# Weakly-Supervised Semantic Segmentation for Histopathology Images Based on Dataset Synthesis and Feature Consistency Constraint


**Abstract**

Weakly-Supervised Semantic Segmentation (WSSS) for computational pathology has great potential to alleviate the time-consuming and labor-intensive burden of manual pixel-level annotations. Currently, numerous studies attempt to use image-level labels to achieve pixel-level segmentation to reduce the need for fine annotations. However, most of these methods are based on class activation map (CAM), which suffers from inaccurate segmentation boundaries. To address this problem, Fang proposed a novel weakly-supervised tissue segmentation framework named PistoSeg, which is implemented under a fully-supervised manner by transferring tissue category labels to pixel-level masks. Firstly, a dataset synthesis method is proposed based on Mosaic transformation to generate synthesized images with pixel-level masks. Next, considering the difference between synthesized and real images, this paper devises an attention-based feature consistency, which directs the training process of a proposed pseudo-mask refining module. Finally, the refined pseudo-masks are used to train a precise segmentation model for testing. Although it achieves significant performance on nature images relying on CAM, it cannot perform well on histopathological images due to the homogeneous features of different tissue types. Moreover, some medical contrastive language-image pretraining (CLIP) have great representation capability for histopathology but they have not been fully used to capture homogeneous features in histopathological images. To fill this cap, we propose a improved framework named PLIMSeg, which aims to leverage contrastive language-image pretraining for WSSS. Specifically, PLIMSeg introduces Pathology Language Image Matching (PLIM) into histopathology WSSS, aiming to utilize the strong representation capability of pre-training language image models for classification. Moreover, we design three PLIM-based losses to refine CAMs, among which tissue and tissue-prompt alignment, background, and tissue-prompt suppression losses are designed to guide the model to excite more reasonable and complete tissue for CAM of each category. Tissue and background-prompt suppression loss is proposed to prevent the model from activating closely related background regions. These designs enable the proposed PLIMSeg to generate a more complete and compact activation map for the target object. Our approach outperforms than baseline on BCSS-WSSS, setting a new state-of-the-art for WSSS of histopathology images.



**Keywords:** Weakly-Supervised Semantic Segmentation, Histopathology images.

# 1 Introduction

Benefit from the rapid development of computational pathology, pathologists have recently seen a lot of AI technologies in tedious tasks such as cancer diagnosis, sub-typing, and others. In computational pathology, automatic tissue segmentation is one of the most important studies, as tumor generation and development are closely related to the tumor microenvironment (TME), which is formed by the interaction of various types of tissues [3]. TME is a highly complex ecosystem formed with different types of tissues, including tumor epithelial, cancer cells, fibroblasts, inflammatory cells, tumor-infiltrating lymphocytes, tumor-associated stroma, etc [9]. So accurate differentiation and segmentation of histopathological tissues is beneficial for predicting the prognosis of cancer patients, guiding oncologists in the medication decision, and helping determine the clinical therapy, which are essential in cancer treatment [10].

Currently, most automated tissue segmentation pipelines are based on fully-supervised methods, which require fine annotations at the pixel level [9,22]. However, the massive difference between the centimeter scale of histopathology sections and the sub-micron scale of cells make the acquisition of pixel-level annotations very time-consuming and laborious. One of the effective solutions is weakly-supervised semantic segmentation (WSSS), which relies solely on image-level [1, 17], bounding box-level [6], point-level [4], or scribble-based supervision [20]. This work aims to only use image-level labels to supervise the learning of a semantic segmentation model.

Existing WSSS approaches are typically based on class activation map (CAM) [25]. The basic idea of CAM is to use localization clues brought by classification models to generate pixel-level pseudo masks. However, since classification tasks only need to focus on the most discriminative regions, a well-known drawback of CAM is its inability to depict exact object boundaries. Furthermore, histopathology images are more homogeneous than natural images, as tissue sections are formed by irregular and arbitrary repetitions of cells and tissues, making the morphological features of different tissue types more similar to each other than natural images, which amplifies the boundary uncertainty.

