Segmenting Vessels Encapsulating Tumor Clusters via Fine-Grained Visual Prompt

Jiahui Yu[⋆1,2], Tianyu Ma[⋆3], Shenjian Gu[2], Yuping Guo[2], Feng Chen[4], Xiaoxiao Li[5], and Yingke Xu[(✉)1,2]

[1] Zhejiang University, Hangzhou 310027, China
yingkexu@zju.edu.cn
[2] Binjiang Institute of Zhejiang University, Hangzhou 310027, China
[3] Ai CoVision, Inc., Hangzhou 310050, China
[4] The First Affiliated Hospital of Zhejiang University, Hangzhou 310011, China
[5] The University of British Columbia, Vancouver BC V6T 1Z4, Canada

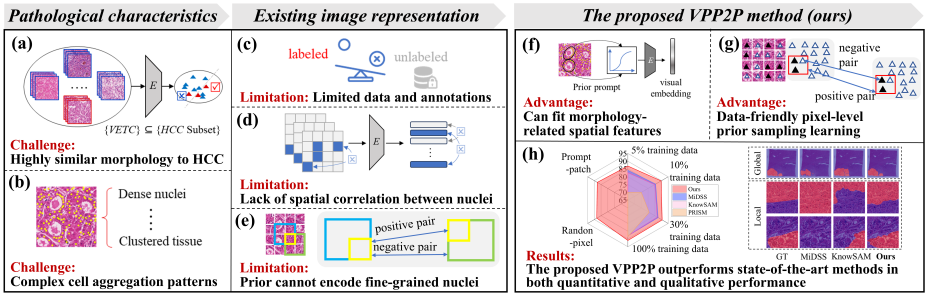

Fig. 1. Comparison between previous work and ours. The morphological patterns (a) and complex characteristics (b) of VETC pose challenges for segmentation. Previous works utilized limited annotations (c), relying on feature encoders (d) or visual prompt (e) to learn patch-level representations. We fitted pathology-specific prompt (f) and propagated them to the pixel level (g) for sampling and learning. The results (h) demonstrate that our method achieves superior performance.

Abstract. Segmenting hepatocellular carcinoma (HCC) and vessels encapsulating tumor clusters (VETC) are new paradigm for prognostic analysis. However, the clustered morphology of VETC nuclei, which is difficult to represent at the patch level, makes segmentation highly challenging. Recent visual prompt-based methods incorporating nucleus prior knowledge have shown promise but assume patch pixels lack spatial correlation, failing to capture nuclei morphology at the pixel level. To address this, we propose a Patch-to-Pixel Visual Prompt (VPP2P) framework, which models VETC morphological features by propagating visual prompts from patches to pixels. Built on contrastive learning, our semi-supervised approach samples positive and negative pairs within patches

[1] ⋆ Contributed equally

to enhance feature learning. Experiments show that VPP2P achieves performance comparable to fully supervised methods using only 10% of the training data. With 30% of the training data, VPP2P attains a Dice score of 90.52%, outperforming state-of-the-art visual prompt-based methods by an average margin of 6.6%. To the best of our knowledge, this is the first semi-supervised deep learning approach for VETC morphological analysis, offering new insights into HCC clinical research. Code is available at https://github.com/sm8754/VPP2P.

Keywords: Whole slide images · VETC · Semi-supervised · Visual prompt.

1 Introduction

The segmentation of hepatocellular carcinoma (HCC) and vascular encapsulating tumor clusters (VETC) can assist in analyzing the prognostic status of patients, which holds significant importance for cancer patients [1, 2]. In recent years, deep learning methods have achieved remarkable results in the tasks of HCC and VETC segmentation [3, 4]. The primary challenges in VETC segmentation stem from the extremely limited availability of annotated data and the unique morphological features of VETC images.

