Hierarchical Prompt Contrastive Learning for Weakly Supervised Histopathology Segmentation

Zengmao Wang
School of Computer Science, Wuhan University
Wuhan, Hubei, China
wangzengmao@whu.edu.cn

Ming Hu
School of Computer Science, Wuhan University
Wuhan, Hubei, China
minghu@whu.edu.cn

Chaoyang Zhou
School of Computer Science, Wuhan University
Wuhan, Hubei, China
zhoucy@whu.edu.cn

Shiying Sun
Institute of Automation, Chinese Academy of Sciences
Beijing, China
sunshiying2013@ia.ac.cn

Qingping Zheng
School of Computer Science, Zhejiang University
Hangzhou, Zhejiang, China
zhengqingping2024@zju.edu.cn

Lefei Zhang
School of Computer Science, Wuhan University
Wuhan, Hubei, China
zhanglefei@whu.edu.cn

Yongchao Xu
School of Computer Science, Wuhan University
Wuhan, Hubei, China
yongchao.xu@whu.edu.cn

Bo Du
School of Computer Science, Wuhan University
Wuhan, Hubei, China
dubo@whu.edu.cn

Abstract

Weakly supervised histopathology segmentation is a hot topic in the computer vision community by generating pixel-level masks with image-level labels due to the expensive and time-intensive pixel-level annotations. The inter-class homogeneity and intra-class heterogeneity are challenging issues for histopathological image segmentation. Recent works addressing these issues mainly focus on learning intra-class similarity, resulting in the spurious features of each category being captured and the discriminative ability of feature representation for each category being decreased. In this paper, we propose a hierarchical prompt contrastive method for histopathological image segmentation with hierarchical text prompt contrastive loss and hierarchical class prompt contrastive loss. Specifically, for hierarchical text prompt contrastive loss, we align the features of different layers from the image encoder to the adaptive text prompt for each category by adding linear layers following a frozen text encoder. Then we can generate several classification activation maps to produce high-confidence pseudo segmentation masks, and they are aligned to text features for contrastive learning with the foundation model to generate more confident segmentation masks. While in hierarchical class prompt contrastive loss, the subclass prompts are obtained by dividing each category into several clusters. Then an intra-class contrastive strategy and an inter-class contrastive strategy are developed by aligning the subclass prompt to the image. These contrastive losses are adopted to guarantee that the structures within each category are preserved and the discriminative ability of different categories is improved. Extensive experiments show that the proposed method outperforms the state-of-the-art methods on the popular benchmark datasets.

This work is licensed under a Creative Commons Attribution 4.0 International License.
WWW Companion '26, Dubai, United Arab Emirates.
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2308-7/2026/04
https://doi.org/10.1145/XXXXXX.XXXXXX

CCS Concepts

• **Computing methodologies → Image segmentation**; • **Applied computing → Health informatics**.

Keywords

Weakly supervised semantic segmentation, Contrastive learning, Histopathological image segmentation

ACM Reference Format:
Zengmao Wang, Ming Hu, Chaoyang Zhou, Shiying Sun, Qingping Zheng, Lefei Zhang, Yongchao Xu, and Bo Du. 2026. Hierarchical Prompt Contrastive Learning for Weakly Supervised Histopathology Segmentation. In *Companion Proceedings of the ACM Web Conference 2026 (WWW Companion '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

1 Introduction

Semantic segmentation of histopathological images is essential for computational pathology[13]. It enables the quantitative analysis of tumor microenvironments to support cancer grading, prognosis evaluation, and therapy assessment[1, 10]. Generally, it is necessary to collect a large amount of pixel-level annotations to train the deep network for precise segmentation[11, 27, 50]. However, this process is expensive and time-consuming due to the requirements of accurate boundaries annotation for complex structures in the image. The annotation challenges promote the development of label-efficient learning approaches for segmentation[17].

Weakly supervised semantic segmentation (WSSS) is the popular approach by utilizing image-level labels to generate pixel-level predictions[38]. The core technique of the weakly supervised methods for segmentation is to employ Class Activation Maps (CAMs)

to produce pseudo-masks for training a segmentation model[47]. However, standard activation maps typically highlight only the most discriminative regions required for classification and overlook less prominent yet important parts of the object[22]. These issues are amplified in histopathology, where tissue structures often lack clear contours and exhibit low contrast between classes. Meanwhile, since histopathological images often contain multiple labels, training a model to classify an image into several categories tends to make it focus excessively on the most discriminative regions for each class. These characteristics of histopathology images make it difficult for conventional activation map methods to capture the entire tissue structure.

Recently, many works have been developed by injecting textual or visual priors to address the intrinsic shortcomings of CAM[47]. Text-based methods introduce language knowledge to align high-level semantic with pixel-level details thereby guiding the model to attend to broader tissue structures[45]. While visual-based methods utilize class prototypes to capture intra-class variations and mitigate inter-class similarities[31]. In these methods, they all almost focus on improving the intra-class similarity and extending the inter-class distinctiveness to alleviate the inter-class homogeneity and intra-class heterogeneity in histopathological images[7, 24]. However, since the intra-class heterogeneity is mainly due to the diverse representations of the same category, the excessive similarity of the diverse representations may capture the spurious features of each category and decease the discriminative ability of the features, resulting in the generalization ability for segmentation is limited.

In this paper, we propose a hierarchical prompt contrastive method to preserve the intrinsic structures within each category and improve the inter-class distinctiveness simultaneously, termed as HPC. In the proposed HPC, the hierarchical text prompt contrastive loss and the hierarchical class prompt contrastive loss are designed to generate more accurate pseudo segmentation masks and improve the discriminative ability for each category respectively. For hierarchical text prompt contrastive loss, we use several transformation networks to transform the prototype of the learnable text prompt obtained from a frozen text encoder to match the image representation of the image encoder at different layers. Thus several CAMs[47] can be generated, and the high-confidence pseudo segmentation mask can be obtained with the consistency of the CAMs. With CAMs, we can produce both pseudo foreground segmentation mask and pseudo background segmentation mask. Then we adopt a frozen foundation model[36] to extract the robust foreground feature and background feature respectively, and design the contrastive loss between the robust features and text prompt to further learn more accurate pseudo segmentation masks. For hierarchical class prompt contrastive loss, it includes an intra-class contrastive loss and an inter-class contrastive loss. To capture the intra-class heterogeneity, we divide each category into several subclasses to produce subclass prototypes. The intra-class contrastive loss and inter-class contrastive loss are designed based on the subclass prototypes. The intra-class contrastive loss is mainly used to make the images in a subclass be similar while far away from the other subclass prototypes, and the inter-class contrastive loss is adopted to make the images to be similar with the subclass prototypes that belong to the same category and far away from the subclass prototypes of the other categories. In this way, we do not need to make the images within the same category be very similar with each other, while they are all far away from the different categories. Therefore, the generalization ability can be effectively improved. The experiments on the popular benchmark datasets show that HPC outperforms the state-of-the-art methods. The main contributions can be summarized as:

- We propose a hierarchical prompt contrastive method for weakly supervised histopathology segmentation method by preserving the structures within the same category and improving the distinctiveness of different categories simultaneously.
- The hierarchical text prompt contrastive loss is proposed to generate more accurate pseudo segmentation masks and the hierarchical class prompt contrastive loss is designed to further balance the inter-class homogeneity and intra-class heterogeneity.
- Extensive experiments on several popular and large-scale benchmark datasets show the proposed method outperforms the state-of-the-art methods.

