PC-VLG: Position-Correlated Vision–Language Graph Alignment for Semi-Supervised Medical Image Segmentation

Luyi Qiu[1], Min Dang[2], Feng Pan[3], Yifan Wang[4], and Adams Wai-Kin Kong[1]*

[1] College of Computing and Data Science, Nanyang Technological University, Singapore
adamskong@ntu.edu.sg
[2] School of Computer Science and Technology, Xidian University, Xi'an, China
[3] Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
[4] School of Cyber Science and Engineering, Southeast University, Nanjing, China

Abstract. Semi-supervised medical image segmentation aims to reduce annotation costs, yet achieving anatomically consistent segmentation under limited supervision remains challenging. Most methods rely on visual representations or coarse vision–language alignment strategies that assume spatially agnostic correspondence between modalities, overlooking localization cues in visual features and textual descriptions. Consequently, pixel–text or mask–text matching yields ambiguous cross-modal alignment, particularly for anatomically adjacent or visually similar structures. We propose PC-VLG, a position-correlated vision–language graph framework for semi-supervised medical segmentation that establishes object-level positional correspondence between visual anatomy and textual semantics. A Vision–Language Graph Alignment Network injects multi-modal embeddings into Position-Correlated Cross-Modal Graph modules, where anatomical structures are modeled as graph nodes enriched with visual, textual, and spatial cues, and inter-object relationships are encoded as geometry-aware edges to enable structured and consistent cross-modal alignment beyond conventional pixel–text or mask–text paradigms. Experiments on multi-object datasets demonstrate that PC-VLG outperforms state-of-the-art methods. Code at here.

Keywords: Medical imaging · Semi-supervised segmentation · Segment anything model · Vision-language alignment · Inter-object positional cues

1 Introduction

Accurate medical image segmentation requires costly pixel-level annotations, motivating semi-supervised learning (SSL) with limited labeled data. Most SSL

* Corresponding author: adamskong@ntu.edu.sg

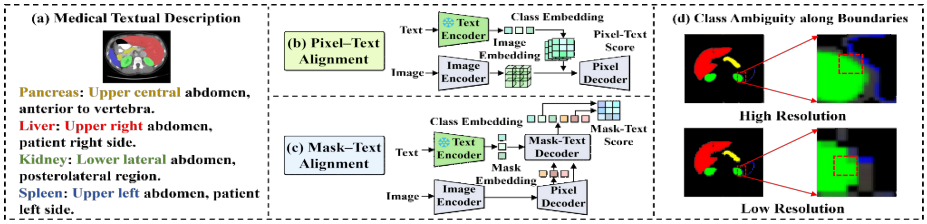

Fig. 1. Overview of text–image alignment and boundary ambiguity. (a) Text description. (b) Pixel–text alignment. (c) Mask–text alignment. (d) Boundary class ambiguity.

methods rely on consistency regularization [12, 17], which enforces prediction invariance under augmentations or perturbations. However, these approaches depend on feature capacity, lack priors, and often yield unstable predictions in complex regions [32]. Recent Segment Anything Models (SAM) demonstrate that large-scale pretraining captures useful inductive biases, leading to medical adaptations such as Med-SAM [20] and SAM-Med2D [34]. Nevertheless, visual representations remain insufficient to distinguish anatomically similar structures.

In parallel, vision–language representation learning has shown that incorporating textual semantics can enhance visual understanding. Representative models such as CLIP [24] and BLIP [14] demonstrate that aligning visual features with language enables context-aware alignment and effective cross-modal representations. Inspired by these advances, textual information has been increasingly explored as complementary guidance for medical image analysis, including LViT [16], ConTEXTual-Net [13], and STPNet [26]. However, most existing approaches overlook explicit localization cues in medical text and imagery that describe spatial positions of anatomical structures. As illustrated in Fig. 1(a), anatomically adjacent organs such as the liver and spleen may share similar visual appearances yet occupy distinct anatomical locations described in medical textual description. Without incorporating these localization cues into vision–language alignment, features from different organs can be matched to the same textual embedding, resulting in ambiguous cross-modal correspondence.

