

PAVL: Enhancing Feature Prototype Alignment with Vision–Language Model for Semi-Supervised Medical Image Segmentation

Junchang Kuang¹*, Jiale Chen¹*, Siyang Feng¹*, Zaiyi Liu^{2,3}, Huan Lin^{2,3}, Rushi Lan^{4,5}, and Xipeng Pan^{1,5}(✉)

¹ School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China
pxp201@guet.edu.cn

² Department of Radiology, Guangdong Provincial People’s Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, 510080, China

³ Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangdong Provincial People’s Hospital, Guangdong Academy of Medical Sciences, Guangzhou, Guangdong, 510080, China

⁴ International Joint Research Laboratory of Spatio-temporal Information and Intelligent Location Services, Guilin University of Electronic Technology, Guilin 541004, China

⁵ Guangxi Key Laboratory of Image and Graphic Intelligent Processing, Guilin University of Electronic Technology, Guilin 541004, China

Abstract. Semi-supervised medical image segmentation (SSMIS) has garnered significant attention for its ability to reduce radiologists’ manual annotation burden. However, inadequate distinction between different categories in the feature space can lead to difficulty in distinguishing visually similar classes. Vision–language models like CLIP excel at disambiguating such categories through rich semantic representations, but their coarse-grained embeddings struggle to capture fine structural details in complex medical scans. To address these challenges, we propose a CLIP-based SSMIS framework, which comprises three key components: unified representation matching (URM), prototype alignment (PA) and embedding fusion CLIP (EFCLIP). Specifically, URM aligns spatial feature representations between teacher and student networks to mitigate the impact of high-uncertainty voxels and ambiguous boundaries. Building on this, PA selects high-confidence voxels based on voxel-level uncertainty to construct robust class prototypes. Furthermore, EFCLIP integrates CLIP’s text embeddings with voxel-level visual features, fusing global semantic priors with local spatial structure to enhance the discriminative ability for visually similar voxels categorize. Extensive experiments on one public and one in-house medical image datasets demonstrate that PAVL consistently outperforms state-of-the-art SSMIS methods.

* Junchang Kuang, Jiale Chen and Siyang Feng contributed equally to this work.

Keywords: Semi-supervised learning · Vision-language model · Feature alignment · Medical image segmentation.

1 Introduction

Semi-supervised medical image segmentation (SSMIS) aims to mitigate manual annotation burdens by leveraging abundant unlabeled data alongside limited labeled samples for model training. Existing SSMIS methods [2, 15] can be categorized into three groups: consistency regularization, pseudo-labeling and adversarial learning. Consistency regularization leverages consistent predictions to constrain model training [11, 12, 21]. For example, Wu *et al.* [21] proposed a method based on mutual consistency constraints among multiple decoders. Pseudo-labeling leverages model predictions on unlabeled data as supervisory signals to guide training [20, 25]. Bai *et al.* [1] proposed a bidirectional copy-paste strategy to improve pseudo-label quality and segmentation performance in SSMIS. Adversarial learning is a strategy that uses perturbed samples to encourage the model to learn robust features [9, 26]. Lei *et al.* [9] leveraged adversarial consistency learning and dynamic convolutional attention to improve robustness and generalization in SSMIS. However, these approaches often overlook voxel-wise uncertainty, failing to capture subtle features in critical regions. Moreover, insufficient discriminability between different classes within the feature space may lead to poorly delineated boundaries. Vision-language models like contrastive language-image pre-training (CLIP) [17] have demonstrated significant effectiveness in resolving inter-class ambiguity. Li *et al.* [10] proposed VCLIPSeg to leverage voxel-level CLIP semantics and dynamic pseudo-labeling for semi-supervised segmentation with limited annotations. However, existing CLIP-based SSMIS methods [10, 19] fail to fully exploit the fine-grained spatial and structural features inherent in unlabeled data.