2 Related Work

2.1 Weakly Supervised Semantic Segmentation

Weakly-supervised semantic segmentation (WSSS) aims to generate pixel-level masks derived from weak annotations, such as image-level labels[13, 42], point[28, 43], bounding boxes[18, 21], or scribbles[35]. Consequently, a significant body of research has focused on generating high-quality pseudo labels from these sparse cues, facilitating the development of numerous CAM-based[47] pipelines including Grad-CAM[30], pixel-affinity propagation[2], and saliency-based localization[23]. In the context of histopathological image segmentation, early CAM-driven frameworks, such as HistoSegNet[4], OEEM[25], and MLPS [13], largely adopt a classify-to-segment paradigm. In this approach, an image-level classifier produces initial pseudo masks, which are subsequently refined via post-processing techniques or noise-aware mining, or alternatively strengthened through stage-wise pseudo supervision prior to training the segmentation model. More recently, end-to-end architectures like TransWS[44] have introduced Transformer-based[33] branches and joint optimization strategies to obtain more integral activation maps and structurally complete pseudo masks.

Recently, vision-language models such as CLIP[29] have advanced WSSS by incorporating textual supervision, with adaptation methods like CLIMS[38] and ExCEL[41] effectively reducing the discrepancy between image-level labels and pixel-level predictions. In the specific domain of histopathology, methods generally fall into text-driven and prototype-driven strategies. Text-driven approaches such as TPRO[45] leverage external medical knowledge and text prompting to guide localization features. Conversely, prototype-driven methods like PBIP[31] avoid hand-crafted prompts and instead construct image banks via clustering to capture intra-class visual variations. However, current multimodal adaptations face a dilemma where fixed textual descriptions are often too coarse to capture severe intra-class heterogeneity, whereas prompt-free prototype weaken explicit semantic anchoring and often leave the feature space insufficiently structured to resolve inter-class homogeneity. Unlike TPRO which relies on fixed text descriptions that may be too coarse for diverse tissue morphologies, or PBIP which lacks explicit semantic anchoring, our framework

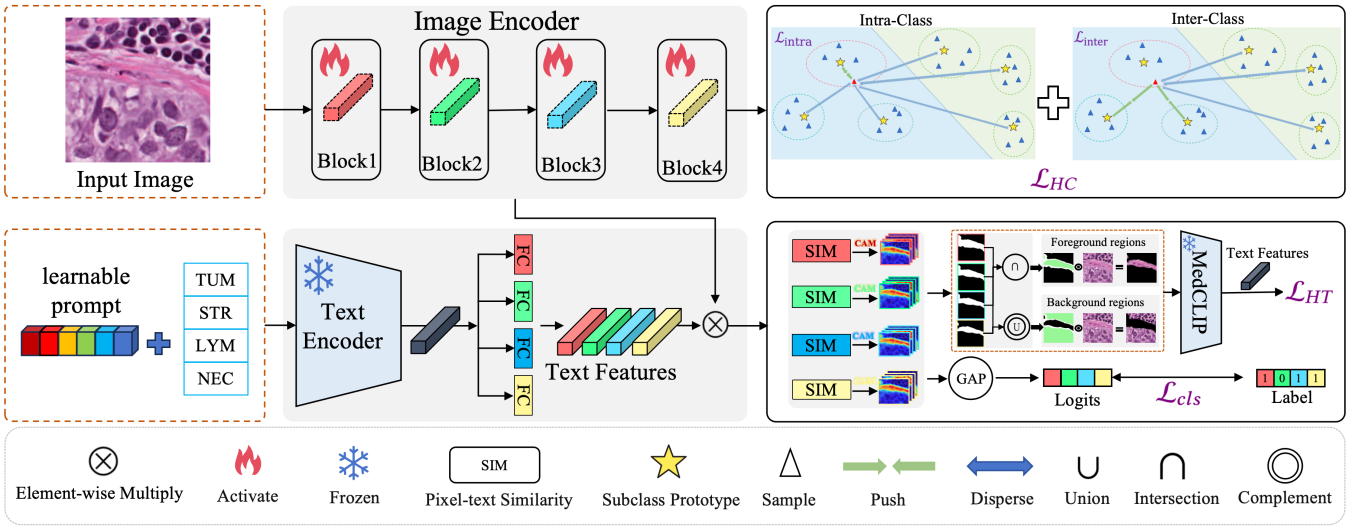


Figure 1: Overview of the proposed HPC. Text features are processed through four FC layers to compute layer-wise similarity maps with image features across four hierarchical levels. The HT module then extracts consistent foreground-background regions for contrastive learning with text , while the final-layer image features are aligned with subclass prototype via HC.

synergizes learnable context prompts with prototype contrast to simultaneously enforce semantic stability and fine-grained feature compactness for reliable pseudo-mask generation.

2.2 Contrastive Learning

Contrastive learning (CL) aims to learn discriminative representations by pulling semantically consistent positives closer while pushing inconsistent negatives apart, typically instantiated via the InfoNCE objective [32]. This has spurred the exploration of diverse strategies for designing positive and negative constructions in visual representation learning, encompassing instance-level pretraining frameworks like SimCLR [6] and MoCo [14], as well as label-aware formulations such as supervised contrast [19]. To better serve dense prediction, recent studies further extend CL from global instance embeddings to local pixel or region features, enabling spatially sensitive representations for segmentation. Representative approaches include DenseCL [34] and PixPro [40], as well as region-aware formulations that regularize segmentation features through region-level contrast[15].

Recently, contrastive alignment has been further strengthened by vision-language pretraining, where models such as CLIP [29] and MedCLIP [36] provide semantically grounded text anchors, and prompt learning such as CoOp [49] and CoCoOp[48] enables adapting textual contexts to specific domains without drifting from pretrained semantics. In weakly supervised semantic segmentation, contrastive objectives are also explored to refine activation quality, such as learning class-agnostic activation maps through cross-image contrast [39]. However, existing CL-based pipelines remain sensitive to noisy pair assignments and false negatives, and region contrast can amplify errors when pseudo regions are ambiguous or boundary-misaligned [15, 34, 39]. Moreover, coarse text descriptions are often insufficient to model the severe intra-class heterogeneity and inter-class similarity common in histopathology, leading to an under-structured feature space and low-quality pseudo masks [8]. Distinct from these methods, our framework anchors visual learning with frozen MedCLIP semantics and learnable context prompts [36, 49]. It further enhances pseudo-mask quality by restricting contrast to consensus-based high-purity regions and by capturing intra-class multi-modality via dual-granularity prototype contrast.

