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

Zijie Fang[1*], Yang Chen[1*], Yifeng Wang[2*], Zhi Wang[1†], Xiangyang Ji[3], Yongbing Zhang[2†]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Harbin Institute of Technology (Shenzhen)
[3]Department of Automation, Tsinghua University
vison307@gmail.com, cy21@mails.tsinghua.edu.cn, wangyifeng@stu.hit.edu.cn,
wangzhi@sz.tsinghua.edu.cn, ybzhang08@hit.edu.cn

Abstract

Tissue segmentation is a critical task in computational pathology due to its desirable ability to indicate the prognosis of cancer patients. 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, which suffers from inaccurate segmentation boundaries. To address this problem, we propose 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. Experiments based on WSSS4LUAD and BCSS-WSSS validate that PistoSeg outperforms the state-of-the-art methods. The code is released at https://github.com/Vison307/PistoSeg.

Introduction

Benefiting from the rapid development of deep learning, computational pathology has recently freed pathologists from tedious tasks such as cancer diagnosis, subtyping, 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 (Arneth 2019). Accurate differentiation and segmentation of histopathological tissues can indicate the prognosis of cancer patients, guide oncologists in the medication decision, and help determine the clinical therapy, which are essential in cancer treatment (Hinshaw and Shevde 2019).

Currently, most automated tissue segmentation pipelines are based on fully-supervised methods, which require fine annotations at the pixel level. However, the massive difference between the centimeter scale of histopathology sections and the sub-micron scale of cells makes the acquisition of pixel-level annotations very time-consuming and laborious. In addition, only professional pathologists or those with clinical backgrounds are competent for annotation, making it hard to utilize popular annotation methods for natural images such as crowdsourcing (Wazny 2017), which further increases the difficulty of obtaining pixel-level annotated data.

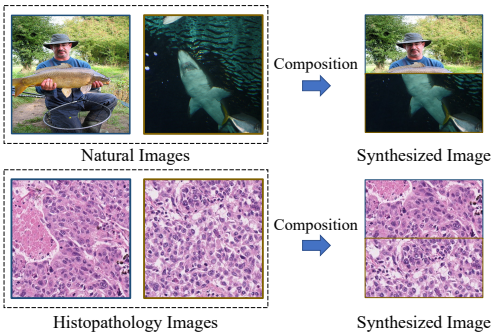

Figure 1: Synthesis of histopathology images is convenient due to the homogeneity of tissue sections, which is impossible for natural images because of not only the difficulty of discriminating background from the foreground in natural images but also the indivisible nature of natural objects.

According to previous research (Han et al. 2022a), coarse-grained image labels can reduce over 95% of the annotation time compared to pixel-level annotations. This opens a new research direction named weakly-supervised semantic segmentation (WSSS). Currently, many WSSS methods are based on the class activation map (CAM) (Zhou et al. 2016). 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 (Ahn and Kwak 2018). Furthermore, histopathology images are more homogeneous than natural images, as tis-

*These authors contributed equally.
†Co-corresponding authors: Yongbing Zhang and Zhi Wang.

sue 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.

Actually, the homogeneity of histopathology images makes the synthesis of them easier than natural images, as shown in Figure 1, 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 (Jia et al. 2017), 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 under 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.
- Various experiments validate the PistoSeg achieves the best performance beyond the state-of-the-art methods on the WSSS4LUAD and BCSS-WSSS datasets.

Related Works

WSSS for Natural Images

Semantic segmentation based on fully-supervised learning encounters the scarce data source 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 (Wang et al. 2020), scribbles (Lin et al. 2016), points (Bearman et al. 2016), and bounding boxes (Oh, Kim, and Ham 2021). Among them, image-level labels are the most effortless to obtain and thus have received much attention.

Since CAM (Zhou et al. 2016) was proposed, numerous WSSS methods at the image level have been designed for natural images, which primarily focus on addressing the boundary ambiguity problem in CAM. For example, SC-CAM (Chang et al. 2020) divided objects into subcategories by clustering at the feature level and trained a classification network with the subcategory information, forcing the classification network to learn better boundaries. AffinityNet (Ahn and Kwak 2018) was proposed based on random walking, which propagated local class responses to adjacent regions with similar semantic features to realize the boundary refinement. Wang et. al proposed SEAM (Wang et al. 2020), whose core idea was that if an image passes through an affine transformation, its segmentation mask should go through the same transformation, which brings a regularization for model training. However, all the methods mentioned above were designed for natural images. Due to the higher homogeneity of histopathology images, the performance is usually unsatisfactory when directly applying them to histopathology images.