To address these limitations, we propose a novel SSMIS framework based on feature prototype alignment with vision-language (PAVL). Specifically, we introduce a unified representation matching (URM) strategy to enforce spatial feature consistency between the teacher and student models. Building on this, we propose a prototype alignment (PA) module that assigns uncertainty-aware weights to candidate voxel features, effectively suppressing noisy samples. Moreover, a novel prototype learning paradigm is proposed to mitigate the issue of inter-class similarity. This approach enhances inter-class discriminability, thereby obtaining robust and precise boundary information. Furthermore, we propose embedding fusion CLIP (EFCLIP), which integrates CLIP’s text embeddings with the student model’s voxel-level visual features, preserving global semantic priors while reinforcing local spatial structure information, thereby enhancing the discriminative ability for visually similar voxel categories. Our contributions can be summarized as follows:

- We propose URM strategy to align spatial feature representations between teacher and student models. We also propose a PA module that weights

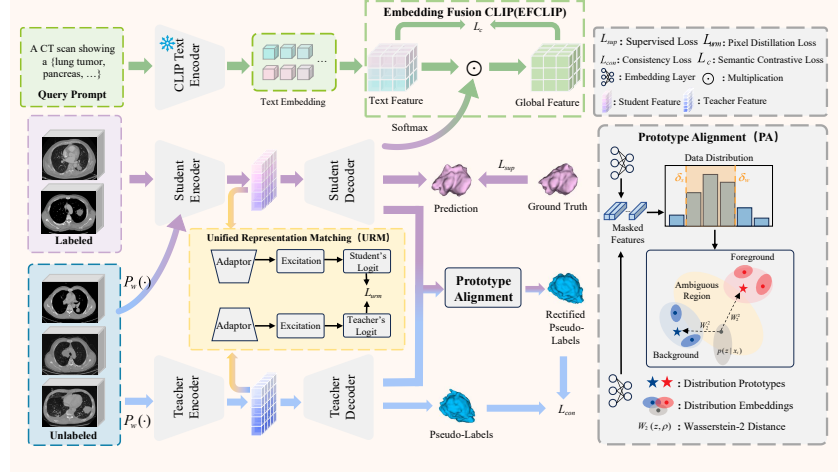


Fig. 1. Overview of PAVL framework, which comprises three main components: URM performs voxel-level knowledge distillation to ensure robust prior transfer; PA weights features by voxel uncertainty to construct accurate class prototypes; and EFCLIP fuses CLIP text–patch embeddings with voxel features.

candidate features by voxel level uncertainty to construct more accurate class prototype distributions.

- EFCLIP is proposed to fuse CLIP’s text embeddings with voxel-level visual features, thereby preserving global semantics and enhancing local spatial detail to improve the discriminative ability for similar voxel categories.
- Extensive experiments on one public dataset and one in-house dataset demonstrate that PAVL outperforms seven advanced SSMIS methods, highlighting its potential for clinical translation.

2 Methodology

The overall framework of PAVL is illustrated in Fig. 1. We briefly outline the SSMIS setting. Let $D_l = \{(x^i, y^i)\}_{i=1}^N$ denote the labeled set, and the unlabeled data is $D_u = \{(x^i)\}_{i=N+1}^{N+M}$, where N and M are the numbers of labeled and unlabeled images, respectively. Each image is formulated as $x^i \in \mathbb{R}^{C \times H \times W \times D}$, and the corresponding ground-truth mask y^i belongs to the set $\{0, 1\}^{H \times W \times D}$. $P_w(\cdot)$ denotes the weak augmentation scheme. The following sections detail the three key components: URM, PA, and EFCLIP.

2.1 Unified Representation Matching

Existing SSMIS methods often overlook precise voxel-level feature alignment, which degrades performance on high-uncertainty images and ambiguous boundary regions. To address this limitation, we introduce a simple but effective

method named unified representation matching (URM). As illustrated in Fig. 1, the student and teacher networks produce embedding features x_s and x_t in a shared space, which are then aligned by a common channel adapter $\phi(\cdot)$:

$$\hat{x}_s = \phi(x_s), \hat{x}_t = \phi(x_t), \quad (1)$$

the adapter consists of three convolutional blocks: a 1×1 convolution to adjust channel dimensionality, a 3×3 convolution for spatial fusion and a final 1×1 convolution to restore the target channels, each block is followed by batch normalization and ReLU. The aligned features $\hat{x}_s$ and $\hat{x}_t$ are then passed through a shared Excitation module $g(\cdot)$ to recalibrate channel weights based on global statistics [7]. Finally, both recalibrated features enter a shared classifier head $h(\cdot)$, producing voxel-wise logits:

$$a_s = h(g(\hat{x}_s)), a_t = h(g(\hat{x}_t)). \quad (2)$$

To encourage the student to replicate the teacher’s discriminative ability at each voxels, we introduce a voxel distillation loss L_{urm} between the two logit maps:

$$L_{urm} = 1 - \frac{1}{N} \sum_{i=1}^N \frac{\langle l_s(i), l_t(i) \rangle}{\|l_s(i)\|_2 \|l_t(i)\|_2}, \quad (3)$$

where $l_s(i)$ and $l_t(i)$ denote the logits a at voxel i produced by the student and teacher networks, respectively. Averaging cosine similarity across voxels encourages directional alignment between student and teacher logits, improving boundary delineation.

2.2 Prototype Alignment

In scenarios with limited memory and computational resources, using all voxels to update class prototypes is impractical. We argue that only highly reliable voxels can produce more representative prototypes. Therefore, for each image, the decoder extracts a feature map $f \in \mathbb{R}^{W \times H \times D}$, which is then projected into the prototype space via a projection head to produce the projected feature map $j(x, y)$ at each spatial location (x, y) . To filter for reliable voxels, we evaluate each embedding’s reliability based on its variance in the projected feature map j . Specifically, we adopt a probability-guided prototype sampling to select representative features set $M(X)$:

$$M(X) = \{j(x, y) | \sigma^2(j(x, y)) \in [Q(S, \delta_s), Q(S, \delta_w)]\}, \quad (4)$$

where $\sigma^2(j(x, y))$ denotes the variance of each projected feature $j(x, y)$, and $S = \{\sigma^2(j(x, y)) | (x, y) \in X\}$ is the set of variances of all candidate embeddings. $Q(S, \delta)$ denotes the value at the δ -quantile of the variance set S , δ_s and δ_w serve as the lower and upper percentile thresholds [22]. By selecting only those features whose variances fall within the interval $[Q(S, \delta_s), Q(S, \delta_w)]$, we dynamically filter for appropriately reliable embeddings and use them to update class prototypes,

thereby enhancing the prototypes' representational capacity. In traditional prototype classifiers, averaging all voxel embeddings ignores uncertainty, which may cause noise from high-uncertainty samples prototype away from the true distribution. To address this issue, we model the embedding of the i -th voxel as a multivariate Gaussian distribution:

$$p(z | x_i) = \mathcal{N}(z; u_i, \text{diag}(\sigma_i^2)), \quad (5)$$

where $z \in \mathbb{R}^D$ denotes the random embedding vector in the latent space for voxel x_i , u_i and σ_i denote the mean and variance vectors of that voxel in latent space. Next, following the probability-guided prototype sampling introduced above, we select the N_p voxels with the smallest variances (the highest-confidence candidates) to form the prototype set $\{x^{(1)}, x^{(2)}, \dots, x^{(N_p)}\}$. We then compute the class- c prototype by a weighted average of these voxels' mean vectors, using the inverse variances as weights:

$$\rho_c = \frac{\sum_{i=1}^{N_p} \frac{1}{\sigma_i^2} u_i}{\sum_{i=1}^{N_p} \frac{1}{\sigma_i^2}}. \quad (6)$$

To quantitatively measure how well an individual voxel's distribution matches the class prototype distribution, we employ the squared Wasserstein-2 distance:

$$W_2^2(p(z|x_i), p(z|X_c)) = \|u_i - \rho_c\|_2^2 + \|\sigma_i - \sigma_c\|_2^2, \quad (7)$$

where W_2^2 represents the Wasserstein-2 distance. Unlike the Euclidean distance, the Wasserstein-2 metric captures both the location and the spread of distributions, penalizing not only mean shifts but also variance discrepancies. This property enables more robust voxel-prototype alignment, particularly in boundary-ambiguous or noise-prone regions. Finally, the rectified pseudo-labels $\hat{y}_r$ by PA are used to compute the consistency loss:

$$L_{con} = \frac{\sum_{i=1}^{|X|} M_i \|\hat{y}_r(i) - \hat{y}_t(i)\|_2^2}{\sum_{i=1}^{|X|} M_i + \varepsilon}, \quad (8)$$

where $\hat{y}_t$ represents pseudo-labels from the teacher model, M_i denotes binary indicator that flags voxel i as ambiguous. ε is a small positive constant added to ensure numerical stability.

