upper bound of 92.27%. Furthermore, we conduct additional experiments on the more challenging Pancreas-CT dataset. As ARCO [23] is not open-source, we include a new approach called CauSSL [13] for comparison, which incorporates a causal diagram into the SSL model. Table 2 presents the comparative results of all methods using 20% labeled data. These results indicate that our method effectively enhances semi-supervised medical image segmentation by incorporating CLIP information.

Qualitative Analysis. We further conduct a qualitative comparison among these methods. Examples from the LA dataset with only 5% labeled data are presented in Fig. 2. The 1st row displays a 3D view, while the 2nd row shows the results of a 2D slice. To ensure a fair comparison, no post-processing methods are applied to any of the methods. In the 1st row, our approach, which integrates semantic information from CLIP, produces smoother prediction results. The 2nd row shows a slice where fully supervised methods only predict a small portion. Other semi-supervised techniques, with the assistance of unlabeled data, notably improve their predictions but still exhibit discernible omissions, particularly in

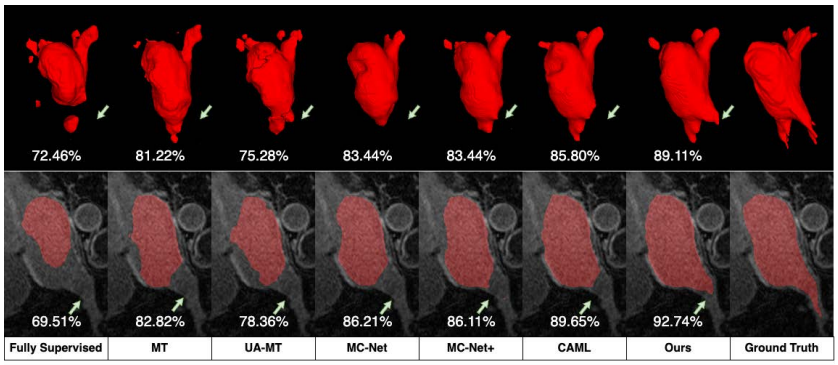

Fig. 2. Qualitative analysis of the proposed method on the LA dataset.

the lower right corner. In contrast, our approach learns the semantic relationships among pixels, resulting in significant improvements.

3.3 Ablation studies

VWPM and VTCM: In this section, we analyze the effectiveness of the proposed VWPM and VTCM, by employing the MC-Net [20] with DLB as the baseline. We evaluate the performance at three annotation levels from the LA dataset and report the dice scores in Table 3. When utilizing 5% labeled data, VWPM and VTCM individually increase the Dice scores to 87.33% and 87.41%, respectively. Furthermore, the combined use of VWPM and VTCM leads to a further increase in dice scores, reaching 88.51%. Similar enhancements are observed when using 10% and 20 % labeled data.

Table 3. Ablation study of our proposed VWPM and VTCM on the LA dataset.

VWPM	VTCM	5%	10%	20%
-	-	86.93	89.39	90.99
✓	-	87.33	89.76	90.84
-	✓	87.41	89.55	91.01
✓	✓	88.51	90.59	91.15

DLB: To validate the effectiveness of the DLB, we add the DLB into the MC-Net [20] and CPS [2] for evaluation and the results are reported in Table 4. As shown in Table 4, both MC-Net and CPS effectively leverage unlabeled data, increasing Dice scores from 50.13% to 82.48% and 84.67%, respectively. When introducing our DLB, MC-Net's Dice score improved by 4.45%. Similarly, CPS increased from 84.67% to 85.52%. This indicates that DLB is more effective in generating pseudo-labels.

Table 4. Ablation study of our proposed DLB on the LA dataset.

Method	Dice (%)↑	Jaccard (%)↑	95 HD (%)↓	ASD (voxel)↓
V-Net	50.13	39.32	32.04	10.27
MC-Net	82.48	70.79	18.71	4.49
+DLB	86.93 (↑ 4.45)	77.10 (↑ 6.31)	9.22 (↓ 9.49)	2.19 (↓ 2.30)
CPS	84.67	73.77	16.60	5.04
+DLB	85.52 (↑ 0.85)	75.11 (↑ 1.34)	12.14 (↓ 4.46)	2.67 (↓ 2.37)

4 Conclusion

In this paper, we introduce a voxel-wise CLIP-Enhanced model for semi-supervised medical image segmentation. To improve the SSL model, we develop VWPM,

which integrates CLIP embeddings at the voxel level. VTCM is used to constrain image features, thereby reducing the influence of noisy labels. Additionally, we propose a dynamic labeling branch to better utilize unlabeled data. Through extensive experimentation, we show the effectiveness of our approach in transferring the capabilities of CLIP and outperforming the state-of-the-art (SOTA) methods at different annotation levels.