A major challenge in segmentation tasks is the limited availability of annotated data. The mainstream approach to addressing this issue relies on semi-supervised learning, primarily through self-training and consistency regularization [5, 6]. Self-training methods generate pseudo-labels for unlabeled data and use them alongside true labels to enhance supervision [7, 8]. Consistency regularization, on the other hand, ensures that models produce stable predictions by applying various perturbations or transformations to unlabeled samples [9–11]. Despite their success in many domains, these methods struggle with segmentation granularity in VETC segmentation [12, 13]. The key limitation is that they extract features at the patch level while overlooking the morphological structure of VETC nuclei, which appear in clustered patterns that are only discernible at the pixel level. These models assume that pixels within a patch are spatially independent, an assumption that fails due to the uneven distribution of pixels within clusters. Visual prompting has emerged as a potential solution by incorporating prior knowledge about VETC morphology into the training process. For instance, QAP trains a nucleus segmentation model while encoding the spatial distribution of nuclei, and visual prompts leverage spatial and morphological attributes to guide the backbone network [14]. However, these approaches still rely on the flawed assumption of spatial independence within patches, failing to capture specific spatial nuclear patterns. Therefore, a key research direction is exploring how to apply visual prompts at a finer-grained level to improve segmentation accuracy.

To address these limitations, we propose VPP2P, a semi-supervised **V**isual **P**rompt framework that models morphological features from **P**atch to **P**ixel. VPP2P utilizes nucleus location information as visual prompts and further samples positive and negative pairs within the patch, propagating the visual prompts

to the pixel level to model complex morphological features inside the patch. Experiments show that with 30% training dataset, VPP2P achieves a Dice of 90.52%, outperforming state-of-the-art visual prompt-based methods by 6.6%, and an HD95 of 9.94%, at least 2% lower, demonstrating superior performance.

Our contributions can be summarized as follows. 1) We introduce a novel visual prompt method in VETC segmentation. This method introduces semantic supervisory signals by analyzing the locations and quantities of nuclei, thereby addressing the bottleneck of insufficient prior morphological knowledge for VETC fitting. 2) We devise a novel semi-supervised network architecture that further sampling positive and negative pairs within patches to propagate the visual prompts to pixel level. This architecture addresses the issue of lacking semantic supervisory signals. 3) We pioneer pixel-level semantic supervision signals in facilitating the training of the VETC segmentation model to achieve impressive performance, which offers fresh perspectives on the development of modern VETC segmentation schemes.

2 Method

As shown in Fig.2, given an input $I \in \mathbb{R}^{H\times W\times 3}$, two stages perform distinct learning processes. The supervised stage trains the teacher model using labeled data with cross-entropy $\mathcal{L}_{\text{Sent}}$. In the unsupervised stage, strong and weak augmentations of I are fed into the student and teacher models, respectively. Both branches share an extractor, projector, and predictor. Augmented images are also input to the Morphology-Aware Prompt Generator (MAPG), whose parameters are independently updated to generate visual prompts $V = \{v_1, v_2, \cdots, v_n\}$. Visual prompts are dynamically fused with deep features to refine contrastive cues through cross-layer positive/negative pair matching.

