Text-driven Multiplanar Visual Interaction for Semi-supervised Medical Image Segmentation

Kaiwen Huang[1], Yi Zhou[2], Huazhu Fu[3], Yizhe Zhang[1], Chen Gong[1], Tao Zhou[1] (✉)

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, China.
taozhou.ai@gmail.com
[2] School of Computer Science and Engineering, Southeast University, China.
[3] Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore.

Abstract. Semi-supervised medical image segmentation plays a critical method in mitigating the high cost of data annotation. When labeled data is limited, textual information can provide additional context to enhance visual semantic understanding. However, research exploring the use of textual data to enhance visual semantic embeddings in 3D medical imaging tasks remains scarce. In this paper, we propose a novel text-driven multiplanar visual interaction framework for semi-supervised medical image segmentation (termed Text-SemiSeg), which consists of three main modules: Text-enhanced Multiplanar Representation (TMR), Category-aware Semantic Alignment (CSA), and Dynamic Cognitive Augmentation (DCA). Specifically, TMR facilitates text-visual interaction through planar mapping, thereby enhancing the category awareness of visual features. CSA performs cross-modal semantic alignment between the text features with introduced learnable variables and the intermediate layer of visual features. DCA reduces the distribution discrepancy between labeled and unlabeled data through their interaction, thus improving the model's robustness. Finally, experiments on three public datasets demonstrate that our model effectively enhances visual features with textual information and outperforms other methods. Our code is available at https://github.com/taozh2017/Text-SemiSeg.

Keywords: Semi-supervised medical segmentation, TMR, CSA, DCA

1 Introduction

Semi-supervised medical image segmentation [7,10,12], which blends limited labeled data with extensive unlabeled data, has garnered increasing attention for alleviating manual annotation burdens. Existing semi-supervised methods are primarily divided into consistency learning [4,5] and pseudo-labeling approaches [8,29]. Consistency learning enhances the model's generalization ability by ensuring consistent predictions across various perturbed inputs. For example, Huang *et al.* [11] employed cutout and slice misalignment as input perturbations

for consistency learning. Additionally, Xu *et al.* [26] designed different network architectures to introduce perturbations at the feature level. Pseudo-labeling generates and iteratively refines pseudo-labels from model predictions on unlabeled data. Wang *et al.* [22] introduced a confidence block to assess the quality of pseudo-labels and used a thresholding to select high-confidence pseudo-labels. Han *et al.* [8] generated high-quality pseudo-labels by obtaining class prototypes from labeled data and calculating the feature distance similarity between unlabeled data and the class prototypes. However, these methods often exhibit a phenomenon where they achieve satisfactory results on a specific medical modality or organ but lack generalizability to other tasks.

Recent work based on Visual-Language Models (VLMs), such as Contrastive Language-Image Pre-training (CLIP) [20], has achieved remarkable results by aligning images with their corresponding text descriptions through contrastive learning. In the medical imaging domain, Liu *et al.* [15] pioneered the application of CLIP to 3D medical imaging tasks, demonstrating that textual information can provide additional semantic context to supplement voxel-based visual features. Additionally, VCLIPSeg [14] applies CLIP to semi-supervised 3D medical segmentation. However, VCLIPSeg simply replicates the textual information to match the dimensionality of the 3D data, which does not align with the 2D-oriented training paradigm of CLIP. Moreover, CLIP lacks learnable parameters to adapt to this cross-dimensional operation from 2D to 3D. **Thus, how can a semi-supervised text-visual model be designed to better accommodate 3D medical image segmentation?**