WSSS for Medical Images

Fine annotations of medical images require expert knowledge and are more difficult to obtain. Some scholars tried to adopt multiple instance learning (MIL) into WSSS for medical images. For example, Xu et. al proposed multiple clustered instance learning to segment cancer and non-cancer regions (Xu et al. 2014). CAMEL was proposed by considering histopathology images as bags and latticed patches as instances (Xu et al. 2019). Jia et. al proposed DWS-MIL to segment cancerous regions in histopathology images (Jia et al. 2017). Besides MIL, CaCL (Wang et al. 2019) utilized variational autoencoder for binary WSSS of immunohistochemistry-stained images. However, since these methods only consider negative and positive labels, only binary segmentation (e.g., cancer and non-cancer) can be achieved, whose clinical importance is largely restricted as the TMA contains interactions of multiple tissue types.

Others applied CAM-based methods on WSSS of medical images as well. For instance, HistoSegNet (Chan et al. 2019) utilized GradCAM (Selvaraju et al. 2017) with a series of designed post-processings for histopathology image segmentation. Han et. al proposed WSSS-Tissue with a progressive dropout mechanism to propel CAMs to focus on the indiscriminative regions (Han et al. 2022a). Chen et. al introduced category and anatomy causality and proposed C-CAM (Chen et al. 2022). However, C-CAM is designed for MRI images whose anatomical relationships (e.g., tissue locations) are relatively fixed and not present in histopathology images. Li et. al proposed OEEM, which forced the segmentation model to learn from the credible supervision signals by assigning higher weights to samples with lower losses (Li et al. 2022). Although the above research can alleviate the fuzzy boundary problem to some extent, the gap between classification and segmentation makes this issue cannot be solved entirely. In contrast, PistoSeg takes a different approach by training a segmentation model on a synthesized dataset in a fully-supervised manner, which avoids this problem fundamentally. Besides, the pseudo-masks of PistoSeg

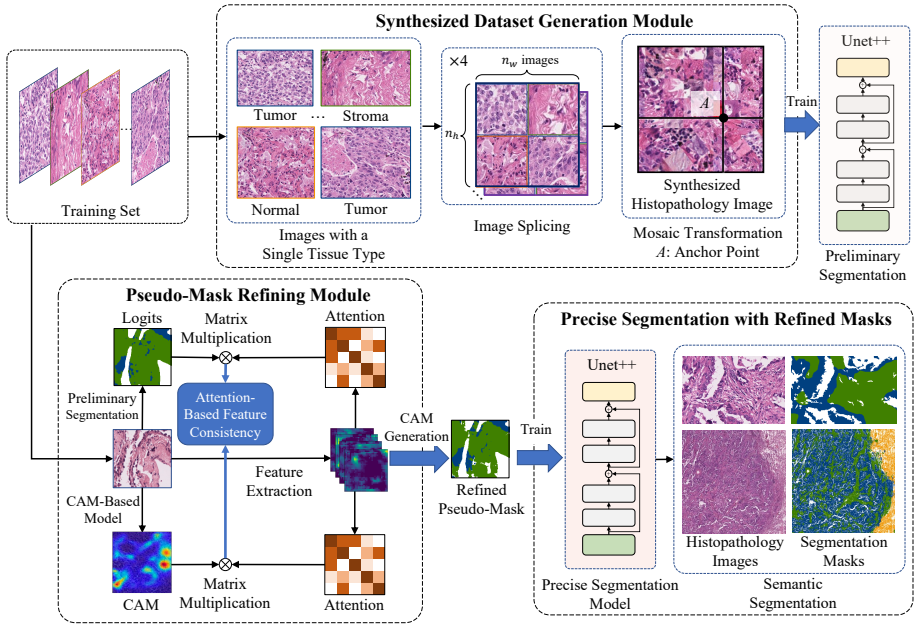

Figure 2: An overview of the PistoSeg framework. Firstly, based on the Mosaic transformation, a synthesized dataset generation module generates synthesized histopathology images with pixel-level masks, which are utilized to train a preliminary segmentation model. Next, assisting with an attention-based 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.

can also be refined with CAM-based methods with a proposed feature consistency.

Methodology

Framework Overview

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 (Bochkovskiy, Wang, and Liao 2020) 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 the 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 is utilized to realize WSSS for the test dataset.

Synthesized Dataset Generation Module

The most straightforward approach to obtain a dataset with pixel-level annotations is directly utilizing images containing a single tissue type. However, as there are usually multiple types of tissues in real TMEs, this strategy is far from satisfactory. To efficiently utilize such single tissue type information for segmentation tasks, the PistoSeg splices images containing only one tissue type to generate synthesized images with multiple tissue types based on Mosaic transformation, which will be described in detail in the following.