Current vision–language methods perform alignment at the pixel–text [24] or mask–text [10] level via implicit attention mechanisms [7] (Fig. 1(b)). Due to computational constraints, pixel–text alignment is typically conducted on low-resolution features, where encoder stride and pooling blur inter-class boundaries and degrade fine-grained correspondence [24] (Fig. 1(c)). Furthermore, mask–text matching yields global textual embeddings that encode holistic semantics rather than localized structural cues. This global–local discrepancy in representation granularity may introduce semantic inconsistencies and weaken anatomical discrimination [7]. These limitations reveal a structural bottleneck: the absence of explicit modeling of position-correlated object-level correspondence in vision–language medical segmentation. Consequently, anatomically adjacent structures with similar appearances remain difficult to disambiguate.

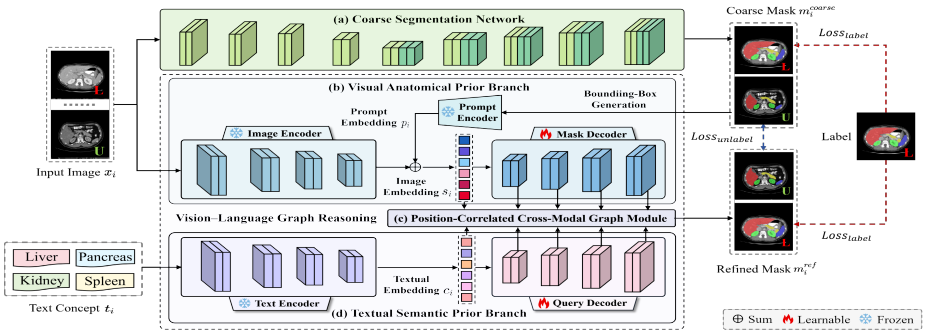

Fig. 2. Overview of the **PC-VLG**, including (a) a coarse segmentation network, (b) a visual anatomical prior branch, (c) a position-correlated cross-modal graph module, and (d) a textual semantic prior branch. 'L' and 'U' denote labeled and unlabeled data.

Motivated by this observation, we identify two challenges: (i) leveraging localization cues in medical text to establish position-level correspondence with visual anatomy and (ii) enabling cross-modal alignment beyond pixel–text or mask–text matching. These limitations arise because existing vision–language segmentation frameworks lack explicit modeling of object-level positional relations. To address this issue, we propose **PC-VLG**, a Position-Correlated Vision–Language Graph framework for semi-supervised medical segmentation that establishes object-level positional correspondence between visual anatomy and anatomical textual semantics through graph-based cross-modal alignment.

Our contributions are summarized as follows: (1) We propose PC-VLG, which models position-correlated object-level correspondence between visual and textual representations for semi-supervised medical segmentation. (2) We develop a Vision–Language Graph Alignment Network that represents anatomical structures as graph nodes with geometry-aware relational modeling for structured cross-modal alignment. (3) Experiments on multi-object benchmarks demonstrate consistent improvements over recent state-of-the-art methods.

2 Method

2.1 Overview of PC-VLG

The Position-Correlated Vision–Language Graph framework (PC-VLG, Fig. 2) addresses cross-modal alignment in semi-supervised medical segmentation by modeling positional relationships between visual and textual cues.

Given an input image, the Coarse Segmentation Network [21] (Fig. 2(a)) produces coarse predictions that provide structural cues for vision-language alignment. Based on coarse predictions, the Vision-Language Graph Alignment Network (VLGR-Net, Fig. 2) employs a Consistency-Aware Prompt Generator (CAPG) [23] to extract bounding-box prompts. The prompts are encoded in the

Visual Anatomical Prior branch (Fig. 2(b)) and fused with image embeddings to inject pretrained anatomical knowledge for decoding. Meanwhile, VLGR-Net incorporates a Textual Semantic Prior branch (Fig. 2(d)) to encode textual descriptions and capture semantics. To establish position-correlated correspondence, PC-VLG introduces Position-Correlated Cross-Modal Graph modules, where anatomical structures are represented as graph nodes with visual, textual, and positional features, while inter-object relations are modeled by geometry-aware edges. This enables structured cross-modal alignment beyond pixel-text alignment. During training, VLGR-Net transfers position-aware vision-language priors to the Coarse Segmentation Network through consistency regularization.

2.2 Vision–Language Graph Alignment Network

The Vision–Language Graph Alignment Network (VLGR-Net) comprises a Visual Anatomical Prior (VAP, Fig. 2(b)) branch, a Textual Semantic Prior (TSP, Fig. 2(d)) branch, and Position-Correlated Cross-Modal Graph (PC-CMG, Fig. 2(c)) modules that model correspondence between anatomy and semantics.