The homogeneity of histopathology images makes the synthesis of them easier than natural images, although it may cause problems for CAM-based methods. Since the same type of tissues tends to cluster together, numerous histopathology images have only a single tissue category besides the background (tissue-free regions). This prior knowledge inspires us with a new approach to achieve WSSS for histopathology images. Firstly, synthesized images can be built using histopathology images with a single tissue category, which can inherit pixel-level annotations from the image-level labels, making all synthesized images have pixel-level annotations as well. Then, a segmentation model can be trained in a fully-supervised manner to generate pseudo-masks for the original training set. Since there is no classification during the procedure, the problem of CAM can be eliminated.

In general, this paper proposes a framework named PistoSeg for weakly-supervised tissue segmentation based on dataset synthesis, providing a new direction for the research community. Besides, this paper devises a novel attention-based feature consistency, directed by which a pseudo-mask refining module is proposed to take advantage of CAM-based methods, making our proposed framework easy to be integrated with existing WSSS methods. The main contributions of this paper are in three aspects.

- This paper proposes a novel WSSS framework named PistoSeg, which fulfills tissue segmentation by

dataset synthesis to transfer tissue category labels to pixel-level masks. Therefore, weakly-supervised segmentation is implemented in a fully-supervised manner. To the best of our knowledge, this is the first method that brings data synthesis into WSSS for histopathology images, providing a new direction for the research community.

- An attention-based feature consistency is proposed to constrain the features of the same image after different pseudo-mask strategies for more efficient feature extraction. Directed by this consistency, a pseudo-mask refining module is designed to further improve the segmentation performance on the basis of dataset synthesis.

To utilize the representation capability Contrastive Language-Image Pretraining (CLIP) [18], we propose an improved framework for weakly supervised histopathology semantic Segmentation (PLIMSeg). PLIMSeg utilizes Pathology Language Image Matching (PLIM) to provide additional supervision to refine the initial CAM by performing alignment between tissue/background and prompts, where PLIM is pre-trained on the large-scale Quilt-1M dataset [12]. To provide additional supervision, we introduce three PLIM-based losses to perform tissue/background and prompts alignment, i.e., Tissue and Tissue-Prompt Alignment ($L_{TTP}$), Background and Tissue-Prompt Suppression($L_{BTP}$) and Tissue and Background-Prompt Suppression ($L_{TBP}$). In order to minimize the distance between different tissues and their text labels, we leverage tissue and tissue-prompt alignment to match the tissue features with the corresponding tissue-prompt. Aiming to suppress the relation between the background and the tissue label, the background and tissue-prompt suppression maximizes the distance between the features of background and tissue text labels. Similarly, we introduce tissue and background-prompt suppression to suppress the relation between tissue features and background prompts.

## 2 Related works

### 2.1 WSSS for Pathology Images.

Semantic segmentation based on fully supervised learning encounters the limited data problem due to the difficulty of obtaining pixel-level annotations. To solve this problem, a common approach is replacing fine-grained annotations with weak labels, such as image-level labels [21], scribbles [14], points [4], and bounding boxes [16]. Among them, image-level labels are the most effortless to obtain and thus have received much attention. Since CAM [25] was proposed, numerous WSSS methods have been designed for natural images, which primarily focus on addressing the boundary ambiguity problem in CAM. For instance, HistoSegNet [5] utilized GradCAM [19] with a series of designed post-processings for histopathology image segmentation. WSSS-Tissue [8] was proposed with a progressive dropout mechanism to propel CAMs to focus on the indiscriminative regions. Li et al. [13] proposed OEEM, which forced the segmentation model to learn from the credible supervision signals by assigning higher weights to samples with lower losses. Zhong et al. [24] proposed a novel two-stage weakly supervised segmentation framework based on high-resolution activation maps and interleaved learning (HAMIL). Zhang et al. [23] proposed a text-prompting-based weakly supervised segmentation method (TPRO), which uses text to introduce additional information. It employed MedCLIP [15] as label encoder and BERT as knowledge encoder. Since the MedCLIP [15] is fune-tuned by radiology dataset, it is not suitable for pathology images. Fang et al. [7] proposed a novel weakly-supervised tissue segmentation

framework named PistoSeg, with dataset synthesis and feature consistency constraint. It is the first method that brings data synthesis into WSSS for histopathology images, but it only use image-level label, which cannot provide fine segmentation supervision. Therefore, in this work we will employ this architecture for CAM and logits generation, and add a text-driven evaluator for further supervision. Although the above research can alleviate the fuzzy boundary problem to some extent, the gap between classification and segmentation makes this issue still an unsolved problem.