Single Tissue Type Image Splicing In the beginning, $n_h \times n_w$ images with a single tissue type are selected from the training set. Each image is cropped to the size of $H_p \times W_p$. Next, these images are gridded following a raster order to form a spliced image $I'_i \in \mathbf{R}^{(n_h H_p)\times(n_w W_p)\times 3}$. After repeating the above steps four times, four spliced images, denoted as I'_1 to I'_4, are obtained.

Mosaic Transformation The four spliced images are then utilized to form one synthesized image shaped $(n_h H_p) \times (n_w W_p)$ based on Mosaic transformation. Firstly, random rotation and scaling are applied to each spliced image. Then, an anchor point $A = (H_A, W_A)$ is generated (shown in Figure 2), where $\alpha n_h H_p \leq H_A \leq \beta n_h H_p$ and $\alpha n_w W_p \leq W_A \leq \beta n_w W_p$ ($0 < \alpha < \beta < 1$ are two hyperparameters. We set $\alpha = 0.2$ and $\beta = 0.8$ in implementation). Next, I'_1 to I'_4 are randomly cropped according to the anchor point,

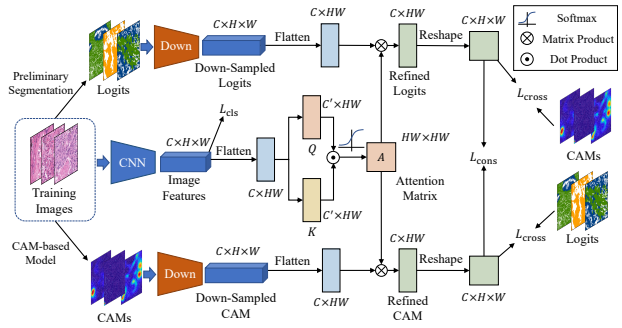

Figure 3: The pseudo-mask refining module. As the pseudo-mask and CAM are different transfers of the same image, their potential features should be closed to each other, bringing the regularization for the training of the module.

generating four intermediate images I_1'' to I_4'', whose shapes are $H_A \times W_A$, $(n_h H_p - H_A) \times W_A$, $H_A \times (n_w W_p - W_A)$, and $(n_h H_p - H_A) \times (n_w W_p - W_A)$, respectively. Finally, the four cropped intermediate images are placed in the four corners to form a synthesized histopathology image I_k, i.e.,

$$I_k = \begin{bmatrix} I_1'' & I_3'' \\ I_2'' & I_4'' \end{bmatrix}. \tag{1}$$

Since each synthesized image I_k is made up of images with single tissue type, pixel-level masks are available for the synthesized image, which can be utilized as the ground truth for training a preliminary segmentation network.

Pseudo-Mask Refining Module

Although the preliminary segmentation model can directly generate pseudo-masks of the training set, artifacts that differ from the actual tissue morphology exist in the synthesized dataset, which makes the pseudo-masks not optimal. Considering the inputs of CAM are without artifacts, this paper proposes an attention-based feature consistency to train a pseudo-mask refining module to take the advantages of CAM-based methods. It should be noted that any CAM generation mechanism can be applied here, including but not limited to pure CAM (Zhou et al. 2016), GradCAM (Selvaraju et al. 2017), and GradCAM++ (Chattopadhay et al. 2018). The structure of the pseudo-mask refining module is shown in Figure 3, whose fundamental presumption is that the feature representations of the logits and the CAMs of a same image should follow consistency.

Attention-based Feature Consistency In this paper, an attention-based feature consistency is proposed to regularize the feature similarity of the same image under different transfers, i.e., the logits of the pseudo-mask M_{pseudo} transferred by preliminary segmentation and the CAMs M_{CAM} transferred by CAM-based models. Specifically, to constrain the consistency of features under these two different transfers, we project them to the attention feature space.

Technically, suppose the concatenated pyramid features of a training image extracted by a convolution neural network (CNN) are $F \in \mathbf{R}^{C \times H \times W}$, where C denotes the channels and H, W are the height and width of the feature map. Firstly, the features are flattened along the spatial dimensions, and then two fully connected layers are applied to map the features to the query and key spaces as

$$Q = W_Q F, \qquad K = W_K F, \tag{2}$$

where $W_Q \in \mathbf{R}^{C' \times C}$ and $W_K \in \mathbf{R}^{C' \times C}$ are trainable parameters. In implementation, C' is set to 192.

Then, a spatial-aware attention matrix $A \in \mathbf{R}^{HW \times HW}$ is calculated by the softmax of the inner product of the query and key as

$$A = \text{Softmax}(Q^\intercal K), \tag{3}$$

where A_{ij} measures the similarity between the i-th and j-th feature. Therefore, we treat A as the attention feature space.