3 Methodology

3.1 Overview

In this section, we will introduce the proposed method in detail, and the flowchart of the proposed method is shown in Figure 1. In the proposed method, we use the image and text to address the inter-class homogeneity and intra-class heterogeneity by developing two kinds of hierarchical prompt contrastive learning strategies, i.e., hierarchical text prompt contrastive learning and hierarchical class prompt contrastive learning. The hierarchical text prompt contrastive learning is mainly used to generate high-confidence pseudo segmentation masks. The learnable prompt[49] is adopted to represent each category with a frozen text encoder. Generally, the texts align to the images with the final outputs of the image encoder and the text encoder. However, the representations of each layer in the image encoder are usually with different types of information. To guarantee the confidence of the pseudo segmentation masks, we design several independent linear layers as transformation networks to transform the representation of the text prompt to adaptively match the image representation at each layer in a hierarchical manner. Thus, each layer can produce a CAM[47], and we regard the consistency of these CAMs as the high-confidence foreground masks. For the background masks, we regard the regions without the regions that are composed by the union of the CAMs. Then we

align the segmentation masks with high-confidence pseudo labels to the corresponding text prompts by contrastive learning with a foundation model, such as MedCLIP[36], which is frozen and used to extract the features of segmentation masks for contrastive learning to further refine the masks. In hierarchical class prompt contrastive learning, we divide each category with image representations into several subclasses due to the intra-class heterogeneity, and thus each category can have several subclass prototypes. With the subclass prototypes, we design an intra-class contrastive loss and an inter-class contrastive loss. The inter-class contrastive loss is adopted to align the image to its corresponding prototype and simultaneously maximize the distances between the image and other subclass prototypes. While for the intra-class contrastive loss, for a given image, we force it to be similar to the other subclass prototypes within the same category and far away from the subclass prototype of the other categories. With the balance of intra-class contrastive loss and inter-contrastive loss, we can guarantee that the intrinsic structure within the category can be kept and the discriminative ability of the image can be effectively improved.

3.2 Hierarchical Text Prompt Contrastive Learning

Text prompts are very useful in weakly supervised learning[38]. Without loss of generality, we also use the text prompts to generate high-confidence masks[45] and design a hierarchical text prompt contrastive learning method. The text prompts are adopted to match the information at different layers of the image encoder. Generally, in a deep network, the shallow layers are effective at capturing the details, while the deep layers can learn the semantic information effectively. To capture the various information, a learnable prompt is used[49]. Suppose there are C categories in a histopathological image dataset, the text prompt of each category can be initialized with a sequence of M context tokens and a class token, represented as $[T]_1, [T]_2, \ldots, [T]_M, [CLS]$. In this way, the text prompt can be enhanced with optical prefix contexts. Then, given a text encoder $\mathcal{E}_{text}$, the representation of the text prompt for the kth category can be denoted as P_k. We should note that given an image encoder $\mathcal{E}_{image}$, for the image x_i in category k, the representations at $\mathcal{E}_{image}$ different layers are different. Hence, we transform P_k to adaptively match the representations at different layers. For convenience, we denote the transformation network as $\mathcal{T}_n$, where n is the nth layer, and it corresponds to the layers in $\mathcal{E}_{image}$. We denote the number of layers in $\mathcal{E}_{image}$ as N. Thus the transformation networks can be represented as $\{\mathcal{T}_n\}_{n=1}^N$. With these transformation networks, we can obtain the transformation of the kth category text prompt, denoted as $t_k^n = \mathcal{T}_n(P_k)$. We also denote the feature of the image x_i at the nth layer in $\mathcal{E}_{image}$ as f_k^n. Then the pixel-level similarity map can be obtained between text prompts and the image features, which can be represented as:

$$S_n = t_k^n \cdot f_k^n, n = 1, 2, ..., N \tag{1}$$

With the pixel-level similarity maps $\{S_n\}_{n=1}^N$, we can use the global average pooling for classification at each layer, and the cross-entropy loss is adopted. The classification loss can be represented as:

$$\mathcal{L}_{cls} = \sum_{n=1}^{N} \mathcal{L}_{cls}^n, \ \mathcal{L}_{cls}^n = CrossEntropy(y, \hat{y}_n). \tag{2}$$

where y is the label of the image and $\hat{y}_n$ is the prediction label based on the global average pooling of S_n. In fact, only with the classification loss, the deep network can be effectively optimized, including the transformation networks $\{\mathcal{T}_n\}_{n=1}^N$, the image encoder $\mathcal{E}_{image}$ and the text encoder $\mathcal{E}_{text}$. Generally, the text encoder is frozen to anchor semantic search space to the generalized concepts.

Then, we use the classification ability to generate pseudo segmentation masks, and the classification activation map (CAM) is adopted. For each similarity map $\{S_n\}_{n=1}^N$, we first upsample it to the original image resolution and binarize it via a dynamic threshold[31] to obtain the mask M_k^n for the k-th class, thus a set of candidate binary masks $\{M_k^n\}_{n=1}^N$ can be produced. Generally, the confidence pseudo labels should satisfy inter-layer consistency. Hence, we use the intersections of the binary masks $\{M_k^n\}_{n=1}^N$ to obtain the foreground masks, which can be represented as:

$$M_{fg}^k = \cap M_k^n, n = 1, 2, ..., N \tag{3}$$

where $\cap$ represents the intersection operator. Then we can assign pseudo label k to the region masked by M_{fg}^k. Similarly, to obtain the foreground mask, the background masks also can be obtained. To guarantee the confidence of the background masks, we do not simply use the complementary mask of M_{fg}^k, since the complementary mask of M_{fg}^k has the possibility to be predicted as a foreground mask at a certain layer. Hence, we regard the complementary mask of the union mask obtained by $\{M_k^n\}_{n=1}^N$ as the background mask, which can be represented as:

$$M_{bg}^k = 1 - \cup M_k^n, n = 1, 2, ..., N \tag{4}$$

where $\cup$ is the union between different sets. Thus we can be sure that the region masked by M_{bg}^k does not belong to the category k. Nextly, the contrastive losses are designed to further improve the segmentation ability of the weakly supervised network with pseudo segmentation labels. Following PBIP[31], we do not use the feature obtained from the image encoder to form contrastive loss and employ the image encoder from the pre-trained MedCLIP model[36], which is trained with large-scale medical image-text pairs. Since the representation ability of MedCLIP[36] is much better than that of the image encoder $\mathcal{E}_{image}$, we can efficiently refine $\mathcal{E}_{image}$ to generate more accurate CAMs. For convenience, we define the foreground regions as $\mathcal{B}_{fg}$ and the background region as $\mathcal{B}_{bg}$. Their features are extracted with the frozen image encoder of MedCLIP. Given the set of foreground regions $\mathcal{B}_{fg}$, for the ith foreground segmentation image $x_{fg}^i \in \mathcal{B}_{fg}$, let y_i denote its corresponding pseudo label. The similarity between x_{fg}^i and the text prompt of the jth category is calculated and denoted as $S_{fg}^{j,i}$. Thus, the contrastive loss for the foreground regions can be formulated as:

$$\mathcal{L}_{FG} = -\frac{1}{|\mathcal{B}_{fg}|} \sum_{i=1}^{|\mathcal{B}_{fg}|} \log \frac{\exp(S_{fg}^{y_i,i}/\tau)}{\sum_{j=1}^{C} \exp(S_{fg}^{j,i}/\tau)} \tag{5}$$

where τ is the temperature parameter, $|\mathcal{B}_{fg}|$ represents the total number of foreground segmentation images, and y_i is the pseudo

label of x_{fg}^i. Complementarily, we also design a contrastive strategy for the background regions, which can be represented as:

$$\mathcal{L}_{BG} = -\frac{1}{|\mathcal{B}_{bg}|} \sum_{i=1}^{|\mathcal{B}_{bg}|} \log \frac{\sum_{j \neq y_i}^{C} \exp(S_{bg}^{j,i}/\tau)}{\sum_{j=1}^{C} \exp(S_{bg}^{j,i}/\tau)} \tag{6}$$

where τ is the temperature parameter, $|\mathcal{B}_{bg}|$ is the total number of background segmentation images, and y_i denotes the category index that x_{bg}^i does not contain. With $\mathcal{L}_{FG}$ and $\mathcal{L}_{BG}$, we can make a background region far away from the category that it does not contain. The overall hierarchical text prompt contrastive loss can be obtained by combing the foreground contrastive loss and background contrastive loss, which can be represented as:

$$\mathcal{L}_{HT} = \lambda_{fg}\mathcal{L}_{FG} + \lambda_{bg}\mathcal{L}_{BG} \tag{7}$$

where λ_{fg} and λ_{bg} denote the balancing coefficients for the two optimization terms.