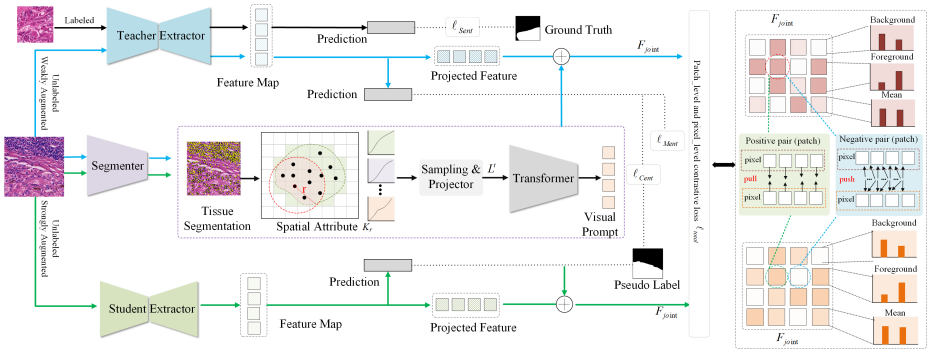

Fig. 2. Overview of our method. VPP2P adopts a dual-branch architecture with supervised and unsupervised learning. The unsupervised branch generates pseudo-labels via the student model and optimizes training with $\mathcal{L}_{\text{Sent}}$, $\mathcal{L}_{\text{Ment}}$, $\mathcal{L}_{\text{Cent}}$, and $\mathcal{L}_{\text{total}}$. The teacher model uses EMA for stability. VPP2P integrates MAPG to analyze nuclei distribution, generating discriminative prompts with independent parameter updates.

2.1 Visual-prompt generation

Inspired by [15], we enhance VETC segmentation by analyzing tissue morphology. We compute spatial proximity between each nucleus (named 'source') and its surrounding nuclei (named 'target') to generate visual prompts V. Specifically, the aggregation/dispersion relationship can be characterized as a function of the number of target objects varying with their spatial distance from source objects. Given the parsed tissue structure information $T_c = \{t_n^c, c = 1, \cdots, C, n = 1, \cdots, N\}$ from an input image I, where t_n^c denotes the n-th object of type c, we count the number of neighboring points between source nuclei $t^m_{i=1,\cdots,n_m} \in T_c$ and target nuclei $t^n_{j=1,\cdots,n_m} \in T_c$ within a circular region of radius r. This process is formulated as follows:

$$K_r = \frac{S}{\lambda}\left(\sum_{j=1}^{n_m} \mathbb{I}\left(\min_{i=1}^{n_m} \|t_i^m - t_i^n\|_2 \le r\right)\right)^2 \cdot \omega_{ij}, \quad \text{where } s_n \text{ and } tn \in [1, C] \quad (1)$$

where S represents the area of the circular region with radius r, $\mathbb{I}(\cdot)$ is the indicator function, C is the total number of nucleus types, $\|\cdot\|_2$ denotes the Euclidean distance between two objects, ω_{ij} is the edge correction factor, and λ is the intensity parameter.

To generate visual prompts V that are highly correlated with image semantics, we discretize the original spatial attribute K_r into compact semantic units $L^t = \{l_1^t, l_2^t, \ldots, l_n^t\}$ using linear quantization based on uniform sampling μ. To enhance feature robustness, we use a learnable projection function ρ to suppress noise while preserving key attributes. A multi-layer Transformer encoder Encoder$(\cdot)$ models implicit relationships via self-attention, generating structured semantic representations $\tilde{L}^t$.

Based on the quantized representation $\tilde{L}^t$, we generate task-specific visual prompts encoding pixel-level VETC spatial attributes from nucleus segmentation. The prompt sequence $V = \{v_0, v_1, \ldots, v_n\}$ through a conditional probability maximization strategy. Specifically, the generation process follows a joint probability decomposition model based on the chain rule, expressed as:

$$P(V|\tilde{L}^t) = P(v_1|\tilde{L}^t)\prod_{i=2}^{n} P(v_i|V_{1,\ldots,i-1}, \tilde{L}^t) \quad (2)$$

Each prompt unit is generated under the constraints of the historical sequence $V_{1,\ldots,i-1}$ and dynamically linked to quantized units $\tilde{L}^t$ via cross-attention. we adopt a multi-layer Transformer decoder Decoder$(\cdot)$, where the i-th output captures temporal dependencies through masked self-attention and integrates attribute semantics from $\tilde{L}^t$. Learnable initial embeddings V' initiate the process, enabling the generation of visual prompts V to combine historical context with quantized semantic features.