To this end, we propose a novel **Text-SemiSeg** framework, which integrates textual information into semi-supervised 3D medical image segmentation. Specifically, we propose a Text-enhanced Multiplanar Representation (TMR) strategy to fully leverage text-visual alignment capabilities from CLIP as it is pre-trained on a large number of 2D images, thereby enhancing the semantic representation of visual features. Additionally, we present a Category-aware Semantic Alignment (CSA) module, which aligns visual embeddings with text features to augment the model's category semantic awareness. To further mitigate the well-known distributional discrepancy between labeled and unlabeled data in semi-supervised tasks, we propose the Dynamic Cognitive Augmentation (DCA) module. This approach reduces these distributional disparities, resulting in improved model robustness. **Our contributions are highlighted as follows: i)** We propose a novel framework that is more suitable for applying CLIP in 3D semi-supervised medical segmentation scenarios. More importantly, this framework can also be applied to other semi-supervised networks. **ii)** We propose TMR and CSA to fully enhance visual features with textual cues. **iii)** We present the DCA strategy to effectively reduce the distribution gap between labeled and unlabeled data. **iv)** Extensive experiments on three 3D medical datasets demonstrate that our model outperforms other state-of-the-art semi-supervised methods.

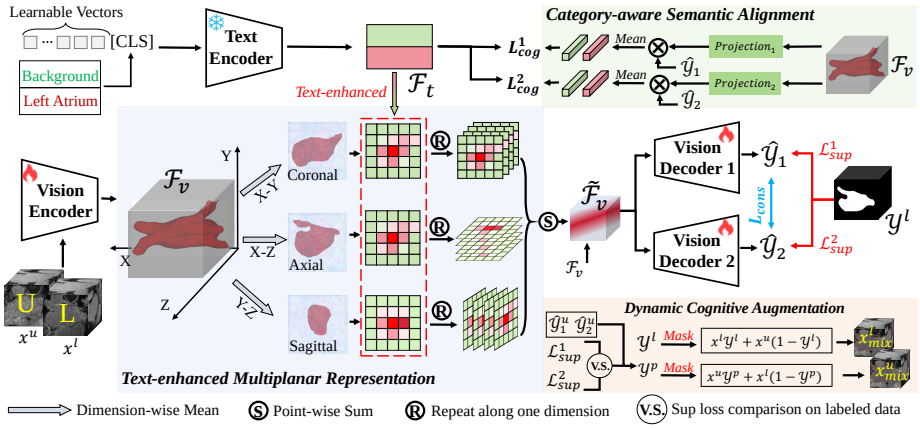

Fig. 1. Overview of our Text-SemiSeg framework, including an MC-Net [25] backbone and a text encoder. The TMR and CSA leverage textual information to enhance visual feature learning, while the DCA adaptively generates mixed samples from both labeled and unlabeled data to reduce the distribution gap between them.

2 Proposed Method

Overview. In the semi-supervised setting, the dataset comprises a labeled subset $\mathcal{D}^L = \{(x_i^l, y_i^l)\}_{i=1}^{N_l}$ and an unlabeled subset $\mathcal{D}^U = \{x_i^u\}_{i=N_l+1}^{N_l+N_u}$, where $N_l \ll N_u$. $x_i \in \mathbb{R}^{H\times W\times D}$ represents the input volume, and $y_i \in \{0,1\}^{H\times W\times D}$ is the corresponding ground-truth map. Fig. 1 illustrates the architecture of our Text-SemiSeg framework. Specifically, the input volume $\{x_i^l, x_i^u\}$ is fed into the encoder to obtain feature embeddings. Meanwhile, the category text is incorporated in a learnable manner [30], yielding text features $\mathcal{F}_t \in \mathbb{R}^{K\times C}$, where K and C represent the number of categories and channels, respectively. In the TMR module, to align with the CLIP training paradigm, we extract feature embeddings from distinct planes, generating three 2D features. We then enhance the visual features of each perspective by incorporating textual information. Subsequently, the enhanced feature embeddings are fed into two decoders for consistency learning. Then, we leverage the CSA module to facilitate the alignment of textual and visual features in their representations. Finally, we incorporate the DCA to reduce the distributional discrepancies between labeled and unlabeled data.