# 3 Method

## 3.1 Overview

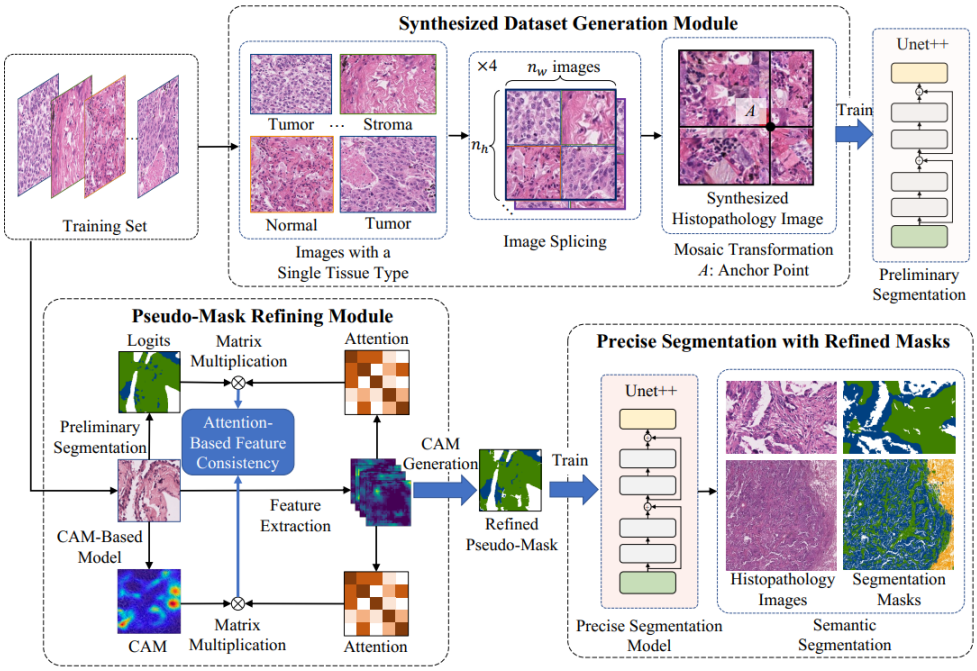



Figure 1. An overview of the PistoSeg framework. Firstly, based on the Mosaic transformation, a synthesized dataset generation module generates synthesized histopathology images with pixel-level masks, which are utilized to train a preliminary segmentation model. Next, assisting with an attention-based feature consistency, a pseudo-mask refining module generates refined masks, which are utilized to train a precise segmentation model for semantic segmentation with higher accuracy.

The goal of WSSS is to train a segmentation model which can predict the pixel-level masks for images in the test dataset with only image-level labels. Considering that if the image-level label of a histopathology image has only one category, all pixels in the image fall in the same category as well. Inspired by this, we propose a PistoSeg, which can be divided into two modules as shown in Figure 2. In the first module, the PistoSeg selects images with one tissue category from the training set. The selected images are then spliced and composed based on the Mosaic transformation to form a synthesized dataset with pixel-level annotations. Finally, a preliminary segmentation model is trained with the synthesized dataset, which is employed to infer pseudo-masks for the whole training set. Next, this paper proposes an attention-based feature consistency, under which the pseudo-mask refining module is trained for generating better pseudo-masks, serving as pixel-

level annotations for the following precise segmentation. The pseudo-mask refining module takes training images as input and employs the preliminary segmentation model and a CAM-based model to transfer the images to logits and CAM. Considering the features of an image after different transfers should be consistent, an attention-based feature consistency is proposed to generate refined pseudo masks. Eventually, a precise segmentation model can be trained based on the refined pseudo-masks, which are utilized to realize WSSS for the test dataset.

### 3.2 Loss

In this section, we briefly review the loss of pistoseg. The initial CAM is generated by an image encoder supervised by a classification loss $L_{\text{cls}}$.

$$L_{\text{cls}} = -\frac{1}{K}\sum_{k=1}^{K} y_k \log \frac{1}{1+e^{-z_k}} + (1-y_k)\log \frac{e^{-z_k}}{1+e^{-z_k}}, \tag{1}$$

where K is the number of tissue categories, $y_k$ and $z_k$ are the image-level label and predicted logits of category k.