(1) Visual Anatomical Prior Branch. The VAP branch incorporates visual anatomical priors via Med-SAM [20], pretrained on large-scale medical images and comprising an image encoder, a prompt encoder, and a mask decoder. The image and prompt encoders are frozen, while the mask decoder is redesigned with dual-branch attention to recover resolution.

Given an image x_i, the Coarse Segmentation Network produces a coarse prediction m_i^{coarse}, which provides region-level cues. The image is simultaneously encoded by the frozen image encoder $\Phi_{\text{img\_enc}}(\cdot)$ to obtain image embeddings:

$$s_i = \Phi_{\text{img\_enc}}(x_i). \tag{1}$$

For labeled data, CAPG [23] extracts bounding-box prompts from masks and predictions, whereas unlabeled samples use confidence-filtered predictions with a threshold of 0.6. The generated prompts are encoded by the prompt encoder:

$$p_i = \begin{cases} \Phi_{\text{prompt\_enc}}\big(\text{CAPG}(m_i^{coarse}, y_i)\big), & x_i \in \mathcal{D}_l, \\ \Phi_{\text{prompt\_enc}}\big(\mathcal{B}(m_i^{coarse})\big), & x_i \in \mathcal{D}_u. \end{cases} \tag{2}$$

The image embeddings s_i and prompt embeddings p_i are fused by the mask decoder $\Phi_{\text{mask\_dec}}(\cdot)$ to generate mask predictions:

$$m_i^m = \Phi_{\text{mask\_dec}}(s_i, p_i), \tag{3}$$

thereby injecting fine-grained anatomical details from pretrained visual priors.

(2) Textual Semantic Prior Branch. The TSP branch (Fig. 3(a)) encodes anatomical semantics from medical text to complement visual features. Textual knowledge is collected from concept–definition pairs (t_i, p_i) and knowledge-graph triplets (t_i, r_{ij}, t_j), where p_i denotes the definition of concept t_i and r_{ij} specifies the relation between t_i and t_j. Triplets are reformulated into relation-augmented descriptions (e.g., $(t_i + r_{ij}, t_j)$ and $(t_i, r_{ij} + t_j)$). All texts are encoded

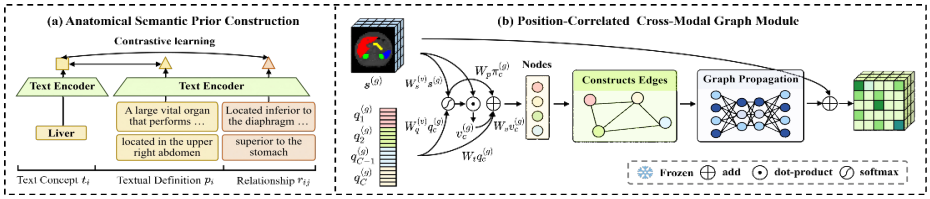

Fig. 3. Overview of the position-correlated cross-modal graph (PC-CMG) module.

by a BERT encoder $\Phi_{\text{text}}(\cdot)$ [8], pretrained on PubMed abstracts [25], yielding domain-specific representations. Given a text sequence t (concept, definition, or relation-augmented form), the encoder produces textual embeddings z_i:

$$z_i = \Phi_{\text{text}}(t), \quad z_i \in \mathbb{R}^d, \quad t \in \{t_i,\, p_i,\, t_i + r_{ij},\, r_{ij} + t_j\}. \tag{4}$$

To enforce semantic consistency, we align concept-definition and relation-augmented embeddings using a contrastive objective. Given paired embeddings $\{(z_i, z_i')\}_{i=1}^N$, the loss pulls matched pairs closer and separates non-matching concepts, where N denotes the number of pairs and k indexes negative samples.

$$Loss_{\text{text}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp(z_i \cdot z_i'/\tau)}{\sum_{k\neq i}\exp(z_i \cdot z_k'/\tau)} + \log\frac{\exp(z_i' \cdot z_i/\tau)}{\sum_{k\neq i}\exp(z_k \cdot z_i'/\tau)}\right],\ \tau{=}0.07 \tag{5}$$

During segmentation, a concept t_i is encoded by the frozen text encoder as $c_i = \Phi_{\text{text}}(t_i)$ and mapped by a six-block transformer query decoder to obtain $q_i = \Phi_{\text{query}}(c_i)$, serving as the semantic cue for position-correlated alignment.