2.3 Embedding Fusion CLIP

CLIP emphasizes global semantic information during pre-training, and applying it to voxel-level segmentation often overlooks local details, making visually similar voxels difficult to distinguish. To address this, we propose an embedding fusion CLIP (EFCLIP) mechanism, which integrates CLIP's text-patch embeddings with spatial features to enhance the discriminability for similar voxel categories. Specifically, we feed the manually constructed query text into the CLIP

Table 1. Quantitative comparisons with existing methods. Metrics are reported in “mean_{std}”. **Bold** and underline represent the top-2 results. * denotes that our improvement has statistical significance ($p < 0.05$ via two-sided paired t-test) compared to others.

Method	Labeled Ratio		BraTS2019			In-house		
	Lab.	Unlab.	DSC(%)↑	JC(%)↑	95HD (mm)↓	DSC(%)↑	JC(%)↑	95HD (mm)↓
V-Net [16]	10%	0	71.94 _{3.16}	62.38 _{2.29}	33.61 _{1.24}	73.09 _{1.43}	60.94 _{1.60}	5.97 _{0.46}
V-Net [16]	20%	0	79.26 _{1.53}	70.31 _{1.46}	20.49 _{0.87}	77.72 _{2.13}	65.55 _{2.50}	5.19 _{1.41}
V-Net [16]	100%	0	86.73 _{0.35}	77.90 _{0.44}	7.14 _{0.62}	82.63 _{0.91}	71.31 _{1.05}	3.55 _{0.47}
MC-Net+ [21] (MedIA’22)	10%	90%	82.40 _{2.39}	72.35 _{1.41}	14.91 _{1.31}	81.73 _{1.84}	69.34 _{2.28}	4.31 _{1.33}
URPC [13] (MedIA’22)	10%	90%	82.44 _{1.94}	72.25 _{1.13}	12.15 _{0.81}	78.59 _{0.72}	66.47 _{0.86}	6.35 _{0.61}
CAML [4] (MICCAI’23)	10%	90%	79.52 _{2.59}	68.89 _{1.77}	16.62 _{2.64}	77.82 _{1.41}	64.62 _{1.73}	5.71 _{1.39}
BCP [1] (CVPR’23)	10%	90%	81.98 _{2.93}	71.89 _{2.05}	15.99 _{1.16}	82.05 _{1.48}	70.18 _{1.05}	4.17 _{1.34}
Semi-SAM [24] (BIBM’24)	10%	90%	78.19 _{1.24}	68.77 _{1.37}	17.16 _{3.26}	72.51 _{1.91}	57.66 _{1.60}	6.89 _{1.26}
MR [18] (MedIA’24)	10%	90%	83.26 _{1.15}	72.82 _{1.88}	11.45 _{1.39}	80.12 _{1.79}	68.95 _{1.42}	5.73 _{0.64}
MCL’S [2] (KBS’25)	10%	90%	82.04 _{1.18}	72.07 _{1.13}	12.09 _{0.96}	82.29 _{0.88}	70.67 _{0.93}	4.01 _{0.87}
Ours	10%	90%	84.68 _{1.02}	74.86 _{1.31}	10.41 _{0.73}	83.52 _{0.61}	71.84 _{0.51}	3.25 _{0.39}
MC-Net+ [21] (MedIA’22)	20%	80%	83.55 _{1.13}	74.09 _{1.29}	9.04 _{0.63}	82.45 _{0.69}	71.34 _{1.49}	4.06 _{0.47}
URPC [13] (MedIA’22)	20%	80%	83.33 _{1.45}	73.44 _{1.32}	9.24 _{0.83}	80.13 _{0.57}	67.61 _{0.52}	4.56 _{0.89}
CAML [4] (MICCAI’23)	20%	80%	83.69 _{0.84}	74.01 _{1.07}	8.96 _{0.82}	80.46 _{0.36}	68.83 _{0.85}	4.74 _{0.61}
BCP [1] (CVPR’23)	20%	80%	82.09 _{1.33}	71.78 _{1.72}	10.67 _{0.94}	83.49 _{0.59}	72.22 _{1.20}	3.19 _{0.41}
Semi-SAM [24] (BIBM’24)	20%	80%	80.93 _{2.50}	70.36 _{1.71}	15.25 _{1.36}	77.99 _{1.84}	65.22 _{1.68}	5.96 _{0.86}
MR [18] (MedIA’24)	20%	80%	84.12 _{0.85}	74.79 _{1.09}	8.52 _{0.24}	82.37 _{0.45}	72.14 _{0.80}	3.95 _{0.29}
MCL’S [2] (KBS’25)	20%	80%	84.76 _{1.06}	75.66 _{0.58}	8.01 _{0.74}	83.36 _{0.61}	72.09 _{0.49}	2.89 _{0.42}
Ours	20%	80%	86.10 _{0.83}	77.81 _{0.77}	7.35 _{0.26}	85.04 _{0.29}	74.47 _{0.25}	2.36 _{0.31}