Next, the logits and CAMs are down-sampled to the same shape as the feature map. Finally, we use matrix multiplication to project M_{CAM} and M_{pseudo} to the attention feature space as

$$\hat{M}_{\text{pseudo}} = M_{\text{pseudo}} \downarrow A, \tag{4}$$

and

$$\hat{M}_{\text{CAM}} = M_{\text{CAM}} \downarrow A, \tag{5}$$

where $\downarrow$ stands for down-sampling and flattening.

Loss Design Since training images have image-level labels, they can be utilized as supervisions for CNN. Here, the multi-label soft margin loss is adopted as

$$L_{\text{cls}} = -\frac{1}{L} \sum_{i=0}^{L-1} y_i \log \frac{1}{1+e^{-z_i}} + (1-y_i) \log \frac{e^{-z_i}}{1+e^{-z_i}}, \tag{6}$$

where L is the tissue category number. Besides, y_i and z_i are the image-level label and predicted logits of category i.

Besides, since the logits and CAMs transferred to the attention space are correspond to the same training image, they should obey consistency with each other. The feature consistency loss is defined as

$$L_{\text{cons}} = \|\hat{M}_{\text{pseudo}} - \hat{M}_{\text{CAM}}\|_1. \tag{7}$$

However, the consistency loss in Eq. 7 will impose great influence on the feature learning process of the attention space, which may deviate the feature representations from the input space of images. To address this problem, a cross regularization is adopted as

$$L_{\text{cross}} = \|M_{\text{pseudo}} - \hat{M}_{\text{CAM}}\|_1 + \|M_{\text{CAM}} - \hat{M}_{\text{pseudo}}\|_1. \tag{8}$$

The total loss is calculated as the summation of the above three losses, i.e.,

$$L = L_{\text{cls}} + L_{\text{cons}} + L_{\text{cross}}. \tag{9}$$

Precise Segmentation with Refined Masks

Finally, a precise segmentation model is trained using refined pseudo-masks. Three different outputs can be utilized as refined masks. The first two can be calculated by Eq. 4 and Eq. 5. The last one can be obtained by generating CAMs from F and transferring the CAMs to the attention space. It is empirically found that using the last refined CAM can achieve the best performance, which will be validated in the Experimental Result section.

Method	Publication	Tumor IoU	Stroma IoU	Normal IoU	mIoU	fwIoU
HistoSegNet (Chan et al. 2019)	ICCV'19	0.6521	0.3975	0.4610	0.5035	0.5424
SEAM (Wang et al. 2020)	CVPR'20	0.6425	0.4019	0.6981	0.5808	0.5459
SC-CAM (Chang et al. 2020)	CVPR'20	<u>0.7982</u>	**0.7268**	0.6529	<u>0.7260</u>	<u>0.7647</u>
C-CAM (Chen et al. 2022)	CVPR'22	0.6140	0.2535	0.6337	0.5004	0.4673
WSSS-Tissue (Han et al. 2022a)	MIA'22	0.7471	0.5804	0.3327	0.5534	0.6667
OEEM (Li et al. 2022)	MICCAI'22	0.7597	0.6104	**0.7462**	0.7054	0.6983
Training with Synthesized Images	-	0.6707	0.3631	0.5291	0.5210	0.5409
Training with Pseudo-Masks of Real Images	-	0.6794	0.3866	0.6545	0.5735	0.5590
PistoSeg	-	**0.8119**	<u>0.7173</u>	<u>0.7246</u>	**0.7513**	**0.7707**

Table 1: Performance comparison on the WSSS4LUAD dataset. We bold the highest and underline the second highest methods.

Method	Publication	TUM	STR	LYM	NEC	mIoU	fwIoU
HistoSegNet (Chan et al. 2019)	ICCV'19	0.3314	0.4646	0.2905	0.0191	0.2764	0.3719
SEAM (Wang et al. 2020)	CVPR'20	0.7437	0.6216	0.5079	0.4843	0.5894	0.6571
SC-CAM (Chang et al. 2020)	CVPR'20	0.7679	0.7061	0.5802	0.6007	0.6637	0.7158
C-CAM (Chen et al. 2022)	CVPR'22	0.7557	0.6796	0.3100	0.4943	0.5599	0.6680
WSSS-Tissue (Han et al. 2022a)	MIA'22	0.7798	0.7295	0.6098	**0.6687**	0.6970	0.7367
OEEM (Li et al. 2022)	MICCAI'22	0.8021	0.7474	**0.6260**	0.6378	<u>0.7033</u>	<u>0.7544</u>
Fully-Supervised	-	<u>0.8107</u>	<u>0.7486</u>	0.5868	0.5987	0.6862	0.7531
PistoSeg	-	**0.8110**	**0.7504**	<u>0.6184</u>	<u>0.6422</u>	**0.7055**	**0.7589**

Table 2: Performance comparison on the BCSS-WSSS dataset. We bold the highest and underline the second highest methods.