2.2 Comparative learning across levels

First, we extract a joint feature map $F_{\text{joint}} = \{f_{h,w} \mid h \in H, w \in W\}$ from the image and visual prompts, where h and w represent the spatial dimensions of fea-

ture map. Subsequently, the feature map is partitioned into n_p non-overlapping patches $\{f_p^{(i)}\}_{i=1}^{n_p}$, where the subscript p denotes patch-level feature. We define positive pairs as spatially aligned teacher-student model pairs at two levels: (1) patch-level pairs $\{f_p^{(i)}, \tilde{f}_p^{+(i)}\}$ and (2) pixel-level pairs $\{f_{px}^{(h,w)}, \tilde{f}_{px}^{+(h,w)}\}$, where f and $\tilde{f}$ denote features from the teacher and student models, respectively.

For negative pairs, we use a heterogeneity-aware sampling strategy. By calculating the foreground pixel proportion FP^2 for each patch, we select pairs with maximal FP^2 divergence (high foreground proportion (HFP) vs. high background proportion (HBP)) to form negative pairs $\{f_p^{(j)}, \tilde{f}_p^{-(k)}\}$. The computation is formulated as:

$$FP^2 = \frac{1}{\Omega_p} \sum_{(h,w)\in\Omega_p} \mathbb{I}\left(M_{\text{stuffed}}(h,w) = 1\right) \tag{3}$$

where Ω_p represents the set of pixel coordinates within a patch, $M_{\text{stuffed}}(h,w)$ denotes pseudo-labels from the student or teacher models, and $\mathbb{I}(\cdot)$ is the indicator function. For pixel $\{f_{px}^{(h,w)}\}_{(h,w)\in\Omega_p}$ within each patch, we establish many-to-many relationships. For negative patches, all pixels are treated as negative pairs $\{f_{px}^{(h,w)}, \tilde{f}_p^{-(h,w)}\}$ to enhance fine-grained discriminative capability, while preserving local consistency among pixel pairs in positive patches.

We employ an exponential equation based on cosine distance to enhance compactness for positive pairs and separation for negative pairs at patch and pixel levels. A contrastive learning process is designed for each pixel feature to pull positive samples closer and push negative samples apart, formulated as:

$$\mathcal{L}_{\text{pixel}} = -\frac{1}{|\Omega_p|} \sum_{(h,w)\in\Omega_p} \log\left(\frac{\exp\left(\left(f_{\text{px}}^{(h,w)}, \tilde{f}_{\text{px}}^{+(h,w)}\right)/\tau\right)}{\exp\left(\left(f_{\text{px}}^{(h,w)}, \tilde{f}_{\text{px}}^{(h,w)}\right)/\tau\right)}\right) \tag{4}$$

where $f_{\text{px}}^{(h,w)}$ denotes the pixel feature at spatial location (h,w), $\tilde{f}_{\text{px}}^{+(h,w)}$ is its corresponding positive sample feature, $\tilde{f}_{\text{px}}^{(h,w)}$ represents the i-th matched pixel-level feature, and τ is a temperature parameter to regulate the similarity distribution.

To further incorporate semantic information at the patch-level, we implicitly integrate hierarchical contrastive learning into the overall computation. The formulation is expressed as:

$$\mathcal{L}_{\text{total}} = -\frac{1}{n_{\text{fb}}} \sum_{i=1}^{n_{\text{fb}}} \log\left(\frac{\exp\left(\left(f_{\text{p}}^{(i)}, \tilde{f}_{\text{p}}^{+(i)}\right)/\tau\right)}{\sum_{j=1}^{n_{\text{fb}}} \exp\left(\left(f_{\text{p}}^{(i)}, \tilde{f}_{\text{p}}^{(j)}\right)/\tau\right)}\right) \tag{5}$$

where $n_{\text{fb}} = n_p - n_m$ denotes the number of patches excluding uniform proportion (UP) patches, and $\tilde{f}_{\text{p}}^{(i)}$ represents the i-th patch-level feature matched with $f_{\text{p}}^{(i)}$.