Then,PistoSeg [7] adopts feature consistency constraints to refine the pseudo masks. Considering that the features of an image after different transfers should be consistent, two attention-based feature consistency loss $L_{\text{cons}}$ and $L_{\text{cross}}$ are utilized. $M_{\text{CAM}}$ and $M_{\text{pseudo}}$ are generated by two pre-trained networks and then refined by multiplication with attention maps from image encoder, named $\hat{M}_{\text{CAM}}$ and $\hat{M}_{\text{pseudo}}$. $L_{\text{cons}}$ makes $M_{\text{CAM}}$ and $M_{\text{pseudo}}$ be consistent, $L_{\text{cross}}$ ensure that the input space of the images is not affected by attention space. $L_{\text{cons}}$ and $L_{\text{cross}}$ are defined as,

$$L_{\text{cons}} = \|\hat{M}_{\text{pseudo}} - \hat{M}_{\text{CAM}}\|_1, \tag{2}$$

$$L_{\text{cross}} = \|M_{\text{pseudo}} - \hat{M}_{\text{CAM}}\|_1 + \|M_{\text{CAM}} - \hat{M}_{\text{pseudo}}\|_1. \tag{3}$$

### 3.3 Our improvement: Pathology Language Image Matching

Conventional CAM tends to focus on the most discriminative region of the image with only image-level label supervision, resulting in incomplete segmentation maps and unnecessary activation of closely related backgrounds. Therefore, PLIMSeg introduces pathology language image matching constraint, uses PLIM as an image-text feature extractor and designs three PLIM-based losses, for WSSS. Details of the proposed PLIM are displayed in Fig. 2. It consists of an image encoder $I(\cdot)$ and a text encoder $T(\cdot)$ based on QuiltNet. QuiltNet is a CLIP fine-tuned by 1M pathology datasets.

Specifically, given an image $X$, the image encoder predicts the initial CAM $M_k$, which represents the probability of each pixel belonging to a category $k$. $M_k$ and $(1-M_k)$ are then multiplied with the input image to mask out the kth histopathology tissue and background area, respectively, which serve as the input of the image encoder of the PLIM. For the initial CAM, the categories of tissue $M_k$ and background $(1-M_k)$ are defined as follows:

$$M_k \in \{R_1, R_2, ..., R_K\}, \tag{4}$$

$$(1-M_k) = \{R_j | j \in [1, K], j \neq k\}, \tag{5}$$

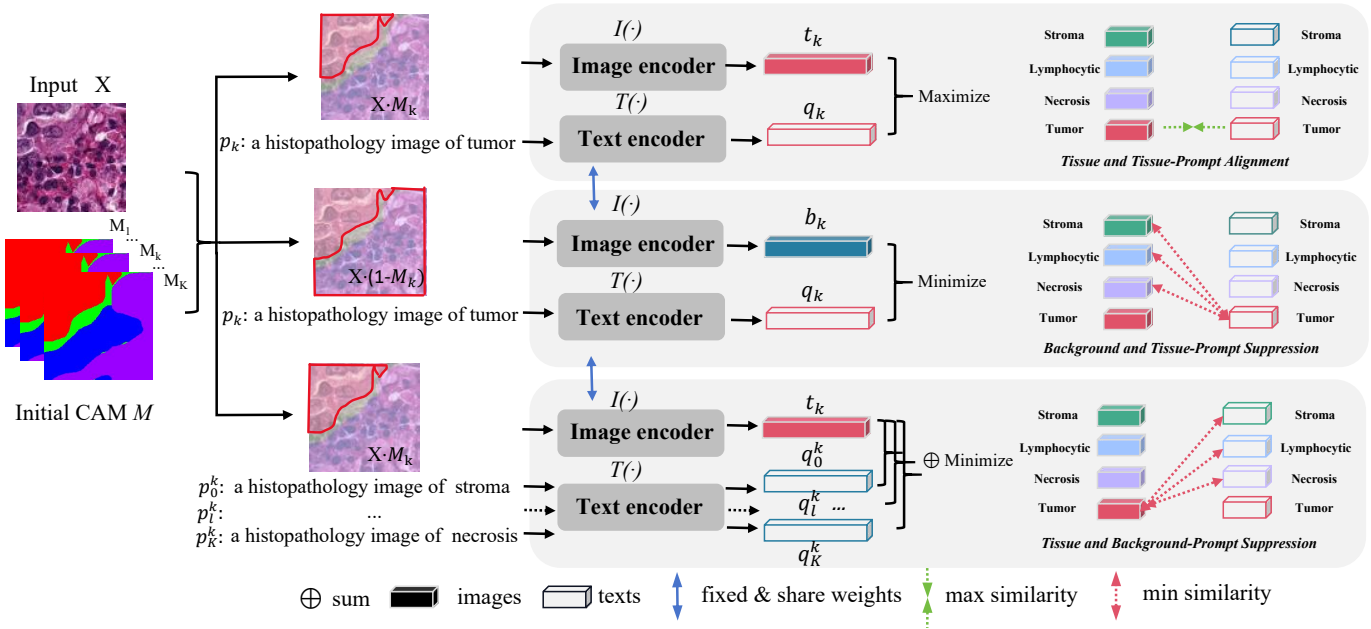



Figure 2. The proposed PLIM and three PLIM-based loss functions, i.e.,Tissue and Tissue-Prompt Alignment ($L_{TTP}$), Background and Tissue-Prompt Suppression ($L_{BTP}$) and Tissue and Background-Prompt Suppression ($L_{TBP}$).

where $R_K$ means the $K$ classes in the dataset. The k-th tissue and background pixels are then mapped to representation vectors $t_k$ and $b_k$ by image encoder $I(\cdot)$:

$$t_k = I(X \cdot M_k), \tag{6}$$

$$b_k = I(X \cdot (1 - M_k)). \tag{7}$$

Following the QuiltNet [12], the prompt of tissue $p_k$ for $(X \cdot M_k)$ is represented as " a histopathology image of { category k}", e.g."a histopathology image of tumor." The closely related background prompts $p_l^k$ for the specific class k is the rest K-1 classes in the dataset. It can be obtained as:

$$p_l^k = \{p_j | j \in [1, K], j \neq k\}. \tag{8}$$

Then the prompt $p_k$ and $p_l^k$ are mapped to representation vectors $q_k$ and $q_l^k$ by text encoder $T(\cdot)$ :

$$q_k = T(p_k), q_l^k = T(p_l^k). \tag{9}$$

The PLIM-based loss strategy aims to assist the generation of CAMs $M$ through the supervision of language supervision. Specifically, Tissue and Tissue-Prompt Alignment loss is designed to activate more accurately identifies the position and the shape of tissue. On the other hand, Background and Tissue-Prompt loss prevents the model from activating background regions related to the tissue by suppressing background and tissue-prompts, ensuring that the model only activate special tissue region. Additionally, the introduction of Tissue and Background-Prompt loss further addresses the issue of false activation in closely-related backgrounds. The final training loss is a weighted combination of these loss functions, guiding the model to activate more reasonable target regions and avoid activating closely-related background regions by adjusting hyperparameter weights.

#### 3.3.1 Tissue and Tissue-Prompt Alignment

Given the k-th tissue representation $t_k$ and its prompt representation $q_k$, we begin by calculating the cosine similarity between image and text representations and then maximize it using the proposed loss $L_{TTP}$:

$$\mathcal{L}_{TTP} = -\sum_{k=1}^{K} y_k \cdot \log(sim(t_k, q_k)), \tag{10}$$

where $sim(a, b)$ indicates the cosine similarity between $a$ and $b$. The more similar $a$ and $b$ are, the closer $sim(a, b)$ is to 1. The generated initial CAMs will gradually approach the target object under the supervision of $L_{TTP}$.

#### 3.3.2 Background and Tissue-prompt Suppression

Given the background representation $b_k$ and its corresponding tissue-prompt representation $q_k$, the $L_{BTP}$ is calculated as follow:

$$\mathcal{L}_{BTP} = -\sum_{k=1}^{K} y_k \cdot \log(1 - sim(b_k, q_k)). \tag{11}$$

When $L_{BTP}$ is minimized, fewer background pixels are reserved in $X \cdot (1 - M_k)$ and more target tissue contents are recovered in $(X \cdot M_k)$. This ensures that more complete object contents are activated in $M_k$.

#### 3.3.3 Tissue and Background-Prompt Suppression

Given the tissue representation $t_k$ and its corresponding prompt representation of background $q_l$, the loss is calculated as:

$$\mathcal{L}_{TBP} = -\sum_{k=1}^{K} \sum_{l=1, l \neq k}^{K} y_k \cdot \log(1 - sim(t_k, q_l^k)). \tag{12}$$

During training, the backbone network will gradually suppress the false activation of closely related background regions in $M_k$ for the minimization of $L_{TBP}$.

