

HomoSeg: A Weak-Label-Guided Semi-Supervised Framework for Tumor Microenvironment Segmentation

Haoyun Zhao¹, Jun Ni^{1,*}, Yibing Zhan², Chongchong Qi⁴, Yuhe Zhu¹, Dapeng Tao^{1,3,4}

¹Yunnan University

²Wuhan University

³Northwestern Polytechnical University

⁴Yunnan United Vision Technology Company Ltd.

*jun.ni@ynu.edu.cn

Abstract. Accurate tumor microenvironment (TME) analysis is critical for prognosis but limited by the high cost of pixel-level annotations. Current weakly supervised methods suffer from localization bias, and although semi-supervised segmentation yields superior boundaries, standard frameworks depend on pixel-level annotations. To resolve this, we propose Homogeneity driven Segmentation, namely HomoSeg, a weak-label-guided semi-supervised framework. It exploits histological spatial homogeneity to extract single-label samples. Superpixel segmentation is utilized to synthesize training samples with dense and reliable pixel-level labels. Subsequently, we design a multi-task Siamese network integrating classification, segmentation, and CAM branches. Through weak-to-strong consistency regularization, supervision signals are propagated from synthetic data to unlabeled data, establishing robust semantic decision boundaries. Finally, to eliminate noise from synthetic data and achieve model lightweighting, we employ an offline distillation strategy to distill the multi-task Siamese network. Experimental results on the LUAD-HistoSeg and BCSS-WSSS datasets demonstrate that HomoSeg significantly outperforms existing methods, providing a robust solution for high-precision Tumor Microenvironment analysis. The code is available at: <https://github.com/zhy1312/HomoSeg>.

Keywords: Weakly Supervised Segmentation · Tumor Microenvironment · Semi-Supervised Segmentation · Knowledge Distillation.

1 Introduction

Precise tumor microenvironment (TME) analysis is critical for cancer diagnosis, prognosis, and treatment planning [8]. Although deep learning has advanced medical image segmentation [11, 15, 21], mainstream fully supervised methods rely heavily on extensive pixel-level annotations. Delineating complex TME

structures on gigapixel whole slide images is time-consuming and requires specialized pathological expertise, making prohibitive annotation costs a key bottleneck for large-scale clinical deployment of deep learning models [7].

As a promising solution, weakly supervised semantic segmentation (WSSS) using only image-level labels has emerged [2, 16]. Most WSSS methods adopt a classify-then-activate paradigm, generating pseudo-masks via class activation map (CAM) [25] for segmentation training. However, CAM-based methods activate only the most discriminative regions, leading to struggles with complex tissue boundaries and a failure to produce clinically precise segmentations.

Semi-supervised semantic segmentation (SSSS) is a compelling alternative to mitigate these shortcomings, including severe localization bias and blurred boundaries [18, 14]. Mainstream SSSS frameworks require high-quality pixel-level annotations for model initialization or supervision. This creates a paradox: SSSS offers superior performance but relies on costly annotations, rendering it inapplicable for strict weakly supervised scenarios without pixel-level labels.

To resolve the conflict between performance demands and annotation costs, we propose a novel weak-label-guided semi-supervised framework. To break the trade-off, we reformulate the task by moving beyond traditional annotation-reliant semi-supervised strategies. Instead, our approach exploits latent consistency supervisory signals using only image-level labels, which enables us to guarantee segmentation accuracy while drastically reducing the annotation burden.

2 Method

As illustrated in Fig. 1, the proposed HomoSeg framework consists of three key components: (i) Homogeneity Driven Data Reformulation, (ii) a Weak-Label Guided Semi-Supervised Framework, and (iii) Distillation for Boundary Refinement. Each of these components is detailed in the following sections.

2.1 Homogeneity Driven Data Reformulation

Unlike natural scenes, TME patches are from gigapixel-level WSIs. Due to the spatial continuity and large scale of biological structures, cropped local patches typically contain only one tissue category. This morphological feature is the fundamental physical premise of our proposed data reformulation strategy.