2.3 Loss function

The model uses pseudo-labels from predictions to form sample pairs, refining them by linking prediction confidence maps with pseudo-labels. This is formal-

ized as:

$$\mathcal{L}_{\text{Ment}} = -\frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \left(p_{h,w}^{T} \log(p_{h,w}^{T}) + (1 - p_{h,w}^{T}) \log(1 - p_{h,w}^{T}) \right) \quad (6)$$

For the labeled branch, we use $\mathcal{L}_{\text{Sent}}$ to measure the discrepancy between predicted probabilities $p_{h,w}^{T}$ and ground truth labels $b_{h,w}$. To improve robustness, we minimize $\mathcal{L}_{\text{Cent}}$ to reduce the distributional difference between student-generated pseudo-labels $\hat{b}_{h,w}$ and teacher predictions $p_{h,w}^{T}$, enforcing invariant semantic feature learning. Both losses are computed using cross-entropy.

Table 1. Compared with SOTA methods.

Methods	Types	Dice	HD95	Dice	HD95
		5% training dataset		**10% training dataset**	
TransNuseg[16]	Full-Sup	63.68±0.25	36.16±0.17	66.29±0.22	34.30±0.19
CausalCLIPSeg[17]		65.18±0.39	33.54±0.44	73.27±0.34	17.22±0.42
PRISM[18]		69.95±0.26	28.02±0.23	74.96±0.31	16.86±0.38
MS-Seg[19]	Semi-Sup	68.58±0.49	27.62±0.46	72.17±0.49	21.18±0.56
LVM-Med[20]		80.16±0.61	16.39±0.72	82.70±0.57	14.66±0.69
CDCL[21]		77.94±0.37	19.84±0.29	80.42±0.25	16.21±0.24
ABD-Seg[22]		79.37±0.50	17.22±0.46	82.73±0.33	14.60±0.32
MiDSS[23]		82.62±0.54	14.56±0.50	84.99±0.51	13.07±0.45
KnowSAM[24]		84.17±0.62	13.74±0.38	85.65±0.34	12.96±0.37
Ours		**86.93±0.47**	**12.15±0.37**	**89.75±0.24**	**10.8±0.32**
Methods	**Types**	**Dice**	**HD95**	**Dice**	**HD95**
		30% training dataset		**100% training dataset**	
TransNuseg	Full-Sup	72.44±0.26	20.87±0.21	84.16±0.10	13.76±0.14
CausalCLIPSeg		79.85±0.37	16.49±0.41	85.58±0.25	12.83±0.19
PRISM		80.81±0.36	16.05±0.39	88.52±0.18	11.29±0.20
MS-Seg	Semi-Sup	72.35±0.37	20.43±0.42	78.39±0.22	17.46±0.36
LVM-Med		83.42±0.40	14.34±0.45	85.26±0.26	13.05±0.30
CDCL		81.93±0.23	14.41±0.27	85.51±0.19	12.75±0.26
ABD-Seg		85.27±0.35	13.18±0.38	86.37±0.25	12.21±0.33
MiDSS		86.87±0.28	11.99±0.14	87.92±0.20	11.52±0.11
KnowSAM		87.48±0.37	11.74±0.32	87.94±0.33	11.49±0.34
Ours		**90.52±0.29**	**9.94±0.18**	**90.97±0.16**	**9.47±0.22**

3 Experiments and Results

3.1 Datasets and Experimental Settings

Datasets. The dataset includes 365 WSIs (20×) of liver biopsy tissues from the First Affiliated Hospital of Zhejiang University School of Medicine. Each WSI