2.1 Text-enhanced Multiplanar Representation

CLIP is a text-visual model trained through contrastive learning between 2D images and text. However, the transformation of pixel-level semantics into voxel-level semantic information is still challenging. To address this issue, we employ two strategies: (1) incorporating learnable variables into the text to enhance the generalizability of text features, and (2) performing semantic enhancement between text features and various planes of 3D feature embeddings. Specifically,

the text description in CLIP no longer uses the format `'A photo of [CLS]'`. Instead, we employ a prompt learning approach [30] as follows:

$$\mathcal{F}_t = [V_1][V_2]\dots[V_M][CLS], \tag{1}$$

where $[V]$ and $[CLS]$ represent learnable variables and class embeddings, respectively. The input volume is fed into the vision encoder to obtain visual embeddings $\mathcal{F}_v$. Subsequently, the visual embeddings are compressed into 2D images via adaptive average pooling along the width, height, and depth dimensions, respectively. This process yields feature embeddings for the coronal plane $\mathcal{F}_v^c$, sagittal plane $\mathcal{F}_v^s$, and axial plane $\mathcal{F}_v^a$. These planes then interact with the text features through a text-enhanced attention operation. For the coronal plane, self-attention is initially applied. Then, text enhancement is performed on the plane, which can be expressed as:

$$\tilde{\mathcal{F}}_v^c = \sigma\left(\frac{\mathbf{Q}(\mathcal{F}_t)\cdot\mathbf{K}(\mathcal{A}(\mathcal{F}_v^c))^\top}{\sqrt{d}}\right)\cdot\mathbf{V}(\mathcal{A}(\mathcal{F}_v^c)), \tag{2}$$

where $\mathcal{A}(\cdot)$ denotes the attention operation. $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the linear projections. $\sqrt{d}$ is a scaling factor, and $\sigma(\cdot)$ represents the softmax activation function. For the sagittal and axial planes, the operations are consistent with those detailed above. Finally, the voxel embeddings are reconstructed from the three text-enhanced planes, and the process can be mathematically expressed as

$$\tilde{\mathcal{F}}_v(x,y,z) = \sum_{x=1}^{H}\sum_{y=1}^{W}\sum_{z=1}^{D}\Big(w_c\cdot\tilde{\mathcal{F}}_v^c(x,y) + w_s\cdot\tilde{\mathcal{F}}_v^s(y,z) + w_a\cdot\tilde{\mathcal{F}}_v^a(x,z)\Big) + \mathcal{F}_v, \tag{3}$$

where w_c, w_s, and w_a are learnable weights that dynamically regulate the contributions of three different planes in the reconstruction of 3D voxel features.

2.2 Category-aware Semantic Alignment

We leverage the proposed CSA module to enhance the network's contextual understanding of text-image pairs. Specifically, the predictions of the two visual decoders are multiplied by the final layer features of the visual encoder to obtain the corresponding category semantic representations. We apply MSE regularization to constrain each category's text feature to be as close as possible to the visual embedding, which can be expressed as:

$$\mathcal{L}_{cog} = \sum\nolimits_{k=1}^{K}(\mathcal{F}_t^k - \text{avg}(\mathcal{F}_v^p\cdot\hat{y}_{1,k}))^2 + (\mathcal{F}_t^k - \text{avg}(\mathcal{F}_v^p\cdot\hat{y}_{2,k}))^2, \tag{4}$$

where $\mathcal{F}_t^k$ represents the text feature of the k-th category. $\hat{y}_{1,k}$ and $\hat{y}_{2,k}$ represent the predictions of the two visual decoders for the k-th category, respectively. $\text{avg}(\cdot)$ denotes the global average pooling operation. $\mathcal{F}_v^p = proj(\mathcal{F}_v)$ represents the output of the visual feature $\mathcal{F}_v$ passing through the projection head, which ensures that it matches the dimensionality of the text features.

2.3 Dynamic Cognitive Augmentation

In semi-supervised tasks, due to the scarcity of labeled data, there is often a distributional discrepancy between labeled and unlabeled data. Reducing this discrepancy can enhance the model's ability to recognize the foreground and reduce the impact of various types of noise. Thus, we use the GT and pseudo-labels to extract the foreground parts from both labeled and unlabeled data and fuse them into each other's background. This process can be expressed as:

$$\begin{cases} x^l_{mix} = \sum_{k=1}^{K} x^l \cdot y^l_k + x^u \cdot (1 - y^l_k), \quad x^u_{mix} = \sum_{k=1}^{K} x^u \cdot y^p_k + x^l \cdot (1 - y^p_k), \\ y^p = \begin{cases} \mathcal{B}(\hat{y}^u_1), & \mathcal{L}^1_{sup} < \mathcal{L}^2_{sup}, \\ \mathcal{B}(\hat{y}^u_2), & \mathcal{L}^1_{sup} > \mathcal{L}^2_{sup}, \end{cases} \end{cases} \tag{5}$$

where $\mathcal{B}(\cdot)$ represents the binarization operation. y^l_k and y^p_k denote the k-th class ground truth and pseudo-label, respectively. $\hat{y}^u_1$ and $\hat{y}^u_2$ represent the predictions of the two decoders for the unlabeled data, respectively. In the process of generating pseudo-labels, we select the decoder output with the lower loss on the labeled data as the pseudo-label for the unlabeled data. Then, when the mixed data is fed into the model for retraining, loss constraints are applied only to the underperforming decoder, encouraging its improvement. The process for obtaining mixed labels can be expressed as follows:

$$y^l_{mix} = \sum\nolimits_{k=1}^{K} y^l \cdot y^l_k + y^p \cdot (1 - y^l_k), \quad y^u_{mix} = \sum\nolimits_{k=1}^{K} y^p \cdot y^p_k + y^l \cdot (1 - y^p_k). \tag{6}$$

2.4 Overall Loss Function

In thi study, the segmentation loss consists of the Dice and cross-entropy (CE) losses, which can be defined by $\mathcal{L}_{seg}(\hat{y}, y) = \mathcal{L}_{dice}(\hat{y}, y) + \mathcal{L}_{ce}(\hat{y}, y)$, where $\hat{y}$ and y denote the prediction and ground truth, respectively. Thus, we have

$$\mathcal{L}_{sup} = \mathcal{L}_{seg}(\hat{y}^l_1, y^l) + \mathcal{L}_{seg}(\hat{y}^l_2, y^l), \quad \mathcal{L}_{unsup} = \mathcal{L}_{seg}(\hat{y}^u_1, \hat{y}^u_2) + \mathcal{L}_{seg}(\hat{y}^u_2, \hat{y}^u_1). \tag{7}$$

Moreover, the mixed loss $\mathcal{L}_{mix}$, which updates only the decoder with poorer performance in each iteration, can be expressed as follows:

$$\mathcal{L}_{mix} = \mathcal{L}_{seg}(\hat{y}^l_{mix}, y^l_{mix}) + \mathcal{L}_{seg}(\hat{y}^u_{mix}, y^u_{mix}), \tag{8}$$

where $\hat{y}^l_{mix}$ and $\hat{y}^u_{mix}$ represent the model's prediction on x^l_{mix} and x^u_{mix}.

Finally, the overall loss function can be formulated by

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \mathcal{L}_{cog} + \mathcal{L}_{mix} + \lambda_u \mathcal{L}_{unsup}, \tag{9}$$

where λ_u represents Gaussian warm-up function [13].

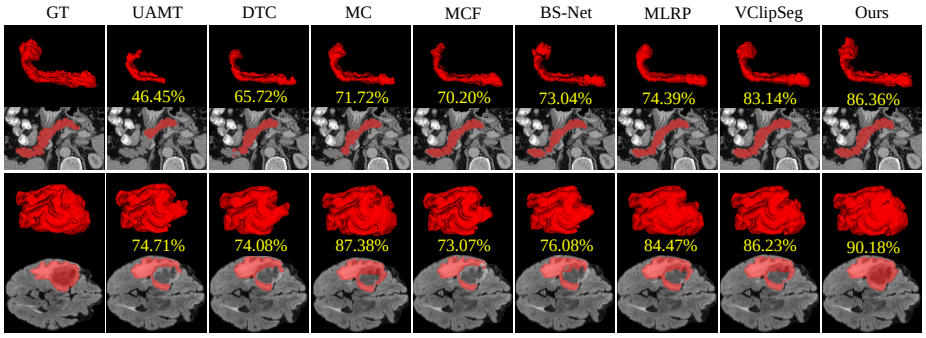

Fig. 2. Visualization results on the Pancreas and BraTS-2019 datasets, with yellow numbers indicating the dice value of the sample.