Single-Label Sample Mining Let the TME dataset be $D = \{(x_i, y_i)\}_{i=1}^N$, with N as the total number of training samples. For each sample, $x_i \in \mathbb{R}^{3 \times H \times W}$ is the input image, and $y_i \in \{0, 1\}^C$ is the multi-hot image-level label for C tissue categories. Given the large biological structure scale relative to patch size, WSI-cropped images often contain only one tissue type, making their weak image-level labels equivalent to global pixel-level masks. The single-label subset is defined:

$$S = \{(x_s, y_s) \in D \mid \|y_s\|_1 = 1\} \quad (1)$$

For $(x_s, y_s) \in S$, the mask m_s is derived by spatially broadcasting y_s to all pixel positions, providing reliable initial supervision.

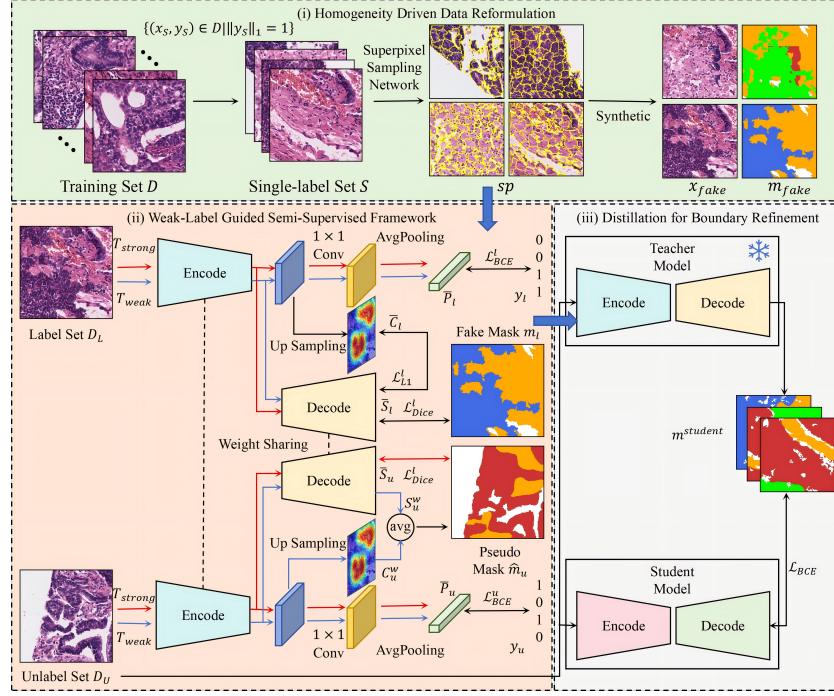


Fig. 1. Overview of the proposed HomoSeg framework.

Topology-Aware Synthetic Generation Relying solely on single-label samples fails to capture complex tissue boundaries. To resolve this, a superpixel-based [23] CutMix strategy is proposed to synthesize multi-label samples. Specifically, we apply superpixel sampling networks to segment images x_s into irregular, morphologically realistic regions. Aggregating connected superpixels generates binary region masks M_{sp} . We then synthesize multi-label data via iterative superposition. For each synthetic sample, we randomly select k distinct tissue categories ($k \in \{2, \dots, C\}$). Starting with a randomly sampled background image x_{bg} , we overlay $k-1$ foreground tissue regions from the remaining categories. Let $T_j(\cdot)$ denote a random geometric transformation. The update rule for the synthetic image x_{fake} and its mask m_{fake} at the j -th iteration is defined as:

$$x_{fake}^{(j)} = M_{aug}^{(j)} \odot T_j(x_j) + (\mathbf{1} - M_{aug}^{(j)}) \odot x_{fake}^{(j-1)} \quad (2)$$

$$m_{fake}^{(j)}[c] = \begin{cases} M_{aug}^{(j)} & \text{if } c = c_j \\ m_{fake}^{(j-1)}[c] \odot (\mathbf{1} - M_{aug}^{(j)}) & \text{if } c \neq c_j \end{cases} \quad (3)$$

where x_j denotes the j -th foreground image belonging to tissue class c_j , $\odot$ denotes element-wise multiplication, and $\mathbf{1} \in \mathbb{R}^{H \times W}$ denotes an all-ones matrix. The initial states are defined as $x_{fake}^{(0)} = x_{bg}$ and $m_{fake}^{(0)}$ as the broadcasted mask of the background class.

Consequently, the original dataset is reformulated into two subsets: the labeled set $D_L = S \cup \{(x_{fake}, m_{fake})\}$ containing samples with pixel-level ground truth, and the unlabeled set $D_U = D$ consisting of the original samples. To ensure balanced supervision during the training phase, we maintain an equal number of samples in both subsets.