(3) Position-Correlated Cross-Modal Graph Module. We introduce PC-CMG (Fig. 3(b)) to model object-level positional relationships via graph nodes and edges derived from multi-layer visual and textual embeddings, enabling explicit structured graph propagation for cross-modal alignment.

PC-CMG constructs graph nodes via query-guided localization with midline-aligned positional encoding. The VAP mask decoder and TSP query decoder are organized into four aligned stages. At stage g, visual features $s^{(g)} \in \mathbb{R}^{H_g \times W_g \times d_g}$ and semantic queries $Q^{(g)} = \{q_c^{(g)}\}_{c=1}^C$ are obtained, where C denotes the number of concept classes. Nodes are derived through query-guided attention using projections $W_q^{(v)}$ and $W_s^{(v)}$ to a shared space, with spatial index $p \in \{1, \ldots, H_g W_g\}$ corresponding to coordinates (x_p, y_p) and satisfying $\sum_p \alpha_{c,p}^{(g)} = 1$:

$$\alpha_{c,p}^{(g)} = \text{softmax}_p\left((W_q^{(v)} q_c^{(g)})^\top (W_s^{(v)} s^{(g)}(p))\right), \quad v_c^{(g)} = \sum_p \alpha_{c,p}^{(g)} s^{(g)}(p). \tag{6}$$

The object centroid is $(x_c^{(g)}, y_c^{(g)}) = \sum_p \alpha_{c,p}^{(g)}(x_p, y_p)$. For laterality-sensitive organs (e.g., left/right kidneys), a midline-aligned coordinate $\tilde{x}_c^{(g)} = x_c^{(g)} - x_{\text{mid}}$ is used, where x_{mid} denotes the anatomical midline from data. The positional

embedding is $\pi_c^{(g)}$=PE$(\tilde{x}_c^{(g)}, y_c^{(g)})$, with PE$(\cdot)$ sinusoidal encoding. Nodes are:

$$n_c^{(g)} = W_v v_c^{(g)} + W_p \pi_c^{(g)} + W_t q_c^{(g)}. \tag{7}$$

PC-CMG constructs relation-aware edges under uncertainty gating. A fully connected graph is formed. For nodes indexed by a and b, edge weights as $ew_{ab}^{(g)}$:

$$ew_{ab}^{(g)} = \mathrm{softmax}_{b\neq a}\Big((W_q^{(e)} n_a^{(g)})^\top (W_k^{(e)} n_b^{(g)}) + W_g g_{ab}^{(g)} + W_r e_{ab}\Big) \cdot \omega_{ab}^{(g)}, \tag{8}$$

where $g_{ab}^{(g)} = \mathrm{ReLU}\big(W_\phi[\Delta x, \Delta y, \|\Delta\|_2]^\top\big)$ encodes relative geometry with $\Delta x = \tilde{x}_a^{(g)} - \tilde{x}_b^{(g)}$ and $\Delta y = y_a^{(g)} - y_b^{(g)}$. $W_q^{(e)}, W_k^{(e)}, W_\phi$ are learnable projections, where $E_r \in \mathbb{R}^{n\times 256}$ encodes n anatomical relations, and r_{ab} indexes the corresponding row. The uncertainty gate is defined as $u_a^{(g)}$.

$$u_a^{(g)} = -\frac{1}{\log(H_g W_g)} \sum_p \alpha_{a,p}^{(g)} \log \alpha_{a,p}^{(g)}, \quad \omega_{ab}^{(g)} = (1 - u_a^{(g)})(1 - u_b^{(g)}). \tag{9}$$

PC-CMG performs graph propagation for progressive refinement. For nodes indexed by a, message passing updates representations as Eq. 10, where W_m is a learnable projection, and FFN$(\cdot)$ denotes a position-wise feed-forward network:

$$n_a^{(g)} \leftarrow \mathrm{FFN}(n_a^{(g)} + \sum_{b\neq a} ew_{ab}^{(g)} W_m n_b^{(g)}). \tag{10}$$

where $b \neq a$ indicates that self-connections are excluded from message aggregation. Refined nodes are projected back to the spatial domain:

$$\hat{s}^{(g)}(p) = s^{(g)}(p) + \sum_a \beta_{a,p}^{(g)} W_o n_a^{(g)}, \quad \beta_{a,p}^{(g)} = \mathrm{softmax}_a((W_b n_a^{(g)})^\top s^{(g)}(p)). \tag{11}$$

where W_o, W_n are learnable projections, $\beta_{a,p}^{(g)}$ denotes the node-to-pixel attention weight between node a and location p. The updated features $\hat{s}^{(g)}$ are added to the next-stage visual features $s^{(g+1)}$ and propagated to the next stage for refinement.