Acknowledgments. This work is supported by National Key R&D Program of China (No. 2023YFC3503004); the National Natural Science Foundation of China (No. 62276221, No. 62376232).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Basak, H., Yin, Z.: Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. pp. 19786–19797 (2023)
2. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. pp. 2613–2622 (2021)
3. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al.: The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of Digital Imaging **26**, 1045–1057 (2013)
4. Gao, S., Zhang, Z., Ma, J., Li, Z., Zhang, S.: Correlation-aware mutual learning for semi-supervised medical image segmentation. In: MICCAI. pp. 98–108 (2023)
5. Guo, Z., Dong, B., Ji, Z., Bai, J., Guo, Y., Zuo, W.: Texts as images in prompt tuning for multi-label image recognition. pp. 2808–2817 (2023)
6. He, W., Jamonnak, S., Gou, L., Ren, L.: Clip-s4: Language-guided self-supervised semantic segmentation. pp. 11207–11216 (2023)
7. Hu, Q., Chen, Y., Xiao, J., Sun, S., Chen, J., Yuille, A.L., Zhou, Z.: Label-free liver tumor segmentation. pp. 7422–7432 (2023)
8. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
9. Lei, T., Zhang, D., Du, X., Wang, X., Wan, Y., Nandi, A.K.: Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network. IEEE Transactions on Medical Imaging (2022)
10. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. pp. 21152–21164 (2023)
11. Luo, X., Hu, M., Liao, W., Zhai, S., Song, T., Wang, G., Zhang, S.: Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision. In: MICCAI. pp. 528–538 (2022)
12. Lyu, Y., Lin, T., Li, F., He, D., Dong, J., Tan, T.: Deltaedit: Exploring text-free training for text-driven image manipulation. pp. 6894–6903 (2023)
13. Miao, J., Chen, C., Liu, F., Wei, H., Heng, P.A.: Caussl: Causality-inspired semi-supervised learning for medical image segmentation. pp. 21426–21437 (2023)

14. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision (3DV). pp. 565–571 (2016)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. pp. 8748–8763 (2021)
16. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results **30** (2017)
17. Wang, L., Wang, J., Zhu, L., Fu, H., Li, P., Cheng, G., Feng, Z., Li, S., Heng, P.A.: Dual multiscale mean teacher network for semi-supervised infection segmentation in chest ct volume for covid-19. IEEE Transactions on Cybernetics (2022)
18. Wang, P., Peng, J., Pedersoli, M., Zhou, Y., Zhang, C., Desrosiers, C.: Cat: Constrained adversarial training for anatomically-plausible semi-supervised segmentation. IEEE Transactions on Medical Imaging (2023)
19. Wu, Y., Ge, Z., Zhang, D., Xu, M., Zhang, L., Xia, Y., Cai, J.: Mutual consistency learning for semi-supervised medical image segmentation. Medical Image Analysis **81**, 102530 (2022)
20. Wu, Y., Xu, M., Ge, Z., Cai, J., Zhang, L.: Semi-supervised left atrium segmentation with mutual consistency training. In: MICCAI. pp. 297–306 (2021)
21. Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., Vesal, S., Ravikumar, N., Maier, A., Yang, X., et al.: A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Medical image analysis **67**, 101832 (2021)
22. Ye, Y., Xie, Y., Zhang, J., Chen, Z., Xia, Y.: Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner. arXiv preprint arXiv:2304.03493 (2023)
23. You, C., Dai, W., Min, Y., Liu, F., Zhang, X., Feng, C., Clifton, D.A., Zhou, S.K., Staib, L.H., Duncan, J.S.: Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. arXiv preprint arXiv:2302.01735 (2023)
24. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: MICCAI. pp. 605–613 (2019)
25. Zeng, L.L., Gao, K., Hu, D., Feng, Z., Hou, C., Rong, P., Wang, W.: Ss-tbn: A semi-supervised tri-branch network for covid-19 screening and lesion segmentation (2023)