2.2 Weak-Label Guided Semi-Supervised Framework

Multi-Task Siamese Network and Perturbation Strategy To effectively leverage the mixed dataset, we adopt a FixMatch-based [12] framework with a shared-weight Siamese network to process labeled and unlabeled samples in parallel. The ResNeSt200 [17] based encoder $\Phi_{enc}(\cdot)$ first extracts high-level features f from the input images. To fully exploit supervision from both pixel-level masks and image-level labels, these features are subsequently fed into three parallel branches: a PSPNet [22] segmentation decoder $\Phi_{dec}(\cdot)$, a classification head $\Phi_{cls}(\cdot)$, and an auxiliary CAM branch $\Phi_{cam}(\cdot)$.

We also define a hierarchical perturbation strategy: weak augmentation $\mathcal{T}_{weak}(\cdot)$ preserves semantic structure for reliable pseudo-labels, while strong augmentation $\mathcal{T}_{strong}(\cdot)$ simulates staining heterogeneity and enhances model robustness.

Establishing Semantic Boundaries with Labeled Data Within our semi-supervised pipeline, effective supervision on the labeled subset is prerequisite to establish precise semantic decision boundaries. This serves as a robust semantic anchor, ensuring that the network can subsequently distinguish between tissues before propagating knowledge to unlabeled data.

Formally, for a labeled sample $(x_l, y_l, m_l) \in D_L$, where x_l is the image, y_l is the label vector, and m_l is the multi-channel mask, we employ a predictive averaging strategy to mitigate stochastic variance. The image x_l is first transformed into a weak view x_l^w and a strong view x_l^s . Both views x_l^k ($k \in \{w, s\}$) are forwarded through the shared encoder Φ_{enc} to extract high-level features $f_l^k = \Phi_{enc}(x_l^k)$. These features are then processed by the three task-specific branches:

$$S_l^k = \Phi_{dec}(f_l^k), \quad P_l^k = \Phi_{cls}(f_l^k), \quad C_l^k = \Phi_{cam}(f_l^k) \quad (4)$$

where S_l^k and C_l^k represent the segmentation logits and CAM activations for labeled samples, respectively, and $P_l^k \in \mathbb{R}^C$ is the classification logit vector.

To minimize the prediction noise derived from distinct perturbations, we compute the ensemble predictions by averaging the outputs from both views:

$$\bar{S}_l = \frac{S_l^w + S_l^s}{2}, \quad \bar{P}_l = \frac{P_l^w + P_l^s}{2}, \quad \bar{C}_l = \frac{C_l^w + C_l^s}{2} \quad (5)$$

Based on these smoothed predictions, the supervised objective $\mathcal{L}_{labeled}$ is formulated to strictly align the model with the ground truth:

$$\mathcal{L}_{labeled} = \mathcal{L}_{Dice}(\bar{S}_l, m_l) + \mathcal{L}_{BCE}(\bar{P}_l, y_l) + \mathcal{L}_{L1}(\bar{C}_l, \bar{S}_l) \quad (6)$$

The first two terms ensure pixel-wise and image-level accuracy, respectively. Crucially, the third term forces the CAM activation to conform to the precise segmentation boundaries, preventing the classification branch from focusing solely on discriminative object parts.

Unlabeled Consistency via CAM-Calibrated Pseudo-Labeling Parallel to the labeled stream, we employ a weak-to-strong consistency paradigm for unlabeled samples $(x_u, y_u) \in D_U$. Undergoing an identical multi-branch feed-forward process as the labeled data, the input views yield S_u^k , P_u^k and C_u^k .