Experimental Result

Dataset

Two weakly-supervised tissue segmentation datasets are adopted in this paper. The first is the WSSS4LUAD dataset (Han et al. 2022b) with 10,091 training images, where 1,181, 1,680, and 1,832 images are only with tumor (TUM), stroma (STR), and normal (NOM) tissues, respectively. The validation dataset comprises 31 small images (the height and width are $200 \sim 500$ pixels) and 9 large images (the height and width are $1500 \sim 5000$ pixels). The test dataset contains 66 small images and 14 large images.

Another is the BCSS-WSSS dataset (Amgad et al. 2019)(Han et al. 2022a) containing 23,422 training images, 3,418 validation images, and 4,986 testing images shaped 224×224. Four tissue categories exist in the BCSS-WSSS dataset, i.e., tumor (TUM), stroma (STR), lymphocytic infiltrate (LYM), and necrosis (NEC). Specifically, 4,738, 2,903, 679, and 1,058 images have only tumor, stroma, lymphocytic infiltrate, and necrosis. It is worth noting that background masks are provided for validation and testing but are unavailable for training for both two datasets.

Experiment Settings and Implementation Details

All experiments are done with an Nvidia RTX 3090 GPU. Codes are implemented with Pytorch 1.12.0 and Pytorch Lightning 1.6.4. Unet++ (Zhou et al. 2018) is selected for preliminary and precise segmentation, with the learning rate being 1e-3 and 5e-4, respectively, and decayed by 0.9 every epoch. Binary cross entropy is employed as the segmentation loss, and AdamW (Loshchilov and Hutter 2019) is adopted to optimize the segmentation models. Training epochs are set to 15. We follow the training settings in Ref. (Wang et al. 2020) for the pseudo-mask refining module with 20 training epochs and 1e-3 of learning rate. We generate 224×224-shaped synthesized image with $n_h = n_w = 7$ and $n_h = n_w = 2$ for WSSS4LUAD and BCSS-WSSS, respectively. As the images are irregularly shaped in WSSS4LUAD, a 224×224 sliding window with 50% overlapping is utilized, and the segmentation outputs of overlapped areas are averaged. For evaluation, category-wise intersection over union (IoU), mean IoU (mIoU), 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.

Comparative Results

Comparative Results on WSSS4LUAD Performance comparison of PistoSeg and other state-of-the-art methods on the WSSS4LUAD dataset is listed in Table 1. The PistoSeg achieves the best mIoU and fwIoU, demonstrating its outstanding performance in general. Although the PistoSeg achieves the second best performance in stroma IoU and normal IoU compared with SC-CAM and OEEM, the performance of SC-CAM on normal IoU is less than 0.7, and the stroma IoU of the OEEM is just slightly higher than 0.6. The weak performance of SC-CAM and OEEM for normal and stroma tissue indicates that they achieve high performance in one category while sacrificing the accuracy of other categories, so their overall performances are poor. In order to better illustrate the effectiveness of training with synthesized images in the PistoSeg, two training strategies are implemented and compared. First, the preliminary segmentation model is trained over the synthesized dataset only. Second, the model is re-trained over the real histopathol-

	Loss	TUM	STR	NOM	mIoU	fwIoU
1	✗✓✗	0.6373	0.2870	0.5369	0.4871	0.4912
2	✗✓✓	0.6249	0.3462	0.6148	0.5286	0.5107
3	✓✓✗	0.2884	0.4493	0.5394	0.4257	0.3616
4	✓✗✓	0.7933	0.6909	0.6658	0.7166	0.7477
5	✓✓✓	**0.8119**	**0.7173**	**0.7246**	**0.7513**	**0.7707**

Table 3: The effect of different loss combinations. The first column is the experiment number. The loss combination from left to right represents L_{cls}, L_{cons}, and L_{cross}.

ogy images with pseudo-masks generated by the first strategy. The performance of these two strategies are denoted by "Training with Synthesized Images" and "Training with Pseudo-Masks of Real Images" in Table 1, respectively. The results show that these two strategies, which does not rely on CAM, can beat some of the current mainstream CAM methods, and the refined PistoSeg can achieve the state-of-the-art performance on WSSS. For the first strategy, the mIoU reaches 0.5210, which exceeds HistoSegNet and C-CAM. The reason for the weak performance of HistoSegNet is that it only uses a series of adjustment and post-processing operations on GradCAM, which leads to poor generalization. Besides, C-CAM is designed for MRI images with solid causal relationships between tissue locations and categories. However, the arbitrary repetition of cells and tissues results in non-causality in histopathology images. For the second strategy, the mIoU reaches 0.5735, which further exceeds WSSS-Tissue. The reason for the weak performance of WSSS-Tissue is that the progressive dropout module drops the most discriminative features extracted by the CNN during the training process, so the model's ability to discriminate different tissues with similar morphological features (e.g., stroma and tumor) is limited, leading to unsatisfactory segmentation performance.