The overall training loss for the proposed text-driven learning framework PLIM can be formulated as:

$$L_{\text{plim}} = \alpha L_{\text{TTP}} + \beta L_{\text{BTP}} + \gamma L_{\text{TBP}}, \tag{13}$$

where $\alpha$, $\beta$ and $\gamma$ are the hyper-parameters weighting the three loss terms.

Overall, the total loss of PLIMSeg is calculated as the summation of the above four losses, i.e.,

$$L = L_{\text{cls}} + L_{\text{cons}} + L_{\text{cross}} + L_{\text{plim}}. \tag{14}$$

## 4 Implementation details

### 4.1 Comparing with the released source codes

The main contributions of our improvement can be summarized as follows:

- We propose a novel framework named PLIMSeg by leveraging the strong representation capability of Contrastive Language-Image Pre-training for histopathology, to improve the quality of pseudo masks for weakly-supervised histopathology image segmentation.

Table 1. Ablation study of losses on the BCSS-WSSS datasets.

| | TUM | STR | LYM | NEC | mIOU |
|---|---|---|---|---|---|
| No language supervision | 0.8116 | 0.7398 | 0.5626 | 0.6673 | 0.6953 |
| +$\mathcal{L}_{TTP}$ | 0.8113 | 0.7400 | 0.5669 | 0.6706 | 0.6960 |
| +$\mathcal{L}_{TTP}$+$\mathcal{L}_{BTP}$ | 0.8087 | 0.7402 | 0.5872 | **0.6810** | 0.7043 |
| +$\mathcal{L}_{TTP}$+$\mathcal{L}_{BTP}$+$\mathcal{L}_{TBP}$ | **0.8139** | **0.7541** | **0.6033** | 0.6716 | **0.7107** |

- We design three PLIM-based loss functions, i.e., $L_{\text{TTP}}$, $L_{\text{BTP}}$ and $L_{\text{TBP}}$, to complementarily learn the complete CAMs and mitigate the false activation through the supervision of the text.

### 4.2 Dataset

The weakly-supervised tissue segmentation dataset BCSS-WSSS [2] is adopted in this paper. BCSS-WSSS is a weakly supervised tissue semantic segmentation dataset extracted from the fully supervised segmentation dataset BCSS, which contains 151 representative H&E-stained breast cancer pathology slides. The dataset was randomly cut into 31826 patches of size 224 × 224 and divided into a training set (23422 patch-level annotations), a validation set (3418 pixel-level annotations), and a test set (4986 pixel-level annotations) according to the official split. There are four tissue classes in this dataset, including Tumor (TUM), Stroma (STR), Lymphocytic infiltrate(LYM), and Necrosis (NEC).

### 4.3 Experimental environment setup

All experiments are done with an NVIDIA A100 GPU. Codes are written with Pytorch 1.12.0 and Pytorch Lightning 1.6.4. We follow the training settings in [7] for the framework with 25 training epochs and learning rate of 1e-3. The default batch size is 16. For evaluation, category-wise intersection over union (IoU), mean IoU (mIoU) and and frequency weighted IoU (fwIoU) are adopted as the metrics. All baselines and comparison methods are implemented strictly following their papers or using their open-sourced codes.

## 5 Results and analysis

### 5.1 Impact of Loss Functions.

In this subsection, we conducted ablation experiments to further evaluate the effectiveness of three individual loss functions. The results are summarized in Table 1. Notably, the impact of incorporating $L_{TTP}$ may not be immediately evident, given its primary function of capturing the complete tissue of CAM, aligning with our goal of operating without language supervision. With the introduction of $L_{BTP}$, a slight improvement is observed. This component effectively attenuates background information, resulting in sharper edges in pseudo masks. Meanwhile, $L_{TBP}$ proves crucial in mitigating false activation originating from closely-related background elements. Employing all three losses simultaneously further improve the effectiveness. The visual results are presented in Fig. 3, demonstrate the impact of incorporating these loss functions (+L denotes the successive addition of each loss function). Here, we use tumor as foreground tissue and other tissues as background. As shown in the second column of Fig. 3, when using only $L_{TTP}$, we observe that 1) only the