3 Experiments and Results

3.1 Experimental Setup

Datasets. We conduct comparative experiments on the Pancreas-CT [6], Brats-2019 [3] and MSD-Lung[1] tumor datasets. • The Pancreas-CT dataset includes 82 3D abdominal CT scans with a resolution of 512 × 512 and slice thickness of 1.5-2.5 mm. We use 62 samples for training and 20 for testing, based on the same setting in [17]. • The BraTS-2019 dataset consists of MRI images from 335 glioma patients, with four modalities: T1, T2, T1CE, and FLAIR. Following [16] setup, we use only the FLAIR images for tumor segmentation. The data is split into 250 samples for training, 25 for validation, and 60 for testing. • The MSD-Lung dataset consists of 63 CT scans. Following Setting [28], we allocated 48 scans for training and 15 for testing.

Implementation Details. All experiments are conducted in a PyTorch 1.8.1 environment with CUDA 11.2 and an NVIDIA 4090 GPU. To ensure fairness, the backbone selection continues to follow previous settings, choosing VNet [19]. We utilize the SGD optimizer with an initial learning rate of 0.01. The batch size is set to 4, and the number of iterations is set to 30k. The number of learnable parameters in CLIP is set to 4. We use a Gaussian warm-up function $\lambda_m(t) = \beta * e^{-5(1-t/t_{max})^2}$ in $\mathcal{L}_{unsup}$, where β is set to 0.1. Following the previous settings [21], during training, we randomly extract 3D patches of size $96 \times 96 \times 96$.

3.2 Comparison with State-of-the-arts

We compare the proposed model with ten state-of-the-art semi-supervised medical image segmentation methods, namely UA-MT [27], DTC [17], MC-Net [25], MC-Net+ [24], URPC [18], MCF [23], BCP [2], BS-Net [9], MLRP [21], and VClipSeg [14]. Table 1 presents a quantitative comparison and our method achieves the best performance across three datasets. Specifically, under the condition of 10% labeled data, our model surpasses the second-best method by

Table 1. Quantitative results on the Pancreas, BraTS-2019 and MSD Lung datasets.