Comparative Results on BCSS-WSSS Table 2 shows the performance of the methods on BCSS-WSSS. Since Han et. al has trained a fully-supervised segmentation network with pixel-level annotations (Han et al. 2022a), we adopt its results denoted as "Fully-Supervised" (we do not train the model ourselves because the annotations of BCSS-WSSS are slightly different from the original BCSS dataset and are not publicly available). The PistoSeg achieves the best or second best performance in all types of IoU. Specifically, the PistoSeg can outperform the fully-supervised model, primarily for LYM and NEC IoU. We argue that the superior performance of PistoSeg for LYM and NEC is because these two tissue types are relatively rare in breast cancer, which causes a long-tail distribution problem. However, since our model utilizes a synthesized dataset, where the proportion of each tissue category can be manually adjusted with re-sampling, the long-tail distribution problem is alleviated. This gives us an insight that the synthesized dataset generation module can be viewed as a form of data augmentation for the segmentation task.

	TUM	STR	NOM	mIoU	fwIoU
CAM	0.7878	0.6875	0.6597	0.7117	0.7430
Refine	**0.8119**	**0.7173**	**0.7246**	**0.7513**	**0.7707**
GCAM	0.7998	**0.7204**	0.6329	0.7177	0.7624
Refine	**0.8063**	0.7053	**0.7510**	**0.7542**	**0.7634**
GCAM++	0.7576	0.6841	**0.7392**	**0.7270**	0.7270
Refine	**0.8029**	**0.7130**	0.6477	0.7212	**0.7615**
SC-CAM	0.7982	**0.7268**	0.6529	0.7260	**0.7647**
Refine	**0.8045**	0.6988	**0.6762**	**0.7265**	0.7575

Table 4: Performance of applying the refining module on different CAM-based methods. GCAM is short for GradCAM.

Ablation Studies

Ablations on Losses of the Refining Module In the pseudo-mask refining module, three different losses, i.e., L_{cls}, L_{cons}, and L_{cross} are utilized. Ablation experiments are performed on the three losses to validate their necessity as listed in Table 3. The first two results show that simply applying L_{cons} will lead to performance degradation, proving that utilizing L_{cons} alone will deviate the feature representations from the input space of images. Similar conclusions can be drawn from experiments 3 and 5. Results of experiments 2 and 5 illustrate that bringing supervision to the image-level labels can enhance the segmentation performance. In addition, not utilizing L_{cons} results in a decrease of mIoU, validating that effective supervision is brought by the consistency of refined logits and CAMs. Generally, the test result achieves the best when all three losses are adopted, demonstrating the reasonability of the loss design.

Applying Pseudo-Mask Refining Module to Different CAMs To verify the effectiveness of our proposed pseudo-mask refining module, we applied it to CAM, GradCAM, GradCAM++, and SC-CAM. Table 4 shows the results. It can be seen that the pseudo-mask refining module can improve the segmentation performance to some extent for different CAM generation methods. For example, all types of IoU can be improved significantly for CAM with pseudo-mask refining. For GradCAM, the mIoU increases by around 0.04 as well. Concerning GradCAM++, although utilizing the module slightly reduces mIoU (less than 0.01), it significantly improves the tumor IoU and stroma IoU, which is also reflected by the improvement of fwIoU. As tumor and stroma appear more frequently in TME, accurate segmentation of them is important. For SC-CAM, adding the mask refining module results in a slight improvement in mIoU, tumor IoU and normal IoU as well. In summary, our proposed pseudo-mask refining module can improve the segmentation performance of existing CAM-based methods.