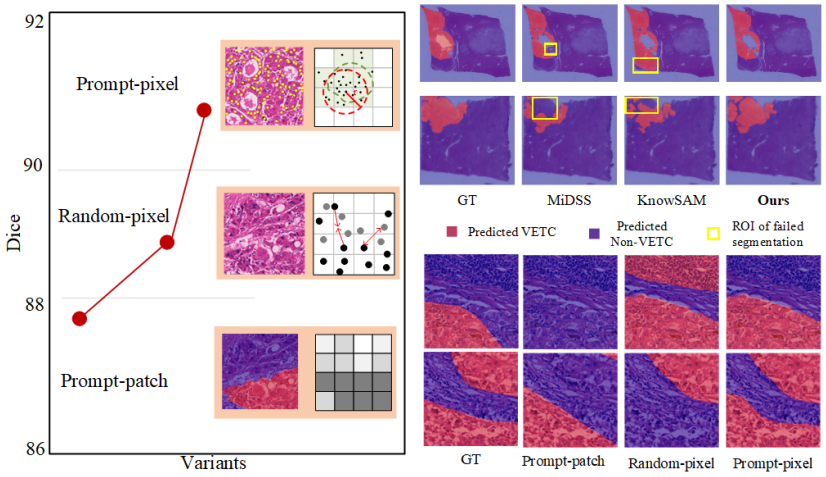

Fig. 3. Visualization results. Red regions: VETC. Blue regions: non-VETC.

was divided into 256×256 patches by a sliding window, filtering out low-quality regions [25]. A 4-fold cross-validation was conducted, and semi-supervised experiments used 5%, 10%, and 30% of training data. Evaluation metrics included the Dice Coefficient for segmentation accuracy and HD95 for boundary precision.

Implementation Details. Data augmentation included Color Jitter, Rotation (±15°), Grayscale, and Noise. DenseUNet [26] serves as the feature extractor, with a projector mapping features to $128 \times 32 \times 32$. The token-level patch size is 4×4. HoVer-Net [27] is used for nucleus segmentation, and ViT-L/16 [28] encode-decode the prompts to 10×1024. By flattening the 32×32 feature map, we concatenate each pixel's 10-dim prompt with its 128-dim morphology vector—yielding a 138-dim representation—and fuse prompts via soft pixel alignment across all 1024 positions, thereby enabling true pixel-wise supervision. Unlike position-sensitive token assignments, the prompt is task-specific yet class-agnostic. Loss weights were set as: $\omega_{\text{Sent}} = 1$, $\omega_{\text{total}} = 0.1$, $\omega_{\text{Ment}} = 0.01$, $\omega_{\text{Cent}} = 0.1$. EMA updates teacher and student models at a 0.99:0.01 ratio. Adam optimizer was used with a 1e-4 learning rate.

3.2 Comparision with SOTA Methods

As shown in Table 1, experiments encompass a wide range of training paradigms, including fully supervised, semi-supervised, vision models, language-vision prompts, SAM-based prompts. Evaluation is performed across varying labeled data levels.

VPP2P outperforms supervised methods with less training data, achieving comparable results using only 10% dataset. With 30% labeled data, it attains a Dice of 90.52%, surpassing PRISM (9.71%), LVm-med (7.1%), and KnowSAM (3.04%) by an average of 6.6%. This benefits from task-specific prompts at a finer pixel level. VPP2P reduces HD95 by 6.06% and 2.16% compared to PRISM and