However, the segmentation branch often yields uncertain or incomplete predictions during early training stages. To mitigate this instability, we exploit the complementary discriminative information from the CAM branch. By fusing these reliable localization cues with the weak-view segmentation predictions, we generate a calibrated pseudo-mask $\hat{m}_u$:

$$\hat{m}_u = \arg \max \left(\frac{S_u^w + C_u^w}{2} \right) \quad (7)$$

Subsequently, this calibrated mask serves as the ground truth to supervise the strong augmentation view S_u^s . Additionally, consistent with the labeled stream, we employ the predictive averaging strategy for the classification branch to compute the smoothed probability $\bar{P}_u = (P_u^w + P_u^s)/2$. The unsupervised objective is formulated as:

$$\mathcal{L}_{unlabeled} = \mathcal{L}_{Dice}(S_u^s, \hat{m}_u) + \mathcal{L}_{BCE}(\bar{P}_u, y_u) \quad (8)$$

Finally, the network is optimized end-to-end. The total loss function $\mathcal{L}_{total}$ integrates both labeled and unlabeled flows via direct summation:

$$\mathcal{L}_{total} = \underbrace{(\mathcal{L}_{Dice}^l + \mathcal{L}_{BCE}^l + \mathcal{L}_{L1}^l)}_{\text{Labeled Flow}} + \underbrace{(\mathcal{L}_{Dice}^u + \mathcal{L}_{BCE}^u)}_{\text{Unlabeled Flow}} \quad (9)$$

This joint optimization establishes precise feature boundaries via labeled data while enhancing robustness against perturbations via unlabeled consistency.

2.3 Distillation for Boundary Refinement

The preceding stage may still be affected by synthetic noise introduced during data generation and the computational overhead of iterative refinement [5]. To alleviate these limitations, we introduce an offline distillation strategy [9].

We leverage the converged semi-supervised network as a Teacher to generate a static, high-quality dataset $D_{student}$ from the original dataset $D = \{(x_i, y_i)\}_{i=1}^N$. For each sample, we compute the ensemble logits $O_i = (S_i + C_i)/2$ to integrate fine-grained boundary sensitivity with discriminative localization. Furthermore, predictions for categories absent in the image-level label y_i are explicitly masked.

Formally, the pixel-wise target mask $m_i^{student}$ is constructed using a conditional strategy. To strictly adhere to the tissue homogeneity prior, single-label

samples are assigned dense ground truths by directly broadcasting their global category index y_i , ensuring absolute supervision accuracy. Conversely, for multi-label samples ($x_i \in D \setminus S$), $m_i^{student}$ is generated by taking the arg max over the masked teacher predictions ($O_i \cdot y_i^{mask}$).

Upon constructing the distilled dataset $D_{student} = \{(x_i, m_i^{student})\}_{i=1}^N$, an efficient student model is trained. The student retains only the segmentation branch while adopting the same backbone and decoder architecture as the teacher. Following common practice, the encoder is initialized with ImageNet-pretrained weights and optimized exclusively using the Dice loss on $D_{student}$:

$$\mathcal{L}_{student} = \mathcal{L}_{Dice}(S_{student}, m^{student}) \quad (10)$$

This distillation process effectively filters out noise arising from synthetic samples in D_L and reduces model complexity.

3 Experiments

3.1 Setting

Dataset Experiments were conducted on two public histopathological datasets. LUAD-HistoSeg [7], collected from Guangdong Provincial People’s Hospital, contains 16,678 training patches, 300 validation patches, and 307 test patches. BCSS-WSSS [1], released by the British Columbia Cancer Agency, is partitioned into 23,422 training patches, 3,418 validation patches, and 4,986 test patches. For both datasets, only image-level labels are available for training, while pixel-level annotations are provided for validation and testing. Category distributions are summarized in Table 1.

Table 1. Category distributions and patch-type statistics in the training sets.

LUAD-HistoSeg		BCSS-WSSS	
Category	Proportion	Category	Proportion
Tumor Epithelial (TE)	78.85%	Tumor (TUM)	65.03%
Tumor-associated Stroma (TAS)	80.60%	Stroma (STR)	69.40%
Lymphocyte (LYM)	4.18%	Lymphocytic Infiltrate (LYM)	23.90%
Necrosis (NEC)	15.76%	Necrosis (NEC)	12.31%
Single-label patches	27.55%	Single-label patches	40.04%

Implementation Details We implemented the framework using Python 3.8.20 and PyTorch 2.4.0. The Adam optimizer was employed with a weight decay of 1×10^{-4} . Specifically, the learning rate was set to 1×10^{-4} for the teacher model and 1×10^{-5} for the student model. Training was conducted with a batch size of 16 for up to 100 epochs. Performance was evaluated using mean intersection over union (mIoU) and frequency-weighted intersection over union (fwIoU), which

Table 2. Comparison with state-of-the-art methods. Best results are highlighted in bold, and second-best results are underlined. (%)