2.3 Progressive Consistency Learning

Given input x_i with label y_i, PC-VLG produces coarse and refined predictions m_i^c, m_i^{ref}. Pseudo-labels are obtained by $\arg\max(\cdot)$ with detached gradients.

$$f_{\mathrm{CSN}}(x_i) = \arg\max(m_i^c), \quad f_{\text{VLGR-Net}}(x_i) = \arg\max(m_i^{ref}). \tag{12}$$

For unlabeled data, cross-branch consistency is enforced using dice loss, whereas labeled predictions are supervised using cross-entropy and dice losses:

$$\begin{aligned} L_{\mathrm{unlabel}}(x_i) =& L_{\mathrm{Dice}}(m_i^c, f_{\mathrm{VLGR}}(x_i)) + L_{\mathrm{Dice}}(m_i^{ref}, f_{\mathrm{CSN}}(x_i)), \\ L_{\mathrm{label}}(x_i, y_i) =& L_{\mathrm{CE}}(m_i^c, y_i) + L_{\mathrm{CE}}(m_i^{ref}, y_i) \\ &+ L_{\mathrm{Dice}}(f_{\mathrm{CSN}}(x_i), y_i) + L_{\mathrm{Dice}}(f_{\mathrm{VLGR}}(x_i), y_i). \end{aligned} \tag{13}$$

The overall objective is defined in Eq. 14, where μ follows a Gaussian warm-up schedule with $\mu_{\max} = 0.1$ [19] to stabilize early-stage pseudo-supervision.

$$L_{\mathrm{Total}} = L_{\mathrm{label}} + \mu\, L_{\mathrm{unlabel}}. \tag{14}$$

Table 1. Ablation study under 10% labeled data on datasets, using Dice for FLARE.

Dataset	Baseline	VAP Branch	TSP Branch	PC-CMG Module	Liver (%)	Kidney (%)	Spleen (%)	Pancreas (%)
FLARE	✓	×	×	×	84.22	67.54	61.65	36.46
	✓	✓	×	×	88.64	82.91	74.23	39.87
	✓	✓	✓	×	90.81	85.37	77.16	44.09
	✓	✓	✓	✓	92.73	88.02	79.59	47.38
Dataset	Baseline	VAP Branch	TSP Branch	PC-CMG Module	DSC (%)	Jaccard (%)	HD95 (mm)	ASD (mm)
ACDC	✓	×	×	×	85.79	73.86	6.24	2.73
	✓	✓	×	×	87.32	78.41	4.77	1.92
	✓	✓	✓	×	88.49	79.26	4.15	1.58
	✓	✓	✓	✓	89.55	79.98	3.30	1.06

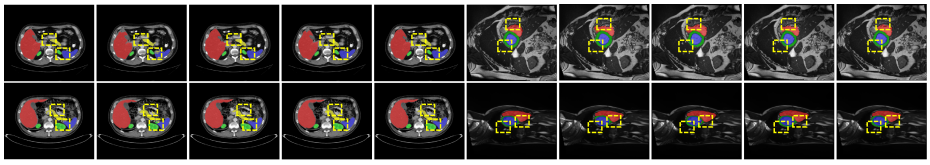

Fig. 4. Comparison on FLARE21 (left) and ACDC (right) under 10% labeled data. (a) Label. (b) Baseline. (c) + VAP. (d) + TSP. (e) + PC-CMG.

3 Experiments

3.1 Experimental Settings

Datasets and Evaluation Metrics. (a) FLARE21 [1] comprises 361 abdominal CT scans annotated for liver, kidney, spleen, and pancreas, split into 260 scans for training, 26 for validation, and 75 for testing. (b) ACDC [4] includes 100 cardiac MR volumes annotated for right ventricle (RV), left ventricle (LV), and myocardium (MYO). Following [28], 75 cases are used for training, 5 for validation, and 20 for testing. Performance is evaluated using DSC, Jaccard index, 95% Hausdorff distance (HD95), and average surface distance (ASD).