text encoder to extract its text feature vector f_t ; meanwhile, softmax is applied to the student model’s voxel-level output to obtain the spatial feature map f_s , we then compute the element-wise product of f_t and f_s to produce the global feature:

$$f_g = f_t \odot f_s, \quad (9)$$

where $\odot$ denotes element-wise multiplication. On this basis, we introduce contrastive loss to align each global feature f_g with the text feature vectors f_t of all C categories, thereby reinforcing bias correction. The semantic contrastive loss is defined as:

$$L_c = -\frac{1}{C} \sum_{j=1}^C \log \frac{\exp(\text{sim}(f_g, f_{t,j})/\tau)}{\sum_{k=1}^C \exp(\text{sim}(f_g, f_{t,k})/\tau)}, \quad (10)$$

where $\text{sim}(\cdot, \cdot)$ denotes a similarity function between vectors, and τ is a temperature hyperparameter. This contrastive objective effectively guides the model to align visual and textual representations at the global level, enhancing the robustness and accuracy of fine-grained semantic segmentation. The overall loss of the framework can be defined as follows:

$$L = L_{sup} + L_{con} + \alpha L_{urm} + \beta L_c, \quad (11)$$

where the supervised loss L_{sup} combines DSC and cross-entropy loss, L_{con} denotes a consistency loss. L_{urm} and L_c are respectively used to optimize the URM and EFCLIP modules, fully leveraging their roles in feature alignment and semantic enhancement, the hyperparameters α and β are introduced to balance their contributions.

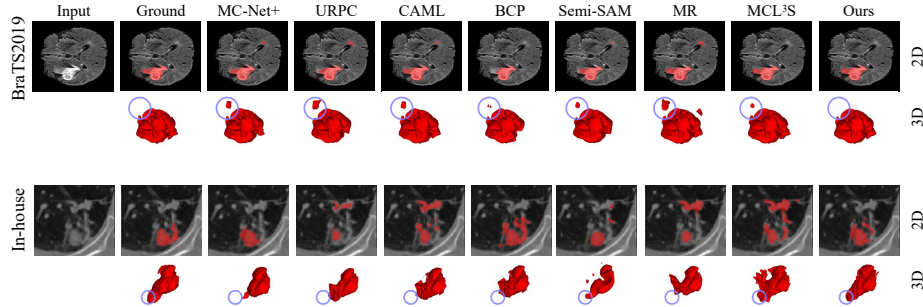


Fig. 2. Visual comparison of results from various methods under the 10% partition protocol on the BraTS19 and in-house datasets.