Method	Venue	LUAD-HistoSeg Dataset						BCSS-WSSS Dataset					
		TE	NEC	LYM	TAS	mIoU	fwIoU	TUM	STR	LYM	NEC	mIoU	fwIoU
HistoSegNet [2]	ICCV'19	19.22	20.53	44.43	44.49	32.17	32.07	55.23	59.53	30.14	58.14	50.76	54.71
SEAM [13]	CVPR'20	60.23	54.88	57.85	62.88	58.96	60.50	78.53	70.92	56.24	61.82	66.88	72.36
SC-CAM [3]	CVPR'20	72.80	75.04	72.45	68.88	72.26	71.41	79.54	73.53	59.09	65.07	69.31	74.33
MLPS [7]	MIA'22	77.13	76.26	72.71	71.42	74.38	74.36	77.94	72.34	61.22	64.24	68.93	73.31
OEEM [10]	MICCAI'22	72.64	65.34	62.35	67.74	67.02	68.90	78.95	73.64	59.63	64.54	69.19	74.16
HAMIL [24]	TMI'23	73.10	63.93	71.39	68.96	67.02	68.90	73.68	69.64	57.61	57.58	64.63	69.69
TPRO [20]	MICCAI'23	77.23	77.15	75.28	71.40	75.26	75.57	76.68	65.83	53.88	63.43	64.96	64.48
PistoSeg [4]	AAAI'23	78.01	72.19	70.46	69.75	72.60	73.50	80.91	73.92	59.53	64.36	69.68	75.10
ARML [6]	MICCAI'24	77.82	81.94	75.32	72.42	76.88	75.79	78.08	72.82	57.82	66.02	68.68	73.31
HisynSeg [5]	TMI'25	78.27	77.23	77.00	72.26	76.19	75.79	81.78	74.51	60.36	67.25	70.97	75.93
ESFAN [19]	MICCAI'25	79.30	<u>81.02</u>	78.60	74.41	78.33	77.52	78.02	73.18	<u>61.80</u>	65.96	69.74	73.83
HomoSeg(teacher)		79.88	80.09	76.59	73.49	77.51	77.09	80.57	73.93	61.47	67.40	70.84	75.28
HomoSeg(student)		81.39	83.98	<u>77.32</u>	<u>74.29</u>	79.25	78.41	82.03	75.57	62.26	70.78	72.66	76.83

reflect category-specific and overall segmentation accuracy while accounting for class imbalances in pathological images.

3.2 Comparison with State-of-the-art Methods

We compared our method with several state-of-the-art methods on the LUAD-HistoSeg and BCSS-WSSS datasets. As shown in Table 2, the IoU results for each category, along with mIoU and fwIoU, demonstrate that HomoSeg improves mIoU and fwIoU by 0.92% and 0.89%, respectively, over the next best method on the LUAD-HistoSeg dataset. On the BCSS-WSSS dataset, HomoSeg outperforms the second-best method by 1.69% in mIoU and 0.9% in fwIoU. Notably, even the pseudo-labels generated by the first-stage teacher model achieve performance comparable to the second-best state-of-the-art method. This performance gain is attributed to HomoSeg’s incorporation of a multi-task semi-supervised learning strategy, which effectively utilizes information from single-label samples and image-level labels. This strategy significantly enhances the model’s ability to recognize complex boundaries and even small objects, particularly in the fine structures of the tumor microenvironment. A more intuitive qualitative analysis is shown in Fig. 2.

3.3 Ablation Study

To evaluate the contributions of each component in the HomoSeg framework, we performed an ablation study on the LUAD-HistoSeg and BCSS-WSSS datasets.