Dataset	Method	10% / 6 labeled data				20% / 12 labeled data			
		Dice ↑	Jaccard ↑	95HD ↓	ASD ↓	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓
Pancreas	VNet [19] (SupOnly)	54.94	40.87	47.48	17.43	75.07	61.96	10.79	3.31
	UAMT [27] (MICCAI'19)	66.44	52.02	17.04	3.03	76.10	62.62	10.87	2.43
	DTC [17] (AAAI'21)	66.58	51.79	15.46	4.16	76.27	62.82	8.70	2.20
	MC-Net [25] (MICCAI'21)	69.07	54.36	14.53	2.28	78.17	65.22	6.90	1.55
	MC-Net+ [24] (MIA'22)	70.00	55.66	16.03	3.87	79.37	66.83	8.52	1.72
	URPC [18] (MIA'22)	73.53	59.44	22.57	7.85	80.02	67.30	8.51	1.98
	MCF [23] (CVPR'23)	67.71	53.83	17.17	2.34	75.00	61.27	11.59	3.27
	BCP [2] (CVPR'23)	75.17	60.99	10.49	2.32	82.91	70.97	6.43	2.25
	BS-Net [9] (TMI'24)	64.61	50.02	24.74	5.28	78.93	65.75	8.49	2.20
	MLRP [21] (MIA'24)	75.93	62.12	9.07	1.54	81.53	69.35	6.81	1.33
	VClipSeg [14] (MICCAI'24)	74.77	60.88	11.02	1.82	80.56	68.13	6.75	1.49
	Ours	**81.27**	**68.78**	**5.98**	**1.44**	**83.50**	**71.94**	**4.52**	**1.26**
BraTS-2019	Method	10% / 25 labeled data				20% / 50 labeled data			
	VNet [19] (SupOnly)	74.43	61.86	37.11	2.79	80.16	71.55	22.68	3.43
	UAMT [27] (MICCAI'19)	79.49	69.22	11.93	1.93	79.72	69.46	11.26	1.97
	DTC [17] (AAAI'21)	77.54	66.91	12.16	2.84	82.38	72.15	11.00	2.16
	MC-Net [25] (MICCAI'21)	81.52	71.89	13.26	4.56	83.79	73.69	9.65	1.76
	MC-Net+ [24] (MIA'22)	79.27	68.97	14.89	4.91	83.47	72.89	9.69	1.92
	URPC [18] (MIA'22)	82.59	72.11	13.88	3.72	82.93	72.57	15.93	4.19
	MCF [23] (CVPR'23)	78.83	68.49	12.25	1.92	80.07	69.55	10.65	2.00
	BCP [2] (CVPR'23)	78.64	68.59	12.88	2.81	80.20	69.66	10.52	2.08
	BS-Net [9] (TMI'24)	79.93	69.37	10.88	2.05	82.03	71.87	10.26	1.96
	MLRP [21] (MIA'24)	84.29	74.74	9.57	2.55	85.47	76.32	7.76	2.00
	VClipSeg [14] (MICCAI'24)	83.12	73.61	10.40	1.65	84.05	74.38	9.75	1.35
	Ours	**85.27**	**75.83**	**8.77**	**1.35**	**86.69**	**77.83**	**7.65**	**1.24**
MSD-Lung	Method	20% / 10 labeled data				40% / 20 labeled data			
	VNet [19] (SupOnly)	36.36	25.78	19.06	32.73	57.67	43.24	11.99	4.62
	UAMT [27] (MICCAI'19)	40.22	28.10	16.79	7.02	61.24	46.51	8.97	3.54
	DTC [17] (AAAI'21)	59.64	44.23	9.00	2.89	63.40	48.47	12.85	7.82
	MC-Net [25] (MICCAI'21)	52.23	39.09	13.85	7.05	56.10	43.69	16.87	10.92
	MC-Net+ [24] (MIA'22)	58.72	44.47	10.51	4.15	59.52	47.25	12.12	5.87
	URPC [18] (MIA'22)	55.19	39.80	9.08	2.42	62.20	47.93	8.22	**2.38**
	MCF [23] (CVPR'23)	53.81	39.23	12.41	4.35	62.86	48.62	8.42	3.67
	BCP [2] (CVPR'23)	59.34	44.83	11.12	3.69	65.01	48.70	9.79	3.47
	BS-Net [9] (TMI'24)	47.33	32.78	9.54	2.92	53.30	40.02	11.71	4.33
	MLRP [21] (MIA'24)	62.12	47.23	8.47	2.68	67.63	53.35	12.30	7.56
	VClipSeg [14] (MICCAI'24)	59.39	44.55	8.81	2.68	67.58	53.19	12.38	7.64
	Ours	**65.02**	**48.90**	**6.25**	**1.28**	**68.21**	**54.02**	**8.08**	2.97

5.34%, 0.98%, and 2.90% in Dice coefficient on the Pancreas, BraTS, and Lung datasets, respectively. Fig. 2 illustrates the qualitative comparison with 10% labeled data. It can be observed that the proposed framework, Text-SemiSeg, effectively enhances visual feature representation through textual information, leading to more precise predictions, particularly in the edge regions.

3.3 Ablation Studies

• **Effectiveness of Key Components:** We used the MC-Net [25] structure as a baseline and conducted ablation experiments on the proposed TMR, CSA, and DCA modules, as shown in Table 2. From the results, it can be observed that each component contributes the promising performance.

Table 2. Ablation study on the key components using 20% labeled data.

Settings			Pancreas (12 labeled data)				BraTS-2019 (50 labeled data)			
TMR	CSA	DCA	Dice ↑	IoU ↑	95HD ↓	ASD ↓	Dice ↑	IoU ↑	95HD ↓	ASD ↓
			78.17	65.22	6.90	1.55	83.79	73.69	9.65	1.76
✓			79.46	67.11	8.30	1.20	84.63	75.02	9.41	1.60
	✓		80.26	67.60	8.96	1.26	84.72	75.31	9.39	1.63
		✓	83.41	71.81	4.44	1.19	85.38	76.17	9.05	1.26
✓	✓		82.10	69.96	5.32	**1.14**	85.19	75.79	9.32	1.54
✓		✓	83.47	71.85	4.59	1.20	86.30	77.11	9.10	1.58
	✓	✓	83.28	71.63	4.59	1.29	85.89	76.56	8.77	1.65
✓	✓	✓	**83.50**	**71.94**	**4.52**	1.26	**86.69**	**77.83**	**7.65**	**1.24**

Table 3. Ablation study of the text-vision enhancement strategy.