3 Experiment

3.1 Dataset

We assess our method on one publicly available dataset as well as one in-house dataset. **(1) BraTS2019 Dataset** [14] comprises 335 multi-modal brain MRI scans. All volumes are co-registered to $240 \times 240 \times 155$ with 1 mm isotropic resolution. Following [3], only FLAIR images are used for binary segmentation. During training, each volume is center-cropped to $96 \times 96 \times 96$ [18]. The dataset is split into 250 cases for training, 25 for validation, and 60 for testing. **(2) An in-house lung adenocarcinoma CT dataset** collected by us, comprising 262 scans of lung adenocarcinoma patients. Initial segmentations are obtained via a nnUNet [8] model trained on internal data and subsequently refined by senior radiologists. Original volumes vary between $512 \times 512 \times 97$ and $512 \times 512 \times 456$ voxels; we resample all to $160 \times 160 \times 120$ to ensure uniform input dimensions. The dataset is split into training, validation, and test sets with a ratio of 7:1:2.

3.2 Implementation Details

All experiments are implemented in PyTorch on an NVIDIA RTX 4090 GPU. For fair comparison, all methods adopt V-Net [16] as the segmentation backbone, trained with SGD (learning rate 0.01, weight decay 0.0001) for 8000 iterations. Performance is measured by DSC, Jaccard (JC), and 95% Hausdorff distance (95HD). We randomly designate 10% or 20% of training data as labeled to simulate semi-supervised settings. All models in Table 1 are run three times with different random seeds. The hyperparameters α and β in Eq 11 are set to 0.2 and 0.3, respectively.

3.3 Comparison with Existing Method

We compare PAVL with seven state-of-the-art SSMIS methods: MC-Net+ [21], URPC [13], CAML [4], BCP [1], Semi-SAM [24], MR [18], and MCL³S [2], The

Table 2. Performance comparison of individual components on the in-house dataset using 20% labeled annotations. symbols indicate: ♣ LoRA-based fine-tuning [6]; ♠ CLIP-only [17]; ◇ knowledge distillation [5]; ♡ semantic representational distillation [23]; ★ euclidean distance.

#	URM	PA	EFCLIP	DSC(%)↑	JC(%)↑	95HD (mm)↓
1				77.72	65.55	5.19
2			✓♣	79.06	68.33	4.21
3			✓♠	78.18	67.96	4.83
4			✓	81.19	70.62	3.60
5	✓◇		✓	81.72	70.93	3.39
6	✓♡		✓	82.30	71.84	3.27
7	✓		✓	83.49	72.95	3.18
8		✓★	✓	81.69	70.23	3.73
9		✓	✓	82.93	72.64	3.41
10	✓	✓	✓	85.04	74.47	2.36

results are listed in Tab. 1. On the BraTS2019 dataset, PAVL achieved DSC scores of 84.68% and 86.10% with 10% and 20% labeled data, approaching the fully-supervised ceiling of 86.73%. Despite these difficulties, PAVL sets new state-of-the-art performance, achieving DSC scores of 75.79% and 81.62% with only 10% and 20% labeled data, surpassing the second-best results by 0.96% and 1.08%, respectively. On the in-house dataset, with 20% labeled data, PAVL improves the JC and 95HD by 2.38% and 0.89, respectively, over the previous MCL³S [2]. Furthermore, under 10% and 20% supervision DSC exceeded the fully-supervised upper bound by 0.89% and 2.41%.

3.4 Ablation Studies

We evaluate the contribution of each core component through ablation studies on the in-house dataset, using V-Net as the baseline. As summarized in Tab. 2, Model#1 is trained with labeled data only. To validate the effectiveness of EFCLIP, we compare it with two mainstream adaptation strategies: using LoRA [6] fine-tuning and CLIP-only guidance (Model#2-Model#4). EFCLIP significantly improves DSC and JC over CLIP-only guidance and LoRA by better integrating hierarchical features, overcoming CLIP’s and LoRA limitations in modeling fine-grained structures. In Model #4, EFCLIP’s fusion of semantic features and textual embeddings raise the DSC score to 81.19% and the JC to 70.62%. Building on EFCLIP, we evaluate URM against two standard distillation strategies, KD [5] and SRD [23] (Model#5-Model#7). URM outperforms both by enforcing precise alignment and consistency between pixel-level features and prototype distributions. In the PA module, Wasserstein-2 distance outperforms Euclidean distance in handling ambiguous boundaries (Model#8-Model#9). Finally, the Model#10 combining all core components, achieves the highest segmentation performance, demonstrating the complementary benefits of each module.