Ablations on Precise Segmentation Inputs As is stated in the Methodology section, $\hat{M}_{\text{pseudo}}$, $\hat{M}_{\text{CAM}}$, and CNN (abbreviation for the refined CAMs generated from the CNN-extracted features) can be utilized as the input to train the precise segmentation model. The performance of each input is listed in Table 5. The results illustrate that using the last as the input (i.e., the refined CAM generated by CNN) outperforms the others. We argue that it is because the CAM

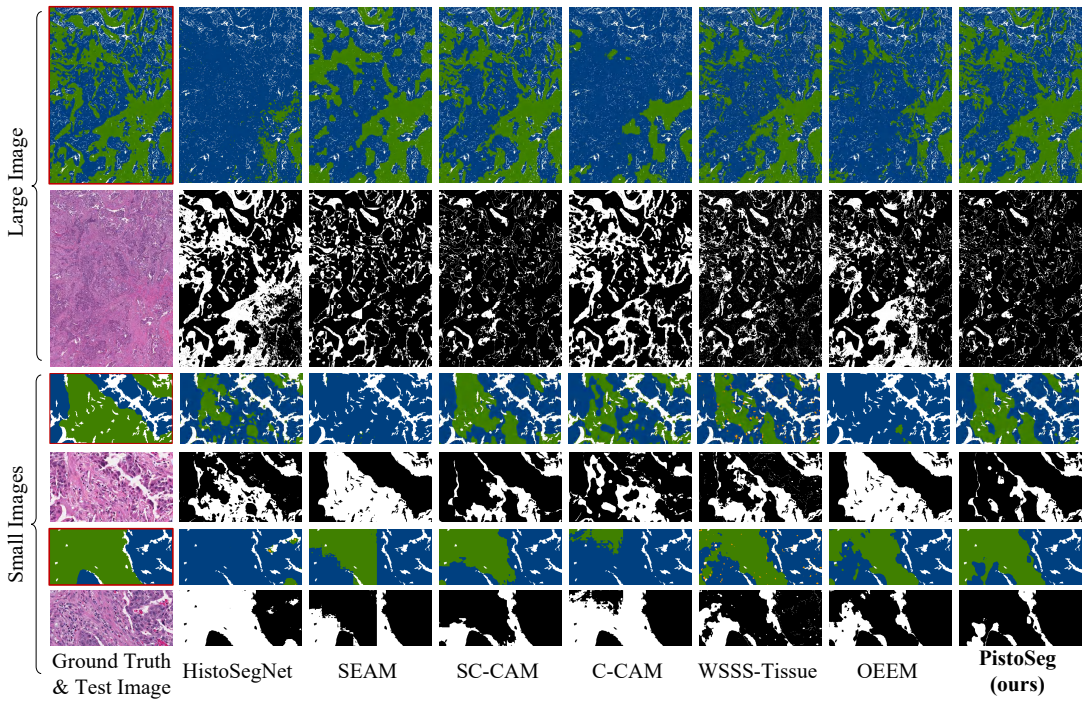

Figure 4: Visualization of segmentation masks on the WSSS4LUAD test dataset of PistoSeg and comparison methods. Blue, green, and yellow represent tumor, stroma, and normal tissue. The black and white masks represent the differences between the ground truth and the predicted mask. Black stands for the same label and white represents difference. Ground truth masks are denoted with red borders.

	TUM	STR	NOM	mIoU	fwIoU
$\hat{M}_{\text{pseudo}}$	0.6913	0.4085	0.5627	0.5542	0.5719
$\hat{M}_{\text{CAM}}$	0.7980	0.6998	0.6915	0.7298	0.7547
CNN	**0.8119**	**0.7173**	**0.7246**	**0.7513**	**0.7707**

Table 5: Performance of the precise segmentation model with different refined pseudo-mask inputs.

generated by the CNN backbone is supervised not only by the image-level labels but also by the attention-based feature consistency between the logits and the CAMs, which pushes the CNN to focus on not only discriminative regions but also boundaries. In addition, as the values of the logits and CAMs are determined mainly by the preliminary segmentation model and the CAM generation model, refining them is more challenging than generating the refined CAM from scratch with the trained CNN.

Qualitative Results

To better understand the merits and drawbacks of PistoSeg and other methods, predicted masks for the test dataset of WSSS4LUAD are shown in Figure 4. As shown in the figure, HistoSegNet tends to misclassify stroma as tumor pixels. In SEAM, the ambiguous boundary phenomenon is severe. As shown in the last image, SC-CAM also faces fuzzy boundaries as the tumor erodes the stroma regions. C-CAM often wrongly predicts stroma tissues as tumors, as shown in the first and the last images. This phenomenon shows the irrationality of C-CAM adopting anatomy casualty for histopathology images. Besides, WSSS-Tissue tends to predict noise normal pixels in tumor and stroma tissues, as shown in the last two images. OEEM usually misclassifies stroma regions and tumor regions as well. Overall, the segmentation masks predicted by PistoSeg are closest to the ground truth annotations.