3.3 Hierarchical Class Prompt Contrastive Learning

The inter-class homogeneity and intra-class heterogeneity widely exist in real-world applications. To improve the separation ability between different categories, we first divide each category into several subclasses with clustering techniques, such as K-means. Thus the intra-class heterogeneity can be effectively captured, and we denote the number of subclasses for each category as u. Meanwhile, the prototypes of the subclasses for category k are denoted as $\{p_b^k\}_{b=1}^u$. We should note that the histopathological images may contain multiple labels, and the prototypes are calculated with the images that only have one label. Then two contrastive losses are designed, i.e., intra-class contrastive loss and inter-class contrastive loss. The intra-class contrastive loss is mainly developed to keep the images within the same subclass much more similar, while they should be far away from the other subclasses of the same category and the subclasses in other categories. Given an image x_i that belongs to the kth category and the bth subclass, the intra-class contrastive loss can be represented as:

$$\mathcal{L}_{intra} = -\log \frac{\exp(f_i \cdot p_b^k/\tau_1)}{\sum_{c=1}^{C}\sum_{j=1}^{u} \exp(f_i \cdot p_j^c/\tau_1)} \tag{8}$$

where τ_1 is the temperature parameter, and f_i is the feature representation of the image x_i with the image encoder $\mathcal{E}_{image}$. With $\mathcal{L}_{intra}$, although it can make the samples within the same subclass much more similar, it also may decrease the discriminative ability of the images that belong to the same category k but do not belong to the same subclass. Then an inter-class contrastive loss is developed to keep the similarity between the image x_i and the subclasses of the same class that x_i do not belong. Meanwhile, we also guarantee the separation ability between different categories. Given an image x_i that belongs to the kth category and the bth subclass, the inter-class contrastive loss can be represented as:

$$\mathcal{L}_{inter} = -\log \frac{\sum_{j=1,\, j\neq b}^{u} \exp\left(\frac{f_i \cdot p_j^k}{\tau_2}\right)}{\sum_{c=1}^{C}\sum_{j=1}^{u} \exp\left(\frac{f_i \cdot p_j^c}{\tau_2}\right) - \exp\left(\frac{f_i \cdot p_b^k}{\tau_2}\right)} \tag{9}$$

where τ_2 is the temperature parameter. We can observe that $\mathcal{L}_{inter}$ and $\mathcal{L}_{intra}$ always keep the image to be far away from the other categories. Hence, the discriminative ability between different categories can be improved. While within the same category, we force the image to be far away from the subclasses that the image does not belong to in $\mathcal{L}_{intra}$ and to be similar in $\mathcal{L}_{inter}$. In fact, it is not conflict. Since the two losses contain different temperature parameters, we can guarantee that the distances between the image and the other categories are farther than that between the image and the subclasses that the image does not belong to. In this way, we not only can preserve the structure within the same category but also improve the separation ability between different categories, and the generalization ability of the deep network for segmentation is improved. The hierarchical class prompt contrastive loss can be represented as:

$$\mathcal{L}_{HC} = \lambda_{intra}\mathcal{L}_{intra} + \lambda_{inter}\mathcal{L}_{inter} \tag{10}$$

where λ_{intra} and λ_{inter} are two trade-off parameters. For the multi-label images, we calculate the distance between the image and the subclass prototypes of the category that the image contains, and choose the subclass that has the smallest distance.

3.4 Optimization Objectives

Our final training objective $\mathcal{L}_{total}$ integrates all preceding modules to jointly optimize the entire framework. It is a unified objective that combines the primary classification loss $\mathcal{L}_{CLS}$, the contrastive losses $\mathcal{L}_{HT}$ and $\mathcal{L}_{HC}$:

$$\mathcal{L}_{total} = \mathcal{L}_{CLS} + \alpha\mathcal{L}_{HT} + \beta\mathcal{L}_{HC} \tag{11}$$

where α and β are hyperparameters that balance the contributions of each component. Then the image encoder $\mathcal{E}_{image}$, learnable prompts, and the transformation networks $\{\mathcal{T}_n\}_{n=1}^N$ can be optimized.

4 Experiments

4.1 Datasets

To assess the performance of our proposed method, experiments were conducted on two extensively utilized histopathology datasets: LUAD-HistoSeg[13] for lung adenocarcinoma and BCSS-WSSS[3, 13] for breast cancer.

4.1.1 LUAD-HistoSeg. This dataset is designed for lung adenocarcinoma [13]. It derives from Whole Slide Images (WSIs) sourced from two cohorts: 29 patients from Guangdong Provincial People's Hospital and 20 patients from the TCGA database. The WSIs were cropped at 10× magnification into non-overlapping 224×224 patches. Annotations are provided for four tissue phenotypes: Tumor Epithelium (TE), Tumor-Associated Stroma (TAS), Necrosis (NEC), and Lymphocyte (LYM). The dataset is divided into a training set of 16,678 patches with image-level labels and validation and testing sets of 300 and 307 patches, respectively, both with pixel-level ground truth for evaluation.

4.1.2 BCSS-WSSS. BCSS-WSSS[3, 13] is a dataset derived from the fully supervised Breast Cancer Semantic Segmentation (BCSS) dataset. It comprises 151 H&E-stained Regions of Interest (ROIs) extracted from breast cancer slides, processed into 31,826 patches

Table 1: Experimental comparison on LUAD-HistoSeg and BCSS-WSSS with best and second-best results highlighted.

Method	Publication	LUAD-HistoSeg						BCSS-WSSS					
		TE	NEC	LYM	TAS	fwIoU	mIoU	TUM	STR	LYM	NEC	fwIoU	mIoU
HistoSegNet[4]	ICCV'19	45.59	36.30	58.28	50.82	48.54	47.75	33.14	46.46	29.05	1.91	37.19	27.64
TransWS[44]	MICCAI'22	57.04	49.98	59.46	58.59	57.41	56.27	44.71	36.49	41.72	38.08	40.61	40.25
OEEM[25]	MICCAI'22	73.81	70.49	71.89	69.48	71.70	71.42	74.86	64.68	48.91	61.03	67.34	62.37
MLPS[13]	MedIA'22	73.90	77.48	73.61	69.53	72.51	73.63	74.54	64.45	52.54	58.67	67.36	62.55
CLIP-ES[26]	CVPR'23	58.44	52.73	60.17	57.09	57.74	57.11	56.09	51.07	47.58	49.34	52.81	51.02
PistoSeg[12]	AAAI'23	**80.07**	73.05	68.18	66.54	72.94	71.96	76.53	**71.39**	53.91	59.64	71.38	65.37
SIPE[5]	CVPR'22	73.14	65.26	66.18	67.23	69.43	67.95	73.29	63.87	49.28	52.49	66.00	59.73
HAMIL[46]	TMI'23	73.46	75.83	72.94	70.86	72.60	73.27	71.65	62.37	51.52	54.29	64.95	59.96
TPRO[45]	MICCAI'23	75.80	80.56	78.14	72.69	75.31	76.80	77.95	65.10	54.55	64.96	69.57	65.64
S2C[20]	CVPR'24	57.19	63.82	43.48	60.25	56.98	56.19	52.42	62.94	40.55	45.34	55.48	50.31
PBIP[31]	CVPR'25	74.08	78.20	76.29	71.27	73.64	74.96	78.79	65.14	**65.34**	67.60	71.12	69.22
HPC (Ours)	-	76.65	**80.67**	**78.73**	**75.59**	**76.84**	**77.91**	**80.99**	67.36	64.08	**70.65**	**73.02**	**70.77**