Implementation Details. All volumes are converted to axial slices, resized to 256×256, and normalized to $[0, 1]$. Random rotation, flipping, and zooming are used for augmentation. Med-SAM image and prompt encoders are frozen for anatomical priors. VLGR-Net is used during training, while inference relies solely on the coarse segmentation network [11]. The VAP decoder contains 4 stages with 256-dimensional features and 8-head attention. PC-CMG adopts 4 graph stages for relation-aware semantic modeling. Organ-specific textual descriptions are derived from PubMed-based anatomical knowledge and relation triplets. The coarse segmentation network uses SGD with learning rate 0.03, while VLGR-Net uses AdamW with learning rate 10^{-4}. Models are trained for 90k iterations with batch size 4. Experiments are implemented in PyTorch on RTX 3090 GPUs.

3.2 Experimental Results

Ablation Study. Table 1 and Fig. 4 report ablation results on the U-Mamba baseline [11]. **(1) FLARE21.** VAP improves Dice across organs, and TSP en-

Table 2. Semi-supervised results on FLARE21. All comparative results are from [1].

Method	with 10% Labeled Data									
	Liver		Kidney		Spleen		Pancreas		Avg	
	DSC (%)	HD95 (mm)	DSC (%)	HD95 (mm)	DSC (%)	HD95 (mm)	DSC (%)	HD95 (mm)	DSC (%)	HD95 (mm)
MT [28] (NIPS)	88.77	12.09	83.34	8.70	72.91	35.89	38.01	17.64	70.76	18.58
SASSNet [15] (MICCAI)	86.94	24.59	63.59	15.10	59.83	51.86	35.36	19.53	61.43	27.76
DTC [18] (AAAI)	87.99	21.63	83.11	18.80	66.04	32.64	35.15	19.31	68.07	23.11
UAMT [33] (MICCAI)	91.65	10.44	84.70	8.08	76.16	20.44	42.01	18.24	73.63	14.30
DUMT [29] (MICCAI)	87.28	13.23	80.47	19.21	68.23	36.17	40.18	20.77	69.04	22.35
URPC [19] (MedIA)	91.09	11.71	85.88	7.41	75.40	20.82	40.89	16.96	73.31	14.23
AAUM [1] (MIA)	90.78	10.85	87.09	9.48	78.13	18.45	45.12	15.98	75.28	13.69
PC-VLG (Proposed)	**92.73**	**9.24**	**88.02**	**8.87**	**79.59**	**17.14**	**47.38**	**14.65**	**76.93**	**12.48**
Method	with 20% Labeled Data									
	Liver		Kidney		Spleen		Pancreas		Avg	
	DSC (%)	HD95 (mm)	DSC (%)	HD95 (mm)	DSC (%)	HD95 (mm)	DSC (%)	HD95 (mm)	DSC (%)	HD95 (mm)
MT [28] (NIPS)	89.82	11.04	85.15	10.89	71.87	25.70	41.55	17.94	72.10	16.39
SASSNet [15] (MICCAI)	88.41	34.01	86.19	11.89	64.11	32.28	40.25	17.16	69.74	23.84
DTC [18] (AAAI)	89.61	25.23	83.31	18.09	62.76	29.05	38.29	17.46	68.49	22.46
UAMT [33] (MICCAI)	89.54	16.60	87.92	7.83	73.07	17.91	48.34	15.66	74.72	14.50
DUMT [29] (MICCAI)	90.11	11.74	85.43	8.64	71.83	25.43	40.94	16.31	72.08	15.53
URPC [19] (MedIA)	91.02	11.16	87.91	8.47	72.06	20.66	46.03	16.33	74.26	14.16
AAUM [1] (MIA)	91.84	11.32	88.72	7.79	78.07	17.38	48.14	15.94	76.69	13.11
PC-VLG (Proposed)	**93.29**	**9.84**	**89.53**	**7.21**	**79.46**	**16.19**	**49.62**	**14.48**	**77.98**	**11.93**

Table 3. Semi-supervised results on ACDC. All comparative results are from [32].