3.5 Discussion

Our PAVL increases training cost (25M parameters) due to the URM adapter, Wasserstein-2 computation, and CLIP encoder, though inference overhead is negligible via pre-cached embeddings. Stability may also degrade at very low label ratios (<5%) due to insufficient diversity for prototype sampling and contrastive alignment.

4 Conclusion

In this paper, we present PAVL, a novel CLIP-based SSMIS framework to address class confusion in the feature space and CLIP’s limited ability to capture fine-grained spatial structures. Specifically, PAVL employs URM strategy to align spatial features between teacher and student models, and leverages PA to build more accurate class prototype distributions based on voxel-level uncertainty. Additionally, EFCLIP deeply integrates CLIP’s global semantic priors with voxel-level local spatial structures to improve the separability of visually similar voxel classes. Extensive experiments on two medical image segmentation tasks demonstrate that PAVL significantly outperforms state-of-the-art semi-supervised methods and opens new avenues for clinical application.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (Nos. 82360356, T2541065, and 82272075); and in part by the Guangxi Natural Science Foundation (Nos. 2024GXNSFFA010014 and AB25069496)

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

References

1. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11514–11524 (2023)
2. Chen, J., Feng, S., Cui, Y., Fan, C., Lin, H., Bian, X., Li, L., Liu, Z., Liu, Z., Lan, R., et al.: Multi-layer feature fusion and coarse-to-fine label learning for semi-supervised lesion segmentation of lung cancer. *Knowledge-Based Systems* **317**, 113451 (2025)
3. Chen, S., Bortsova, G., García-Uceda Juárez, A., Van Tulder, G., De Bruijne, M.: Multi-task attention-based semi-supervised learning for medical image segmentation. In: MICCAI. pp. 457–465. Springer (2019)
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. Springer (2023)
5. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: ICLR. pp. 1–13 (2022)

7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR. pp. 7132–7141 (2018)
8. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* **18**(2), 203–211 (2021)
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* **42**(5), 1265–1277 (2022)
10. 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)
11. Li, W., Ma, L., Zhao, W., Yang, H.: Mutual masked image consistency and feature adversarial training for semi-supervised medical image segmentation. *Knowledge-Based Systems* p. 115349 (2026)
12. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 8801–8809 (2021)
13. 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)
14. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging* **34**(10), 1993–2024 (2014)
15. Miao, J., Chen, C., Yuan, Y., Li, Q., Heng, P.A.: Sam-driven cross prompting with adaptive sampling consistency for semi-supervised medical image segmentation. *Medical Image Analysis* p. 103973 (2026)
16. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
17. 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: International conference on machine learning. pp. 8748–8763. PmLR (2021)
18. 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)
19. Sun, B., Li, K., Liu, J., Sun, Z., Wang, X., He, Y., Zhao, X., Xue, H., Hao, A., Li, S., et al.: Advancing mri segmentation with clip-driven semi-supervised learning and semantic alignment. *Neurocomputing* **612**, 128690 (2025)
20. Wu, H., Wang, C., Cui, Z.: Dual cross-image semantic consistency with self-aware pseudo labeling for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging* (2025)
21. 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)
22. Xie, H., Wang, C., Zheng, M., Dong, M., You, S., Fu, C., Xu, C.: Boosting semi-supervised semantic segmentation with probabilistic representations. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2938–2946 (2023)

23. Yang, J., Zhu, X., Bulat, A., Martinez, B., Tzimiropoulos, G.: Knowledge distillation meets open-set semi-supervised learning. *International Journal of Computer Vision* **133**(1), 315–334 (2025)
24. Zhang, Y., Yang, J., Liu, Y., Cheng, Y., Qi, Y.: Semisam: Enhancing semi-supervised medical image segmentation via sam-assisted consistency regularization. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2024)
25. Zheng, B., Zhao, W., Liu, W., He, Z., Qin, C., Yang, H.: Semi-supervised medical image segmentation via pseudo-labeling refinement and dual-adaptive adjustment schemes. *Pattern Recognition* p. 112310 (2025)
26. Zhu, L., Zhou, J., Goh, R.S.M., Liu, Y.: Advmim: Adversarial masked image modeling for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 56–66. Springer (2025)