Table 2: Comparisons of the Effectiveness of Different Losses in HPC on the BCSS-WSSS Dataset

$\mathcal{L}_{CLS}$	$\mathcal{L}_{HT}$	$\mathcal{L}_{HC}$	TUM	STR	LYM	NEC	fwIoU	mIoU
✓			74.44	60.20	61.16	63.24	66.54	64.76
✓		✓	77.22	65.61	62.42	59.44	70.03	66.17
✓	✓		79.85	66.74	63.28	69.97	72.16	69.96
✓	✓	✓	**80.99**	**67.36**	**64.08**	**70.65**	**73.02**	**70.77**

Table 3: Comparative Analysis of Vision-Language Models on the BCSS-WSSS Dataset

Encoder	TUM	STR	LYM	NEC	fwIoU	mIoU
SegFormer[37]	74.68	65.07	53.74	53.03	67.58	61.63
CLIP[29]	79.65	67.23	64.29	60.46	72.00	67.91
PLIP[16]	79.67	66.86	**65.52**	65.19	72.16	69.31
MedCLIP[36]	**80.99**	**67.36**	64.08	**70.65**	**73.02**	**70.77**

(224×224 pixels). The dataset focuses on four primary tissue classes: Tumor (TUM), Stroma (STR), Lymphocytic infiltrate (LYM), and Necrosis (NEC). Following standard benchmarks, it is divided into three subsets: a training set of 23,422 patches with weak image-level annotations, and validation and testing sets with 3,418 and 4,986 patches, respectively. Unlike the training set, both validation and test sets include dense pixel-level masks for accurate segmentation evaluation.

4.2 Implementation Details

We employ the MixTransformer from SegFormer[37] as the backbone for segmentation, initialized with weights pre-trained on ImageNet[9]. The pretrained MedCLIP serves as our text encoder and image encoder for feature extraction. The training process contains two stages. In the first stage, the model is trained for 10 epochs with a batch size of 10 and a learning rate of 1×10^{-5}. Then in the second stage, pseudo labels generated from the first stage are used to train a fully supervised segmentation model. We use $\alpha = 0.5$ and $\beta = 1$ to balance the objectives. The other loss weights are set to $\lambda_{fg} = 1$, $\lambda_{bg} = 0.5$, $\lambda_{intra} = 1$, and $\lambda_{inter} = 0.2$. For the similarity calibration, the temperature is fixed to $\tau = 1.0$, while $\tau_1 = 0.07$ and $\tau_2 = 0.1$ are used in the corresponding contrastive components. All experiments are performed on a single NVIDIA 3090 GPU. The results are evaluated using category-wise Intersection over Union (IoU), including mean IoU (mIoU) and frequency-weighted IoU (fwIoU). We validated the effectiveness of our proposed method by comparing it with eleven advanced methods: HistoSegNet[4], TransWS[44], OEEM[25], MLPS[13], CLIP-ES[26], PistoSeg[12], SIPE[5], HAMIL[46], TPRO[45], S2C[20]. These comparisons were conducted on the standard LUAD-HistoSeg and BCSS-WSSS datasets. We follow the settings of the comparisons in their original paper.

4.3 Experimental Results

Table 1 shows the experimental results on the benchmark datasets. From the results, we can observe that our method achieves state-of-the-art performance on both LUAD-HistoSeg and BCSS-WSSS datasets in terms of fwIoU and mIoU. Specifically, on LUAD-HistoSeg, it achieves 76.84% fwIoU and 77.91% mIoU, outperforming TPRO by 1.53% and 1.11%, respectively. On BCSS-WSSS, the method attains 73.02% fwIoU and 70.77% mIoU, surpassing PistoSeg and PBIP by 1.64% and 1.55%, respectively. These results indicate that our framework generates more accurate pseudo masks, enabling more reliable segmentation training. Meanwhile, we should note HPC

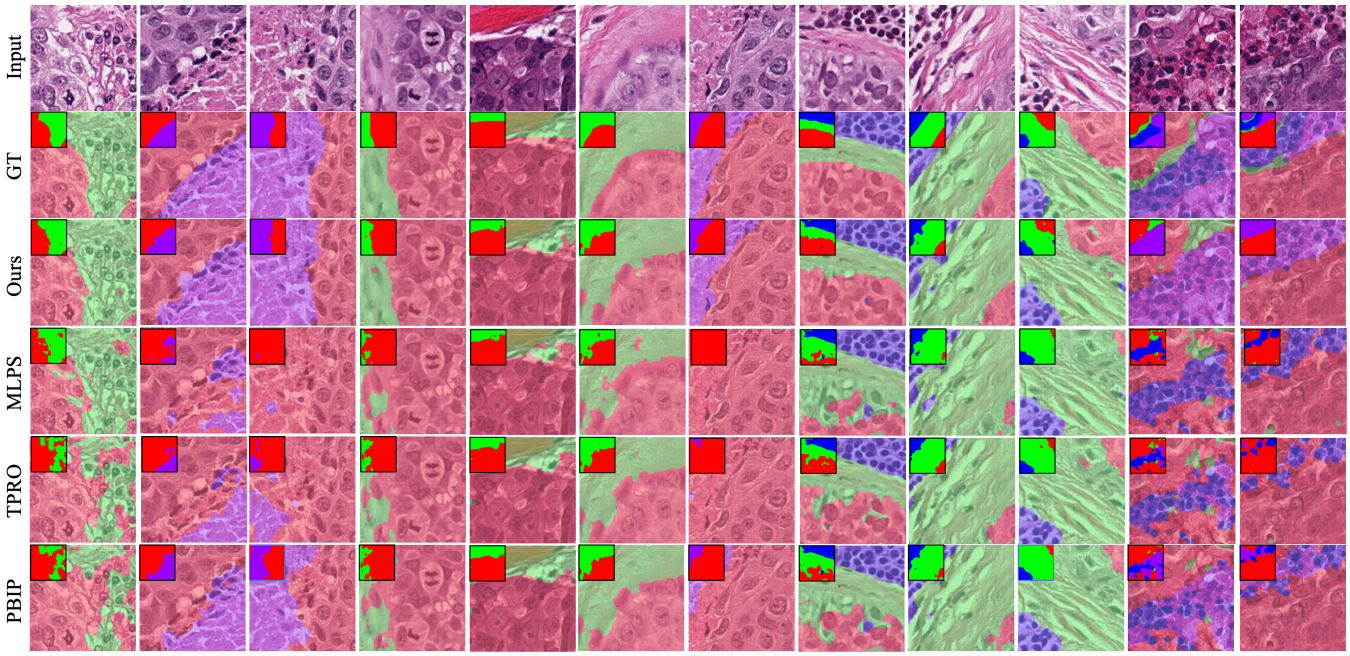


Figure 2: Visualization of patch-level segmentation. Predictions are overlaid onto the input images and displayed in the top-left corner.