Method	with 10% Labeled Data				with 20% Labeled Data			
	DSC (%)	Jaccard (%)	HD95 (mm)	ASD (mm)	DSC (%)	Jaccard (%)	HD95 (mm)	ASD (mm)
UAMT [33] (MICCAI)	81.32	70.58	13.17	3.77	85.62	76.65	9.31	1.53
SASSNet [15] (MICCAI)	84.26	74.21	6.08	1.76	86.98	77.34	5.36	2.52
Tri-U-MT [28] (MICCAI)	83.71	73.99	7.54	2.73	87.04	78.09	5.62	1.63
DTC [18] (AAAI)	82.43	71.15	8.82	3.15	86.13	76.87	6.28	2.31
CoraNet [27] (MedIA)	84.17	74.08	6.18	2.41	86.32	77.03	6.45	2.21
MC-Net+ [31] (MedIA)	86.55	77.13	7.01	2.13	88.35	79.11	5.76	1.73
URPC [19] (MedIA)	84.72	74.26	5.12	1.64	87.18	78.30	5.29	1.64
PLCT [5] (MedIA)	86.48	76.73	6.69	2.32	88.26	79.07	5.84	2.13
MCF [30] (CVPR)	86.95	77.43	5.04	1.86	88.47	79.92	5.61	1.79
CAML [9] (MICCAI)	87.46	78.51	5.01	1.39	89.06	80.49	5.21	1.56
DCNet [6] (MICCAI)	87.64	78.83	4.89	1.26	89.31	80.57	4.86	1.19
PatchCL [3] (CVPR)	87.21	78.46	5.15	1.68	89.06	80.33	5.14	1.61
SC-SSL [22] (TMI)	83.14	73.31	9.47	2.63	86.52	77.81	6.73	1.86
BCP [2] (CVPR)	88.02	79.13	4.68	1.43	89.12	80.42	4.42	1.29
BoCLIS [32] (TMI)	88.96	79.72	4.23	1.21	90.10	81.11	4.14	1.06
PC-VLG (Proposed)	**89.55**	**79.98**	**3.30**	**1.06**	**90.62**	**81.87**	**2.91**	**0.74**

hances pancreas performance. PC-CMG yields gains for large and small objects. Visual comparisons in Fig. 4 show clearer boundaries and reduced confusion among adjacent objects. **(2) ACDC.** VAP and TSP improve DSC and Jaccard, while PC-CMG reduces HD95 and ASD, validating cross-modal alignment.

Comparison with State-of-the-Art Methods. Tables 2 and 3 show that: **(1) FLARE21.** PC-VLG outperforms advanced methods under the same settings, indicating that position-correlated graph modeling and cross-modal alignment enhance anatomical consistency and boundary discrimination. **(2) ACDC.** PC-VLG achieves superior accuracy and boundary precision, indicating improved structural consistency through positional correspondence.

4 Conclusion

We propose PC-VLG, a position-correlated vision-language graph framework for semi-supervised medical image segmentation. PC-VLG establishes object-

level positional correspondence between visual anatomy and anatomical textual semantics through graph modeling. By integrating visual anatomical priors, textual semantic priors, and geometry-aware relational modeling, PC-VLG enables coherent cross-modal alignment beyond pixel-text or mask-text correspondence. Experiments on multi-object datasets demonstrate that PC-VLG consistently outperforms state-of-the-art semi-supervised segmentation methods.

Acknowledgment The authors would like to thank Nanyang Technological University for providing computational resources and support.

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

References

1. Adiga, S., Dolz, J., Lombaert, H.: Anatomically-aware uncertainty for semi-supervised image segmentation. Med. Image Anal. **91**, 103011 (2024)
2. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 11514–11524 (2023)
3. Basak, H., Yin, Z.: Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 19786–19797 (2023)
4. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging. **37**(11), 2514–2525 (2018)
5. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. Med. Image Anal. **87**, 102792 (2023)
6. Chen, F., Fei, J., Chen, Y., Huang, C.: Decoupled consistency for semi-supervised medical image segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. pp. 551–561. Springer (2023)
7. Das, A., Hu, X., Jiang, L., Schiele, B.: Mta-clip: Language-guided semantic segmentation with mask-text alignment. In: European Conference on Computer Vision. pp. 39–56. Springer (2024)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1. pp. 4171–4186 (2019)
9. Gao, S., Zhang, Z., Ma, J., Li, Z., Zhang, S.: Correlation-aware mutual learning for semi-supervised medical image segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. pp. 98–108. Springer (2023)
10. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: European conference on computer vision. pp. 540–557. Springer (2022)

11. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
12. Huang, W., Zhang, L., Wang, Z., Wang, Y.: Gapmatch: Bridging instance and model perturbations for enhanced semi-supervised medical image segmentation. In: Proc. AAAI Conf. Artif. Intell. vol. 39, pp. 17458–17466 (2025)
13. Huemann, Z., Tie, X., Hu, J., Bradshaw, T.J.: Contextual net: a multimodal vision-language model for segmentation of pneumothorax. Journal of Imaging Informatics in Medicine **37**(4), 1652–1663 (2024)
14. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML. pp. 12888–12900 (2022)
15. Li, S., Zhang, C., He, X.: Shape-aware semi-supervised 3d semantic segmentation for medical images. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. pp. 552–561. Springer (2020)
16. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q.: Lvit: language meets vision transformer in medical image segmentation. IEEE Trans. Med. Imaging. **43**(1), 96–107 (2023)
17. Li, Z., Zheng, Y., Shan, D., Yang, S., Li, Q., Wang, B., Zhang, Y., Hong, Q., Shen, D.: Scribformer: Transformer makes cnn work better for scribble-based medical image segmentation. IEEE Trans. Med. Imaging. **43**(6), 2254–2265 (2024)
18. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: Proc. AAAI Conf. Artif. Intell. vol. 35, pp. 8801–8809 (2021)
19. Luo, X., Hu, M., Song, T., Wang, G., Zhang, S.: Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In: International conference on medical imaging with deep learning. pp. 820–833. PMLR (2022)
20. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
21. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
22. Miao, J., Zhou, S.P., Zhou, G.Q., Wang, K.N., Yang, M., Zhou, S., Chen, Y.: Sc-ssl: Self-correcting collaborative and contrastive co-training model for semi-supervised medical image segmentation. IEEE Trans. Med. Imaging. **43**(4), 1347–1364 (2023)
23. Qiu, L., Till, T., Dang, M., Kong, A.W.K.: Spc-seg: A sam-guided progressive consistency framework with anatomical priors for scribble-supervised segmentation. In: ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4291–4295. IEEE (2026)
24. 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 (2021)
25. Roberts, R.J.: Pubmed central: The genbank of the published literature (2001)
26. Shan, D., Li, Z., Li, Y., Li, Q., Tian, J., Hong, Q.: Stpnet: Scale-aware text prompt network for medical image segmentation. IEEE Trans. Image Process. (2025)
27. Shi, Y., Zhang, J., Ling, T., Lu, J., Zheng, Y., Yu, Q., Qi, L., Gao, Y.: Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. IEEE Trans. Med. Imaging. **41**(3), 608–620 (2021)
28. Wang, K., Zhan, B., Zu, C., Wu, X., Zhou, J., Zhou, L., Wang, Y.: Tripled-uncertainty guided mean teacher model for semi-supervised medical image segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. pp. 450–460. Springer (2021)

29. Wang, Y., Zhang, Y., Tian, J., Zhong, C., Shi, Z., Zhang, Y., He, Z.: Double-uncertainty weighted method for semi-supervised learning. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. pp. 542–551. Springer (2020)
30. Wang, Y., Xiao, B., Bi, X., Li, W., Gao, X.: Mcf: Mutual correction framework for semi-supervised medical image segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 15651–15660 (2023)
31. Wu, Y., Ge, Z., Zhang, D., Xu, M., Zhang, L., Xia, Y., Cai, J.: Mutual consistency learning for semi-supervised medical image segmentation. Med. Image Anal. **81**, 102530 (2022)
32. Yang, Y., Zhuang, J., Sun, G., Wang, R.: Boundary-guided contrastive learning for semi-supervised medical image segmentation. IEEE Trans. Med. Imaging. (2025)
33. 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: Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. pp. 605–613. Springer (2019)
34. Zhao, Y., Zhou, T., Gu, Y., Zhou, Y., Zhang, Y., Wu, Y., Fu, H.: Segment anything model-guided collaborative learning network for scribble-supervised polyp segmentation. arXiv preprint arXiv:2312.00312 (2023)