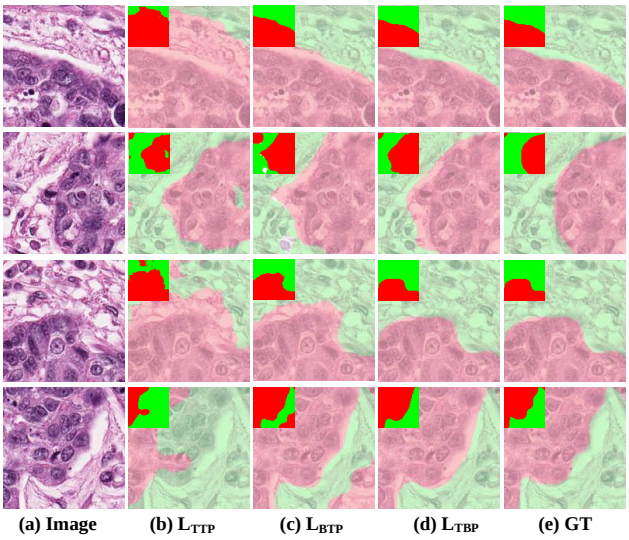



Figure 3. Pseudo masks generated by the proposed PLIM using different combinations of loss functions. Input images are shown in column 1. Columns 2 to 4 present the CAMs generated using $\mathcal{L}_{TTP}$, $\mathcal{L}_{TTP}$ +$\mathcal{L}_{BTP}$, $\mathcal{L}_{TTP}$ +$\mathcal{L}_{BTP}$ +$\mathcal{L}_{TBP}$, and Ground Truth

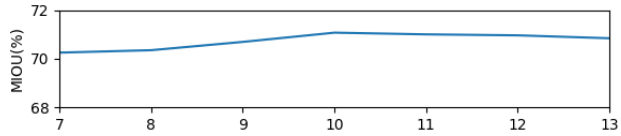
Table 2. Ablation study of different CLIPs on the BCSS-WSSS dataset.

| | TUM | STR | LYM | NEC | MIOU |
|---|---|---|---|---|---|
| CLIP | 0.8062 | 0.7272 | 0.6105 | 0.6525 | 0.6991 |
| MedCLIP | 0.8135 | 0.7447 | 0.5914 | 0.6506 | 0.7001 |
| PLIP | 0.8087 | 0.7349 | **0.6117** | 0.6644 | 0.7049 |
| QuiltNet | **0.8139** | **0.7541** | 0.6033 | **0.6716** | **0.7107** |

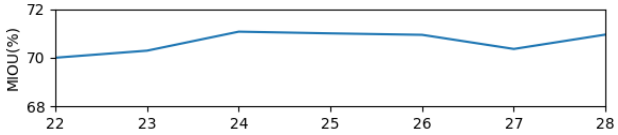
discriminative tumor are activated in the CAMs; 2) closely-related backgrounds, such as stroma, are also activated in the initial CAMs. With the addition of $L_{BTP}$ (third column), there is a significant increase in the size of activated regions, leading to the activation of more complete object regions. However, it should be noted that $L_{TTP}$ + $L_{BTP}$ can falsely activate background regions. As shown in the fourth column of Fig. 3, the inclusion of $L_{TBP}$ effectively addresses the aforementioned issues. It can be observed that $L_{TBP}$ efficiently constrains the size of activated regions and notably excludes closely-related background elements, such as stroma, from the CAMs, closely resembling the ground-truth.

### 5.2 Impact of CLIP.

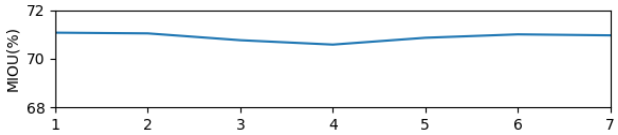
We conducte an ablation experiment to compare the performance of four contrastive language-image pre-training approaches: CLIP [18], MedCLIP [15], PLIP [11], and QuiltNet [12]. The results are presented in Table 2. It is evident that both PLIP and QuiltNet exhibit an overall advantage in terms of mIoU compared to MedCLIP and CLIP. This superiority arises from their fine-tuning with pathology-specific datasets. In the case of MedCLIP, its overall performance is akin to CLIP, as it undergoes fine-tuning on a radiology dataset and lacks prior knowledge of pathology. Notably, the superior results achieved by QuiltNet can be attributed to its utilization of a larger dataset for fine-tuning, surpassing even that of PLIP.