Table 4: Ablation study on mask generation strategies

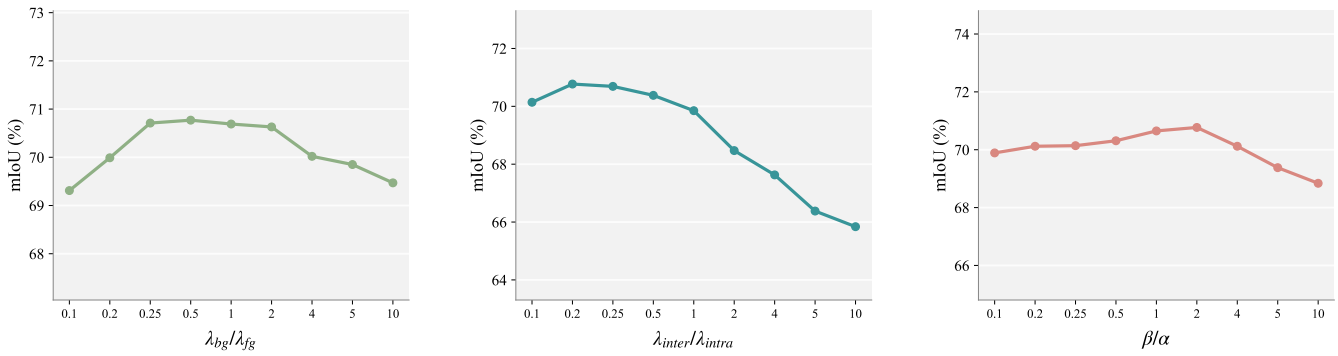
M_1	M_2	M_3	M_4	TUM	STR	LYM	NEC	fwIoU	mIoU
			✓	80.35	66.45	63.64	70.00	72.29	70.11
		✓	✓	80.84	67.15	63.08	**71.05**	72.78	70.53
	✓	✓	✓	80.95	67.04	**64.93**	69.83	72.92	70.69
✓	✓	✓	✓	**80.99**	**67.36**	64.08	70.65	**73.02**	**70.77**

has a good ability to distinguish different categories. Notably, categories such as TUM and STR tend to show larger gains because they exhibit strong intra-class heterogeneity and ambiguous boundaries with neighboring tissues, which often leads to noisy or incomplete pseudo masks. By refining pseudo masks and suppressing noisy regions, HPC provides cleaner supervision and thus improves these challenging classes more noticeably. On the LUAD-HistoSeg dataset, although the PistoSeg method achieves a 3.5% improvement compared with HPC in the TE category, HPC achieves at least a 7% improvement compared with PistoSeg for the other categories. Compared with the other methods except PistoSeg, HPC achieves the best performance for all categories. While on the BCSS-WSSS dataset, it has the similar observations. These observations well demonstrate that the proposed HPC can learn the robust feature representation of each category, and the discriminative representation of each category is improved. To observe the effectiveness of the proposed method directly, we also show some visual segmentation results in Figure 2. In Figure 2, three state-of-the-art methods PBIP, TPRO, MLPS are compared. For the visualization results, we can observe that the proposed method can obtain complete segmentation masks. While for the comparisons, they almost could not obtain the consistent segmentation masks and there are many noisy regions in the segmentation results. This also demonstrates that the proposed HPC can obtain the robust feature representation of each category, and the distinctiveness of each category is well improved.

Overall, the quantitative comparisons and the visualization results confirm that our method achieves state-of-the-art performance on both LUAD-HistoSeg and BCSS-WSSS datasets. Compared with recent approaches such as TPRO and PBIP, the proposed HPC can generate more accurate pseudo segmentation masks. Meanwhile, the distinctiveness of each category and the representation ability of each category are well improved to address the intra-class heterogeneity and inter-class homogeneity.

4.4 Ablation Studies

4.4.1 The contributions of $\mathcal{L}_{HT}$ and $\mathcal{L}_{HC}$. The hierarchical text prompt loss and the hierarchical class prompt loss are two important components in the proposed HPC. We show the ablation results for these two components in Table 2 on BCSS-WSSS dataset[3, 13]. From the ablation results, we can note that both the hierarchical text prompt loss and the hierarchical class prompt loss are important to improve the segmentation ability. The hierarchical class prompt loss and the hierarchical text prompt loss are designed at the image level. The performance of the hierarchical text prompt loss is better than that of the hierarchical class prompt loss. This indicates that it is much more important to improve the distinctiveness of different categories precisely. The ablation studies demonstrate that the



(a) Analysis of $\lambda_{bg}/\lambda_{fg}$. (b) Analysis of $\lambda_{inter}/\lambda_{intra}$. (c) Analysis of β/α.

Figure 3: Hyperparameter sensitivity analysis. We vary the internal weights of (a) $\mathcal{L}_{HT}$ and (b) $\mathcal{L}_{HC}$, and (c) their relative balance.

proposed method can deal with intra-class heterogeneity and inter-class homogeneity well.

4.4.2 Text Encoder and Image Encoder from Different Vision-Language Model. In the proposed method, a frozen text encoder and a frozen image encoder of a vision-language foundation model are adopted. CLIP[29] is a popular vision-language foundation model trained with text-image pairs. Then many variant models based on CLIP have been developed. PLIP[16] and MedCLIP[36] are two popular models that are developed for medical data. PLIP benefits from domain-specific pre-training on pathological data, but its advantages appear skewed toward specific fine-grained morphological features. MedCLIP is trained with much more medical data. To verify the influence of the vision-language model, we use different foundation models for evaluation. We compare the results on BCSS-WSSS dataset with CLIP, PLIP and MedCLIP in Table 3. From the results, we can observe that the performance of MedCLIP is better than that of PLIP while the performance of CLIP is the poorest one. This demonstrates that domain-specific knowledge from the foundation model is important to further improve the performance of the proposed method.

4.4.3 CAM from Different Layers. In the proposed method, we use several CAMs[47] to generate accurate pseudo segmentation masks. There are four layers in the SegFormer[37] backbone. We show the results with the combination of different layers in Table 4. From the results, we can observe that the performance of the proposed method is increasing as the number of CAM masks increases. Although the performance is not significantly improved, the results with the most number of CAMs almost achieve the best or second-best performance. Hence, it is necessary to use more CAMs to generate the high-confidence pseudo segmentation masks and obtain the robust performance.

4.4.4 Hyperparameter Sensitivity. In the proposed method, there are several trade-off parameters. We show a simple analysis results in Figure 3 by changing one and fixing others on BCSS-WSSS. In the experiments, the ratio for different contrastive losses is analyzed, i.e., $\lambda_{bg}/\lambda_{fg}$ in $\mathcal{L}_{HT}$, $\lambda_{inter}/\lambda_{intra}$ in $\mathcal{L}_{HC}$, and β/α between $\mathcal{L}_{HC}$ and $\mathcal{L}_{HT}$. From Figure 3 (a), we can observe that the foreground segmentation masks are more important than the background masks to improve the performance. However, the background masks are essential, and the weight λ_{bg} can be half of that of λ_{fg}. From Figure 3 (b), we can note that the intra-class contrastive loss is important to keep the performance. This indicates that it is important to preserve the structures within the category and λ_{inter} can be set with a small number to further improve the performance. From Figure 3 (c), we can note that the $\mathcal{L}_{HC}$ and $\mathcal{L}_{HT}$ are both important and they can be set with the same values.