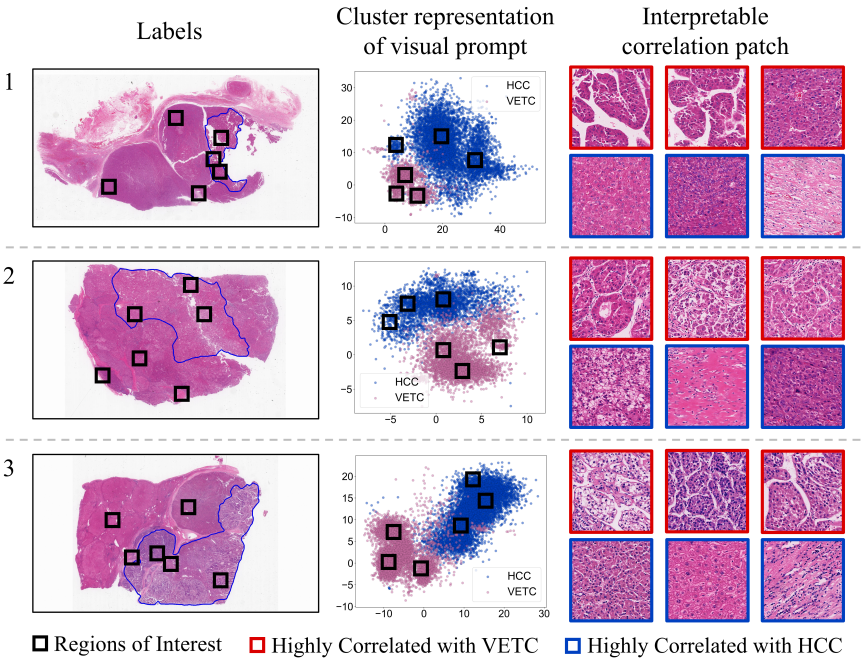

Fig. 4. Feature space analysis and model interpretability.

KnowSAM with just 10% data. As shown in Fig. 3, it enhances fine-grained segmentation, particularly excelling in edge clarity and structure preservation.

3.3 Ablation Study

As shown in Fig.3, we configured three variant methods based on prompt granularity and semantic source for analysis to validate the effectiveness of pixel-level prompting and sampling strategy. The results demonstrate that pixel-level visual prompts significantly outperform patch-level visual prompts and pixel-level random prompting. The visual prompting method successfully captures the clustered morphological features of VETC nuclei. Patch-level visual prompts fail to model the distribution characteristics of nuclei. Pixel-level random prompts are unable to capture the clustered distribution features of nuclei.

3.4 Feature spatial distribution and interpretability

Figure 4 illustrates the distribution of the patch embedding in the semantic space. It clearly delineates the boundaries between VETC and non-VETC regions, indicating that visual prompt method effectively captures the morphological characteristics of VETC. Additionally, our model selects three typical VETC patches through clustering, suggesting that morphological observations of these areas could provide valuable diagnostic references for clinicians.

4 Conclusion

This paper introduces a novel patch-to-pixel visual prompt framework that leverages pixel-level nucleus prior knowledge to enhance VETC segmentation. Contrastive learning structure propagates domain-specific visual prompts from patch to pixel level. Experiments validate the effectiveness, demonstrating superior performance across various benchmarks. Future work will evaluate diverse clinical cohorts to assess generalization, as well as explore the influence of various external segmentation networks on pixel-level feature propagation.

Acknowledgments. This work is supported by the Zhejiang Provincial Natural Science Foundation (LZ23H180002 and LQ23F030001), the Zhejiang Province's Vanguard Geese Leading Plan Project (2024C03067 and 2024C03056), the Key projects for agriculture and social development in Hangzhou (20231203A13), the Cao Guangbiao High-tech Development Fund (2022RC009), the National Key Research and Development Program of China (2021YFF0700305), the Key Projects of Hangzhou Science and Technology Bureau (20231203A13), and the National Natural Science Foundation of China (62406280).

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

References

1. Shams, M.Y., El-Kenawy, E.S.M., Ibrahim, A., Elshewey, A.M.: A hybrid dipper throated optimization algorithm and particle swarm optimization (dtpso) model for hepatocellular carcinoma (hcc) prediction. Biomedical Signal Processing and Control **85**, 104908 (2023)
2. Liu, K., Dennis, C., Prince, D.S., Marsh-Wakefield, F., Santhakumar, C., Gamble, J.R., Strasser, S.I., McCaughan, G.W.: Vessels that encapsulate tumour clusters vascular pattern in hepatocellular carcinoma. JHEP Reports **5**(8), 100792 (2023)
3. Yu, J., Ma, T., Hua, D., Chen, F., Fu, J., Xu, Y.: Semi-supervised instance segmentation in whole slide images via dense spatial variability enhancing. IEEE Journal of Biomedical and Health Informatics (2024)
4. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023)
5. Ju, L., Wu, Y., Feng, W., Yu, Z., Wang, L., Zhu, Z., Ge, Z.: Universal semi-supervised learning for medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 355–365. Springer (2024)
6. Liu, J., Qian, W., Cao, J., Liu, P.: Overlay mantle-free for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 589–598. Springer (2024)
7. Javed, S., Mahmood, A., Qaiser, T., Werghi, N., Rajpoot, N.: Unsupervised mutual transformer learning for multi-gigapixel whole slide image classification. Medical Image Analysis **96**, 103203 (2024)