(a)




(b)




(c)

Figure 4. Sensitivity analyses of hyper-parameters $\alpha$, $\beta$ and $\gamma$. The mIoU values here are reported on BCSS-WSSS dataset.

### 5.3 Sensitivity Analysis.

There are three hyper-parameters in Eq. 13. The sensitivity analyses of these three parameters are performed on BCSS-WSSS dataset and the results are presented in Fig. 4 . It is observed that the performances of our approach are stable with the variation of $\alpha$(from 7 to 13), $\beta$ (from 22 to 28) and $\gamma$ (from 1 to 5), i.e., our approach is not sensitive to hyper-parameters. In our experiments, the default value of $\alpha$, $\beta$ and $\gamma$ are 10, 24 and 1, respectively.

## 6 Conclusion and future work

In this paper, we propose the PLIMSeg to address the limitation of weakly supervised semantic segmentation on histopathology images by incorporating contrastive language image pretraining as additional supervision. We argue that image-level labels alone cannot provide sufficient supervision and that text supervision can provide additional guidance to the model. We employ PLIM as image-text feature extractor and design tissue and tissue-prompt alignment, background and tissue-prompt suppression losses to guide the model to excite more reasonable and complete object regions for CAM of each category. In addition, we design a tissue and background-prompt suppression loss to prevent the model from activating closely-related background regions. The proposed method achieves the best results on BCSS-WSSS dataset, demonstrating the superiority of our method.

## References


[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4981–4990, 2018.

[2] Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A Atteya, Mai AT Elsebaie, Lamia S Abo Elnasr, Rokia A Sakr, Hazem SE Salem, Ahmed F Ismail, Anas M Saad, et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18):3461–3467, 2019.

[3] Borros Arneth. Tumor microenvironment. *Medicina*, 56(1):15, 2019.

[4] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016.

[5] Lyndon Chan, Mahdi S Hosseini, Corwyn Rowsell, Konstantinos N Plataniotis, and Savvas Damaskinos. Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10662–10671, 2019.

[6] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015.

[7] Zijie Fang, Yang Chen, Yifeng Wang, Zhi Wang, Xiangyang Ji, and Yongbing Zhang. Weakly-supervised semantic segmentation for histopathology images based on dataset synthesis and feature consistency constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 606–613, 2023.

[8] Chu Han, Jiatai Lin, Jinhai Mai, Yi Wang, Qingling Zhang, Bingchao Zhao, Xin Chen, Xipeng Pan, Zhenwei Shi, Zeyan Xu, et al. Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. *Medical Image Analysis*, 80:102487, 2022.

[9] Penghui He, Aiping Qu, Shuomin Xiao, and Meidan Ding. Detisseg: A dual-encoder network for tissue semantic segmentation of histopathology image. *Biomedical Signal Processing and Control*, 87:105544, 2024.

[10] Dominique C Hinshaw and Lalita A Shevde. The tumor microenvironment innately modulates cancer progression. *Cancer research*, 79(18):4557–4566, 2019.

[11] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, pages 1–10, 2023.

[12] Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *arXiv preprint arXiv:2306.11207*, 2023.

[13] Yi Li, Yiduo Yu, Yiwen Zou, Tianqi Xiang, and Xiaomeng Li. Online easy example mining for weakly-supervised gland segmentation from histology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 578–587. Springer, 2022.

[14] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016.

[15] Kaushalya Madhawa and Raul Carlomagno. Medclip: Fine-tuning a clip model on the roco medical dataset, 2021.

[16] Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6913–6922, 2021.

[17] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1713–1721, 2015.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[19] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[20] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7158–7166, 2017.

[21] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12275–12284, 2020.

[22] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[23] Shaoteng Zhang, Jianpeng Zhang, Yutong Xie, and Yong Xia. Tpro: Text-prompting-based weakly supervised histopathology tissue segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 109–118. Springer, 2023.

[24] Lanfeng Zhong, Guotai Wang, Xin Liao, and Shaoting Zhang. Hamil: High-resolution activation maps and interleaved learning for weakly supervised segmentation of histopathological images. *IEEE Transactions on Medical Imaging*, 2023.

[25] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.