5 Conclusion

In this work, we have presented a Hierarchical Prompt Contrastive (HPC) framework to explicitly tackle the challenges of inter-class homogeneity and intra-class heterogeneity in weakly supervised histopathology segmentation. In the proposed method, the hierarchical text prompt contrastive loss and the hierarchical class prompt contrastive loss are designed to preserve the intra-class structures and improve the inter-class distinctiveness. With hierarchical text prompt contrastive loss, we use a hierarchical manner to obtain several CAMs to generate more accurate pseudo-segmentation masks to improve the distinctiveness of each category. For the hierarchical class prompt contrastive loss, it preserves the structures within each category by keeping the samples close to their subclass they belong to and making the distances between the samples and the subclasses of the other categories farther than that between the samples and the other subclasses of the same category. Thus, the representation ability within the category can be improved. Extensive experiments on the LUAD-HistoSeg and BCSS-WSSS datasets have validated the effectiveness of our method. Meanwhile, we should note that the MedCLIP is very important to improve the performance of Segformer. In the future, we aim to design more efficient and light-weight deep networks using domain-specific foundation models, making them more practical for real-world applications.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62271357, Grant 62476204, Grant

62225113, Grand 62502448, in part by the National Science and Technology Major Project Grant No. 2024YFF1401003, 2024ZD1600100, 2023YFC2705700, the Innovative Research Group Project of Hubei Province under Grant 2024AFA017, in part by the Natural Science Foundation of Hubei Province under Grant 2023BAB072, in part by Wuhan Natural Science Foundation under Grant 2024040801020236, and in part by the Fundamental Research Funds for the Central Universities and Supercomputing Center of Wuhan University through Supercomputing System.

References

[1] Khalid AbdulJabbar, Shan E Ahmed Raza, Rachel Rosenthal, Mariam Jamal-Hanjani, Selvaraju Veeriah, Ayse Akarca, Tom Lund, David A Moore, Roberto Salgado, Maise Al Bakir, et al. 2020. Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nature medicine* 26, 7 (2020), 1054–1062.

[2] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. 2019. Weakly Supervised Learning of Instance Segmentation With Inter-Pixel Relations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2209–2218. doi:10.1109/CVPR.2019.00231

[3] Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A. Atteya, Mai A. T. Elsebaie, Lamia S. Abo Elnasr, Rokia Adel Sakr, Hazem S. E. Salem, Ahmed F. Ismail, Anas M. Saad, Joumana Ahmed, Maha A. T. Elsebaie, Mustafijur Rahman, Inas A. Ruhban, Nada M. Elgazar, Yahya Alagha, Mohamed H. Osman, Ahmed M. Alhusseiny, Mariam M. Khalaf, Abo-Alela F. Younes, Ali Abdulkarim, Duaa M. Younes, Ahmed M. Gadallah, Ahmad M. Elkashash, Salma Y. Fala, Basma M. Zaki, Jonathan D. Beezley, Deepak Roy Chittajallu, David Manthey, David A. Gutman, and Lee A. D. Cooper. 2019. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinform.* 35, 18 (2019), 3461–3467. doi:10.1093/BIOINFORMATICS/BTZ083

[4] Lyndon Chan, Mahdi S. Hosseini, Corwyn Rowsell, Konstantinos N. Plataniotis, and Savvas Damaskinos. 2019. HistoSegNet: Semantic Segmentation of Histological Tissue Type in Whole Slide Images. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 10661–10670. doi:10.1109/ICCV.2019.01076

[5] Qi Chen, Lingxiao Yang, Jianhuang Lai, and Xiaohua Xie. 2022. Self-supervised Image-specific Prototype Exploration for Weakly Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 4278–4288. doi:10.1109/CVPR52688.2022.00425

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 1597–1607. http://proceedings.mlr.press/v119/chen20j.html

[7] Zhe Chen, Zhao Chen, Jingxin Liu, Qiang Zheng, Yuang Zhu, Yanfei Zuo, Zhaoyu Wang, Xiaosong Guan, Yue Wang, and Yuan Li. 2021. Weakly Supervised Histopathology Image Segmentation With Sparse Point Annotations. *IEEE J. Biomed. Health Informatics* 25, 5 (2021), 1673–1685. doi:10.1109/JBHI.2020.3024262

[8] Ozan Ciga, Tony Xu, and Anne Louise Martel. 2022. Self supervised contrastive learning for digital histopathology. *Machine learning with applications* 7 (2022), 100198.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 248–255. doi:10.1109/CVPR.2009.5206848

[10] Amelie Echle, Niklas Timon Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander Thomas Pearson, and Jakob Nikolas Kather. 2021. Deep learning in cancer pathology: a new generation of clinical biomarkers. *British journal of cancer* 124, 4 (2021), 686–696.

[11] Lei Fan, Dongdong Fan, Yiwen Ding, Yong Wu, Donglin Di, Maurice Pagnucco, and Yang Song. 2025. GrainBrain: Multiview Identification and Stratification of Defective Grain Kernels. *IEEE Trans. Ind. Informatics* 21, 5 (2025), 3627–3635. doi:10.1109/TII.2025.3528555

[12] Zijie Fang, Yang Chen, Yifeng Wang, Zhi Wang, Xiangyang Ji, and Yongbing Zhang. 2023. Weakly-Supervised Semantic Segmentation for Histopathology Images Based on Dataset Synthesis and Feature Consistency Constraint. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, Brian Williams, Yiling Chen, and Jennifer Neville (Eds.). AAAI Press, 606–613. doi:10.1609/AAAI.V37I1.25136

[13] Chu Han, Jiatai Lin, Jinhai Mai, Yi Wang, Qingling Zhang, Bingchao Zhao, Xin Chen, Xipeng Pan, Zhenwei Shi, Zeyan Xu, Su Yao, Lixu Yan, Huan Lin, Xiaomei Huang, Changhong Liang, Guoqiang Han, and Zaiyi Liu. 2022. Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. *Medical Image Anal.* 80 (2022), 102487. doi:10.1016/J.MEDIA.2022.102487

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 9726–9735. doi:10.1109/CVPR42600.2020.00975

[15] Hanzhe Hu, Jinshi Cui, and Liwei Wang. 2021. Region-aware Contrastive Learning for Semantic Segmentation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 16271–16281. doi:10.1109/ICCV48922.2021.01598

[16] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. 2023. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine* 29, 9 (2023), 2307–2316.