As detailed in Table 3, we validated each component by removing the Homogeneity Driven Data Reformulation and relying solely on single-label samples. We also evaluated the impact of weak/strong augmentations and individual loss terms within the multi-task framework. Furthermore, Table 4 shows that offline distillation significantly reduces training time and memory usage while enhancing accuracy. By eliminating redundant auxiliary branches and synthetic noise,

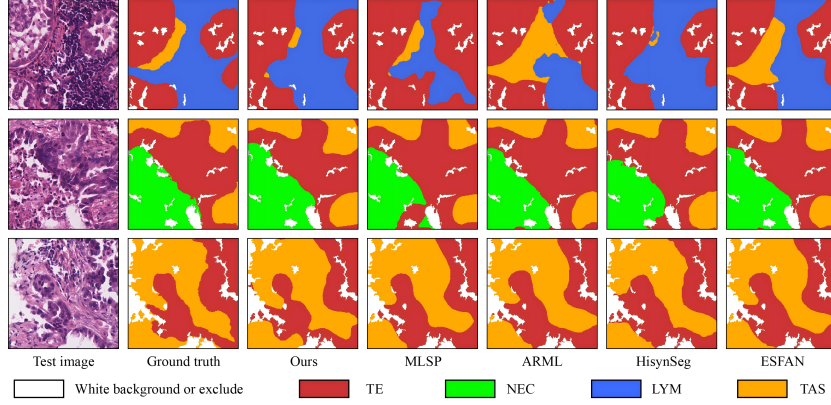


Fig. 2. Visualization of tissue segmentation on LUAD-HistoSeg Dataset.

the lightweight student model effectively focuses on stable, discriminative features, achieving a superior accuracy-efficiency trade-off.

Table 3. Ablation study on the contribution of each component in HomoSeg (%)

x_{fake}	$\mathcal{T}_{weak}$	$\mathcal{T}_{strong}$	$\mathcal{L}_{Dice}$	$\mathcal{L}_{BCE}$	$\mathcal{L}_{L1}$	LUAD-HistoSeg		BCSS-WSSS	
						mIoU	fwIoU	mIoU	fwIoU
	✓	✓	✓	✓	✓	78.46	78.02	69.08	73.67
✓		✓	✓	✓	✓	76.69	76.88	70.86	75.13
✓	✓		✓	✓	✓	73.27	73.00	70.94	74.91
✓	✓	✓	✓			72.48	74.45	70.81	75.10
✓	✓	✓	✓	✓		74.04	75.16	71.27	75.70
✓	✓	✓	✓	✓	✓	79.25	78.41	72.66	76.83

4 Conclusion

Current weakly supervised TME segmentation methods frequently suffer from localization bias and blurred boundaries. To address these challenges, we propose HomoSeg, a weak-label-guided semi-supervised framework. By exploiting spatial homogeneity for single-label mining and employing a multi-task Siamese network, we establish precise semantic boundaries without manual pixel-level annotations. Furthermore, our boundary-refined distillation strategy ensures computational efficiency. Extensive experiments on two public datasets demonstrate that HomoSeg significantly outperforms state-of-the-art methods, offering a robust and efficient solution for clinical TME analysis.

Table 4. Computational efficiency comparison on LUAD-HistoSeg with/without boundary-refined distillation.

Method	Training(s/epoch)	GPU Memory(GB)	Inference(ms/sample)
w/o Distillation	852	17.89	6.75
with Distillation	217	5.74	6.75

Acknowledgments. This work was supported in part by the Yunnan Provincial Key Science and Technology Project (Nos. 202502AS080002, 202402AA310016), Yunnan Yunling Scholar Program (No. K264202230207), National Natural Science Foundation of China (No. 62301456), Kunming University of Science and Technology Medical Joint Project (No. KUST-KH2023026Y), Zhejiang Leading Goose Program (No. 2025CO1129) and Zhejiang Healthcare Major Science and Technology Project (No. WKJ-ZJ-26004).