Conclusion

In this paper, we propose a novel WSSS framework for histopathology images named PistoSeg, which generates pseudo-masks by training a preliminary segmentation model using a synthesized dataset obtained by Mosaic transformation on images with a single tissue type. This approach can avoid the problem of ambiguous boundaries in mainstream CAM-based models of WSSS, giving new insights to the research community of WSSS for histopathology images. Besides, to fully exploit the advantages of existing CAM-based WSSS methods, this paper proposes a pseudo-mask refining module with an attention-based feature consistency. In the future, a potential research direction is to utilize image composition methods or generative models to generate a more true-to-life synthesized dataset, which may push the performance of synthesized dataset training to the same or beyond the CAM-based methods without refining.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61922048 & 62031023), in part by the Shenzhen Science and Technology Project (JCYJ20200109142808034&20220818170353009 and RCYX20200714114523079), and in part by Guangdong Special Support (2019TX05X187).

References

Ahn, J.; and Kwak, S. 2018. 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*, 4981–4990.

Amgad, M.; Elfandy, H.; Hussein, H.; Atteya, L. A.; Elsebaie, M. A.; Abo Elnasr, L. S.; Sakr, R. A.; Salem, H. S.; Ismail, A. F.; Saad, A. M.; et al. 2019. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18): 3461–3467.

Arneth, B. 2019. Tumor microenvironment. *Medicina*, 56(1): 15.

Bearman, A.; Russakovsky, O.; Ferrari, V.; and Fei-Fei, L. 2016. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, 549–565. Springer.

Bochkovskiy, A.; Wang, C.-Y.; and Liao, H.-Y. M. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Chan, L.; Hosseini, M. S.; Rowsell, C.; Plataniotis, K. N.; and Damaskinos, S. 2019. Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10662–10671.

Chang, Y.-T.; Wang, Q.; Hung, W.-C.; Piramuthu, R.; Tsai, Y.-H.; and Yang, M.-H. 2020. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8991–9000.

Chattopadhay, A.; Sarkar, A.; Howlader, P.; and Balasubramanian, V. N. 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 839–847.

Chen, Z.; Tian, Z.; Zhu, J.; Li, C.; and Du, S. 2022. C-CAM: Causal CAM for Weakly Supervised Semantic Segmentation on Medical Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11676–11685.

Han, C.; Lin, J.; Mai, J.; Wang, Y.; Zhang, Q.; Zhao, B.; Chen, X.; Pan, X.; Shi, Z.; Xu, Z.; et al. 2022a. Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. *Medical Image Analysis*, 102487.

Han, C.; Pan, X.; Yan, L.; Lin, H.; Li, B.; Yao, S.; Lv, S.; Shi, Z.; Mai, J.; Lin, J.; et al. 2022b. WSSS4LUAD: Grand Challenge on Weakly-supervised Tissue Semantic Segmentation for Lung Adenocarcinoma.

Hinshaw, D. C.; and Shevde, L. A. 2019. The tumor microenvironment innately modulates cancer progression. *Cancer research*, 79(18): 4557–4566.

Jia, Z.; Huang, X.; Chang, E. I.-C.; and Xu, Y. 2017. Constrained Deep Weak Supervision for Histopathology Image Segmentation. *IEEE Transactions on Medical Imaging*, 36(11): 2376–2388.

Li, Y.; Yu, Y.; Zou, Y.; Xiang, T.; and Li, X. 2022. Online Easy Example Mining for Weakly-supervised Gland Segmentation from Histology Images.

Lin, D.; Dai, J.; Jia, J.; He, K.; and Sun, J. 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3159–3167.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Oh, Y.; Kim, B.; and Ham, B. 2021. 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*, 6913–6922.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626.

Wang, X.; Takaki, S.; Yamagishi, J.; King, S.; and Tokuda, K. 2019. A Vector Quantized Variational Autoencoder (VQ-VAE) Autoregressive Neural F_0 Model for Statistical Parametric Speech Synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 157–170.

Wang, Y.; Zhang, J.; Kan, M.; Shan, S.; and Chen, X. 2020. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12275–12284.

Wazny, K. 2017. "Crowdsourcing" ten years in: A review. *Journal of global health*, 7(2).

Xu, G.; Song, Z.; Sun, Z.; Ku, C.; Yang, Z.; Liu, C.; Wang, S.; Ma, J.; and Xu, W. 2019. Camel: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10682–10691.

Xu, Y.; Zhu, J.-Y.; Chang, E. I.-C.; Lai, M.; and Tu, Z. 2014. Weakly supervised histopathology cancer image segmentation and classification. *Medical Image Analysis*, 18(3): 591–604.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.

Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; and Liang, J. 2018. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. arXiv:1807.10165.