[17] Yuxin Kang, Hansheng Li, Xiaoshuang Shi, Xiao Zhang, Yaqiong Xing, Yuting Wen, Yi Wang, Lei Cui, Jun Feng, and Lin Yang. 2025. Exploring Unbiased Activation Maps for Weakly Supervised Tissue Segmentation of Histopathological Images. *IEEE Trans. Medical Imaging* 44, 6 (2025), 2631–2642. doi:10.1109/TMI.2025.3541115

[18] Hoel Kervadec, Jose Dolz, Shanshan Wang, Eric Granger, and Ismail Ben Ayed. 2020. Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. In *International Conference on Medical Imaging with Deep Learning, MIDL 2020, 6-8 July 2020, Montréal, QC, Canada (Proceedings of Machine Learning Research, Vol. 121)*, Tal Arbel, Ismail Ben Ayed, Marleen de Bruijne, Maxime Descoteaux, Hervé Lombaert, and Christopher Pal (Eds.). PMLR, 365–381. http://proceedings.mlr.press/v121/kervadec20a.html

[19] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html

[20] Hyeokjun Kweon and Kuk-Jin Yoon. 2024. From SAM to CAMs: Exploring Segment Anything Model for Weakly Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 19499–19509. doi:10.1109/CVPR52733.2024.01844

[21] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. 2021. Anti-Adversarially Manipulated Attributions for Weakly and Semi-Supervised Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 4071–4080. doi:10.1109/CVPR46437.2021.00406

[22] Jungbeom Lee, Seong Joon Oh, Sangdoo Yun, Junsuk Choe, Eunji Kim, and Sungroh Yoon. 2022. Weakly Supervised Semantic Segmentation using Out-of-Distribution Data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 16876–16885. doi:10.1109/CVPR52688.2022.01639

[23] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. 2021. Railroad Is Not a Train: Saliency As Pseudo-Pixel Supervision for Weakly Supervised Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 5495–5505. doi:10.1109/CVPR46437.2021.00545

[24] Chao Li, Xinggang Wang, Wenyu Liu, Longin Jan Latecki, Bo Wang, and Junzhou Huang. 2019. Weakly supervised mitosis detection in breast histopathology images using concentric loss. *Medical Image Anal.* 53 (2019), 165–178. doi:10.1016/J.MEDIA.2019.01.013

[25] Yi Li, Yiduo Yu, Yiwen Zou, Tianqi Xiang, and Xiaomeng Li. 2022. Online Easy Example Mining for Weakly-Supervised Gland Segmentation from Histology Images. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part IV (Lecture Notes in Computer Science, Vol. 13434)*, Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li (Eds.). Springer, 578–587. doi:10.1007/978-3-031-16440-8_55

[26] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. 2023. CLIP is Also an Efficient Segmenter: A Text-Driven Approach for Weakly Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 15305–15314. doi:10.1109/CVPR52729.2023.01469

[27] Ziniu Qian, Kailu Li, Maode Lai, Eric I-Chao Chang, Bingzheng Wei, Yubo Fan, and Yan Xu. 2022. Transformer Based Multiple Instance Learning for Weakly Supervised Histopathology Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part II (Lecture Notes in Computer*

Science, Vol. 13432), Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li (Eds.). Springer, 160–170. doi:10.1007/978-3-031-16434-7_16

[28] Hui Qu, Pengxiang Wu, Qiaoying Huang, Jingru Yi, Zhennan Yan, Kang Li, Gregory M. Riedlinger, Subhajyoti De, Shaoting Zhang, and Dimitris N. Metaxas. 2020. Weakly Supervised Deep Nuclei Segmentation Using Partial Points Annotation in Histopathology Images. *IEEE Trans. Medical Imaging* 39, 11 (2020), 3655–3666. doi:10.1109/TMI.2020.3002244

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

[30] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 618–626. doi:10.1109/ICCV.2017.74

[31] Qingchen Tang, Lei Fan, Maurice Pagnucco, and Yang Song. 2025. Prototype-Based Image Prompting for Weakly Supervised Histopathological Image Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 30271–30280. doi:10.1109/CVPR52734.2025.02818

[32] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR* abs/1807.03748 (2018). arXiv:1807.03748 http://arxiv.org/abs/1807.03748

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[34] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. 2021. Dense Contrastive Learning for Self-Supervised Visual Pre-Training. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 3024–3033. doi:10.1109/CVPR46437.2021.00304

[35] Yuheng Wang, Yuhan Zheng, Chloe Yue, Thomas Zhang, Jiayue Cai, Chunqi Chang, Harvey Lui, Sunil Kalia, Z. Jane Wang, and Tim K. Lee. 2026. GloW-VSNet: A scribble-based weakly supervised framework for global-view vitiligo lesion segmentation. *Medical Image Anal.* 109 (2026), 103920. doi:10.1016/J.MEDIA.2025.103920

[36] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 3876–3887. doi:10.18653/V1/2022.EMNLP-MAIN.256

[37] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José M. Álvarez, and Ping Luo. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. (2021), 12077–12090. https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html

[38] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. 2022. CLIMS: Cross Language Image Matching for Weakly Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 4473–4482. doi:10.1109/CVPR52688.2022.00444

[39] Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, and Linlin Shen. 2022. C2am: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 989–998.

[40] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. 2021. Propagate Yourself: Exploring Pixel-Level Consistency for Unsupervised Visual Representation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 16684–16693. doi:10.1109/CVPR46437.2021.01641

[41] Zhiwei Yang, Yucong Meng, Kexue Fu, Feilong Tang, Shuo Wang, and Zhijian Song. 2025. Exploring CLIP's Dense Knowledge for Weakly Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 20223–20232. doi:10.1109/CVPR52734.2025.01883

[42] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. 2020. Causal Intervention for Weakly-Supervised Semantic Segmentation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/07211688a0869d995947a8fb11b215d6-Abstract.html

[43] Hongrun Zhang, Liam Burrows, Yanda Meng, Declan Sculthorpe, Abhik Mukherjee, Sarah E. Coupland, Ke Chen, and Yalin Zheng. 2023. Weakly Supervised Segmentation with Point Annotations for Histopathology Images via Contrast-Based Variational Model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 15630–15640. doi:10.1109/CVPR52729.2023.01500

[44] Shaoteng Zhang, Jianpeng Zhang, and Yong Xia. 2022. TransWS: Transformer-Based Weakly Supervised Histology Image Segmentation. In *Machine Learning in Medical Imaging - 13th International Workshop, MLMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings (Lecture Notes in Computer Science, Vol. 13583)*, Chunfeng Lian, Xiaohuan Cao, Islem Rekik, Xuanang Xu, and Zhiming Cui (Eds.). Springer, 367–376. doi:10.1007/978-3-031-21014-3_38

[45] Shaoteng Zhang, Jianpeng Zhang, Yutong Xie, and Yong Xia. 2023. TPRO: Text-Prompting-Based Weakly Supervised Histopathology Tissue Segmentation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2023 - 26th International Conference, Vancouver, BC, Canada, October 8-12, 2023, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 14220)*, Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu E. Salcudean, James Duncan, Tanveer F. Syeda-Mahmood, and Russell H. Taylor (Eds.). Springer, 109–118. doi:10.1007/978-3-031-43907-0_11

[46] Lanfeng Zhong, Guotai Wang, Xin Liao, and Shaoting Zhang. 2023. HAMIL: High-Resolution Activation Maps and Interleaved Learning for Weakly Supervised Segmentation of Histopathological Images. *IEEE Trans. Medical Imaging* 42, 10 (2023), 2912–2923. doi:10.1109/TMI.2023.3269798

[47] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2921–2929. doi:10.1109/CVPR.2016.319

[48] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 16795–16804. doi:10.1109/CVPR52688.2022.01631

[49] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.

[50] Zhi-Hua Zhou. 2018. A brief introduction to weakly supervised learning. *National science review* 5, 1 (2018), 44–53.