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

References

1. Amgad, M., Elfandy, H., Hussein, H., Atteya, L.A., Elsebaie, M.A., Abo Elnasr, L.S., Sakr, R.A., Salem, H.S., Ismail, A.F., Saad, A.M., et al.: Structured crowd-sourcing enables convolutional segmentation of histology images. *Bioinformatics* **35**(18), 3461–3467 (2019)
2. Chan, L., Hosseini, M.S., Rowsell, C., Plataniotis, K.N., Damaskinos, S.: Histoseg-net: Semantic segmentation of histological tissue type in whole slide images. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 10662–10671 (2019)
3. Chang, Y.T., Wang, Q., Hung, W.C., Piramuthu, R., Tsai, Y.H., Yang, M.H.: Weakly-supervised semantic segmentation via sub-category exploration. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 8991–9000 (2020)
4. Fang, Z., Chen, Y., Wang, Y., Wang, Z., Ji, X., Zhang, Y.: Weakly-supervised semantic segmentation for histopathology images based on dataset synthesis and feature consistency constraint. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 37, pp. 606–613 (2023)
5. Fang, Z., Wang, Y., Xie, P., Wang, Z., Zhang, Y.: Hisynseg: Weakly-supervised histopathological image segmentation via image-mixing synthesis and consistency regularization. *IEEE Transactions on Medical Imaging* **44**(4), 1765–1782 (2025). <https://doi.org/10.1109/TMI.2024.3520129>
6. Feng, S., Chen, J., Liu, Z., Liu, W., Wang, Z., Lan, R., Pan, X.: Mining gold from the sand: Weakly supervised histological tissue segmentation with activation relocation and mutual learning. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 414–423. Springer (2024)
7. Han, C., Lin, J., Mai, J., Wang, Y., Zhang, Q., Zhao, B., Chen, X., Pan, X., Shi, Z., Xu, Z., et al.: Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. *Medical Image Analysis* **80**, 102487 (2022)

8. Kang, J., Lee, J.H., Cha, H., An, J., Kwon, J., Lee, S., Kim, S., Baykan, M.Y., Kim, S.Y., An, D., et al.: Systematic dissection of tumor-normal single-cell ecosystems across a thousand tumors of 30 cancer types. *Nature communications* **15**(1), 4067 (2024)
9. Lei, S., Zhang, S., Tao, D.: Offline behavior distillation. *Advances in Neural Information Processing Systems* **37**, 115098–115118 (2024)
10. Li, Y., Yu, Y., Zou, Y., Xiang, T., Li, X.: Online easy example mining for weakly-supervised gland segmentation from histology images. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 578–587. Springer (2022)
11. Liu, X., Gao, P., Yu, T., Wang, F., Yuan, R.Y.: Cswin-unet: Transformer unet with cross-shaped windows for medical image segmentation. *Information Fusion* **113**, 102634 (2025)
12. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems* **33**, 596–608 (2020)
13. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 12275–12284 (2020)
14. Xiao, Y., Chen, C., Wang, L., Yu, J., Chen, B., Fu, X., Lu, M.: A novel semi-supervised domain adaptive method for cross-modality medical image segmentation. *The Visual Computer* **41**(11), 9139–9154 (2025)
15. Xu, S., Chen, Y., Zhang, X., Sun, F., Chen, S., Ou, Y., Luo, C.: X-unet: a novel global context-aware collaborative fusion u-shaped network with progressive feature fusion of codec for medical image segmentation. *Neural Networks* **192**, 107943 (2025). <https://doi.org/https://doi.org/10.1016/j.neunet.2025.107943>, <https://www.sciencedirect.com/science/article/pii/S089360802500824X>
16. Yi, J., Bi, Q., Zheng, H., Zhan, H., Ji, W., Huang, H., Li, Y., Li, S., Wu, X., Zheng, Y., et al.: D-cam: Learning generalizable weakly-supervised medical image segmentation from domain-invariant cam. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 122–132. Springer (2025)
17. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., et al.: Resnest: Split-attention networks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 2736–2746 (2022)
18. Zhang, H., Ma, P., Liu, J., Lian, J., Ma, Y.: Prototype-augmented mean teacher for robust semi-supervised medical image segmentation. *Pattern Recognition* **166**, 111722 (2025)
19. Zhang, H., Feng, S., Huan, Z., Wang, H., Liu, Z., Lan, R., Pan, X.: Edge-semantic synergy fusion and adaptive noise-aware for weakly supervised pathological tissue segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 160–169. Springer (2025)
20. Zhang, S., Zhang, J., Xie, Y., Xia, Y.: Tpro: Text-prompting-based weakly supervised histopathology tissue segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 109–118. Springer (2023)
21. Zhao, H., Tao, D., Zhan, Y., Ni, J., Chen, Y.: Cpgnet: Multimodal graph learning with hierarchical category guidance for multi-label whole slide image classification. *IEEE Journal of Biomedical and Health Informatics* (2025)

22. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)
23. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2914–2923 (2017)
24. Zhong, L., Wang, G., Liao, X., Zhang, S.: Hamil: High-resolution activation maps and interleaved learning for weakly supervised segmentation of histopathological images. *IEEE Transactions on Medical Imaging* **42**(10), 2912–2923 (2023)
25. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)