8. Ayromlou, S., Tsang, T., Abolmaesumi, P., Li, X.: Ccsi: Continual class-specific impression for data-free class incremental learning. Medical Image Analysis p. 103239 (2024)
9. Gao, H., Wang, H., Chen, L., Cao, X., Zhu, M., Xu, P.: Semi-supervised segmentation of hyperspectral pathological imagery based on shape priors and contrastive learning. Biomedical Signal Processing and Control **91**, 105881 (2024)
10. Wu, X., Xu, Z., Tong, R.K.y.: Few slices suffice: Multi-faceted consistency learning with active cross-annotation for barely-supervised 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 286–296. Springer (2024)
11. Cho, H., Han, Y., Kim, W.H.: Anti-adversarial consistency regularization for data augmentation: Applications to robust medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 555–566. Springer (2023)
12. Yu, J., Ma, T., Chen, H., Lai, M., Ju, Z., Xu, Y.: Marrying global–local spatial context for image patches in computer-aided assessment. IEEE Transactions on Systems, Man, and Cybernetics: Systems (2023)
13. Yu, J., Wang, X., Ma, T., Li, X., Xu, Y.: Patch-slide discriminative joint learning for weakly-supervised whole slide image representation and classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 713–722. Springer (2024)
14. Yin, C., Liu, S., Zhou, K., Wong, V.W.S., Yuen, P.C.: Prompting vision foundation models for pathology image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11292–11301 (June 2024)
15. Renne, S.L., Woo, H.Y., Allegra, S., Rudini, N., Yano, H., Donadon, M., Viganò, L., Akiba, J., Lee, H.S., Rhee, H., et al.: Vessels encapsulating tumor clusters (vetc) is a powerful predictor of aggressive hepatocellular carcinoma. Hepatology **71**(1), 183–195 (2020)
16. He, Z., Unberath, M., Ke, J., Shen, Y.: Transnuseg: A lightweight multi-task transformer for nuclei segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 206–215. Springer (2023)
17. Chen, Y., Wei, M., Zheng, Z., Hu, J., Shi, Y., Xiong, S., Zhu, X.X., Mou, L.: Causalclipseg: Unlocking clip's potential in referring medical image segmentation with causal intervention. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 77–87. Springer (2024)
18. Li, H., Liu, H., Hu, D., Wang, J., Oguz, I.: Prism: A promptable and robust interactive segmentation model with visual prompts. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 389–399. Springer (2024)
19. Dai, D., Dong, C., Xu, S., Yan, Q., Li, Z., Zhang, C., Luo, N.: Ms red: A novel multi-scale residual encoding and decoding network for skin lesion segmentation. Medical image analysis **75**, 102293 (2022)
20. MH Nguyen, D., Nguyen, H., Diep, N., Pham, T.N., Cao, T., Nguyen, B., Swoboda, P., Ho, N., Albarqouni, S., Xie, P., et al.: Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching. Advances in Neural Information Processing Systems **36** (2024)
21. Wu, H., Wang, Z., Song, Y., Yang, L., Qin, J.: Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11666–11675 (2022)

22. Chi, H., Pang, J., Zhang, B., Liu, W.: Adaptive bidirectional displacement for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4070–4080 (2024)
23. Ma, Q., Zhang, J., Qi, L., Yu, Q., Shi, Y., Gao, Y.: Constructing and exploring intermediate domains in mixed domain semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11642–11651 (2024)
24. 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 (2025)
25. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
26. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE transactions on medical imaging **37**(12), 2663–2674 (2018)
27. Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N.: Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical image analysis **58**, 101563 (2019)
28. Chen, R.J., Krishnan, R.G.: Self-supervised vision transformers learn visual concepts in histopathology. arXiv preprint arXiv:2203.00585 (2022)