Settings	Pancreas (12 labeled data)				BraTS-2019 (50 labeled data)			
	Dice ↑	IoU ↑	95HD ↓	ASD ↓	Dice ↑	IoU ↑	95HD ↓	ASD ↓
Repeat	82.35	70.42	7.15	2.03	85.57	76.16	8.50	1.33
Multiplanar	**83.50**	**71.94**	**4.52**	**1.26**	**86.69**	**77.83**	**7.65**	**1.24**

Table 4. Effect of text-enhanced strategy applied to other semi-supervised methods.

Settings	Pancreas (12 labeled data)				BraTS-2019 (50 labeled data)			
	Dice ↑	IoU ↑	95HD ↓	ASD ↓	Dice ↑	IoU ↑	95HD ↓	ASD ↓
UAMT [27]	76.10	62.62	10.87	2.43	79.72	69.46	11.26	1.97
+Text	**77.53**	**64.57**	**10.54**	**1.16**	**84.21**	**74.86**	**9.49**	**1.65**
DTC [17]	76.27	62.82	8.70	2.20	82.38	72.15	11.00	2.16
+Text	**81.04**	**68.70**	**7.74**	**1.27**	**84.97**	**75.48**	**9.14**	**1.66**

• **Effects of Multiplanar Strategy:** To demonstrate the effectiveness of adopting a multiplanar strategy, we compare it to the approach used by VClipSeg, which replicates textual features to match the dimensionality of the visual features. As shown in Table 3, the Dice coefficients improve from 82.35% and 85.57% to 83.50% and 86.69% on the Pancreas and BraTS datasets, respectively. The multiplanar strategy offers several advantages: (1) it is more aligned with CLIP's training paradigm, and (2) it can achieve global text-vision enhancement for 2D planar features using only a small number of parameters.

• **Fairness Discussion:** Currently, the exploration of VLMs in 3D scenarios remains quite limited. While some comparative methods in the experiments do not rely on text, the inference cost of Text-SemiSeg is comparable to other semi-supervised methods. The primary focus of this work is to explore how to better apply VLMs to downstream tasks. Table 4 presents the results of incorporating the text-enhanced strategy into other semi-supervised methods, while excluding other strategies such as DCA. The results clearly demonstrate that our text-enhanced strategy effectively boosts segmentation accuracy across different methods.

4 Conclusion

We have proposed a novel Text-SemiSeg framework for semi-supervised 3D medical image segmentation. Our model fully leverages CLIP's alignment capabilities to enrich visual features with textual cues in a 3D context. Additionally, to enhance the interaction between labeled and unlabeled data, we propose dynamic cognitive augmentation, which significantly boosts the model's generalization ability. Extensive experiments across various segmentation tasks show that our model outperforms existing methods and adapts more effectively to 3D scenarios.

Acknowledgement. This work was in part supported by the National Natural Science Foundation of China (Nos. 62172228, 62476054, and 62201263), and Agency for Science, Technology and Research (A*STAR) Central Research Fund ("Robust and Trustworthy AI system for Multi-modality Healthcare").

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

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature Communications **13**(1), 4128 (2022)
2. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation. In: CVPR. pp. 11514–11524 (2023)
3. Bakas, S.S.: Brats miccai brain tumor dataset (2020). https://doi.org/10.21227/hdtd-5j88, https://dx.doi.org/10.21227/hdtd-5j88
4. Basak, H., Bhattacharya, R., Hussain, R., Chatterjee, A.: An exceedingly simple consistency regularization method for semi-supervised medical image segmentation. In: ISBI. pp. 1–4. IEEE (2022)
5. Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., De Bruijne, M.: Semi-supervised medical image segmentation via learning consistency under transformations. In: MICCAI. pp. 810–818. Springer (2019)
6. 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)
7. Gu, Y., Zhou, T., Zhang, Y., Zhou, Y., He, K., Gong, C., Fu, H.: Dual-scale enhanced and cross-generative consistency learning for semi-supervised medical image segmentation. Pattern Recognition **158**, 110962 (2025)
8. Han, K., Liu, L., Song, Y., Liu, Y., Qiu, C., Tang, Y., Teng, Q., Liu, Z.: An effective semi-supervised approach for liver ct image segmentation. IEEE Journal of Biomedical and Health Informatics **26**(8), 3999–4007 (2022)
9. He, A., Li, T., Yan, J., Wang, K., Fu, H.: Bilateral supervision network for semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging (2023)

10. Huang, K., Zhou, T., Fu, H., Zhang, Y., Zhou, Y., Gong, C., Liang, D.: Learnable prompting sam-induced knowledge distillation for semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging **44**(5), 2295–2306 (2025)
11. Huang, W., Chen, C., Xiong, Z., Zhang, Y., Chen, X., Sun, X., Wu, F.: Semi-supervised neuron segmentation via reinforced consistency learning. IEEE Transactions on Medical Imaging **41**(11), 3016–3028 (2022)
12. Jiao, R., Zhang, Y., Ding, L., Xue, B., Zhang, J., Cai, R., Jin, C.: Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. Computers in Biology and Medicine **169**, 107840 (2024)
13. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
14. Li, L., Lian, S., Luo, Z., Wang, B., Li, S.: Vclipseg: Voxel-wise clip-enhanced model for semi-supervised medical image segmentation. In: MICCAI. pp. 692–701. Springer (2024)
15. 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. In: ICCV. pp. 21152–21164 (2023)
16. Luo, X.: SSL4MIS. https://github.com/HiLab-git/SSL4MIS (2020)
17. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: AAAI. vol. 35, pp. 8801–8809 (2021)
18. Luo, X., Wang, G., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Metaxas, D.N., Zhang, S.: Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. Medical Image Analysis **80**, 102517 (2022)
19. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV. pp. 565–571. IEEE (2016)
20. 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. In: ICML. pp. 8748–8763. PMLR (2021)
21. Su, J., Luo, Z., Lian, S., Lin, D., Li, S.: Mutual learning with reliable pseudo label for semi-supervised medical image segmentation. Medical Image Analysis **94**, 103111 (2024)
22. Wang, X., Yuan, Y., Guo, D., Huang, X., Cui, Y., Xia, M., Wang, Z., Bai, C., Chen, S.: SSA-Net: Spatial self-attention network for covid-19 pneumonia infection segmentation with semi-supervised few-shot learning. Medical Image Analysis **79**, 102459 (2022)
23. Wang, Y., Xiao, B., Bi, X., Li, W., Gao, X.: Mcf: Mutual correction framework for semi-supervised medical image segmentation. In: CVPR. pp. 15651–15660 (2023)
24. 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)
25. Wu, Y., Xu, M., Ge, Z., Cai, J., Zhang, L.: Semi-supervised left atrium segmentation with mutual consistency training. In: MICCAI. pp. 297–306. Springer (2021)
26. Xu, M.C., Zhou, Y.K., Jin, C., Blumberg, S.B., Wilson, F.J., deGroot, M., Alexander, D.C., Oxtoby, N.P., Jacob, J.: Learning morphological feature perturbations for calibrated semi-supervised segmentation. In: MIDL. pp. 1413–1429. PMLR (2022)
27. 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. Springer (2019)

28. Zeng, Q., Lu, Z., Xie, Y., Xia, Y.: Pick: Predict and mask for semi-supervised medical image segmentation. International Journal of Computer Vision pp. 1–16 (2025)
29. Zhang, Z., Tian, C., Bai, H.X., Jiao, Z., Tian, X.: Discriminative error prediction network for semi-supervised colon gland segmentation. Medical Image Analysis **79**, 102458 (2022)
30. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)