Weakly Supervised Gland Segmentation with Class Semantic Consistency and Purified Labels Filtration

Siyang Feng[1], Huadeng Wang[1], Chu Han[1,2], Zhenbing Liu[1], Hualong Zhang[1], Rushi Lan[1,3*], Xipeng Pan[1*]

[1]School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China
[2]Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangzhou 510080, China
[3]International Joint Research Laboratory of Spatio-temporal Information and Intelligent Location Services, Guilin University of Electronic Technology, Guilin 541004, China
rslan@guet.edu.cn, pxp201@guet.edu.cn

Abstract

Image-level weakly supervised semantic segmentation (WSSS) reduces the dependence on high-quality data annotation, which plays a crucial role in computational pathology. Benefit from the ability to localize the objects with only binary labels, Class Activation Map (CAM) is a widely used method to initial pseudo masks. However, due to the low contrast among different tissues in histopathological images, most existing CAM-based methods perform poorly in gland segmentation. We retrospect this process and find that class consistency and semantic consistency can guide the network to effectively distinguish confusing pixels and generate fine-grained pseudo masks. Specifically, for class consistency, we propose Consistency Correlation Attention (CCA) to encourage the network to focus on the contribution of class features to semantic dependencies. For semantic consistency, we propose Multi-scale Pyramid Fusion Pooling (MPFP) to aggregate coarse-to-fine global semantic information from CAMs at multiple spatial resolutions, thus identifying class localization. Additionally, we introduce a Purified Labels Filtration (PLF) strategy during the segmentation phase to mitigate the noisy supervision signal and improve the segmentation quality of the model. Extensive experiments show that the our method achieves new state-of-the-art results on three publicly available gland datasets. Furthermore, our method demonstrates impressive domain adaptation capability, achieving satisfactory results with only a small portion of samples when faced with unseen domain data.

Introduction

Histopathological image analysis is the gold standard for cancer diagnosis. Quantitative measurement of glands from digitized whole slide images (WSIs) is crucial to determine grades of several cancers, such as colorectal adenocarcinoma (Kim et al. 2020), breast cancer (Rechsteiner, Dietrich, and Varga 2023), prostate cancer (Epstein et al. 2016), and endometrial adenocarcinoma (Stolnicu et al. 2021). Manual identification of these glands not only requires specialized clinical expertise, but also is often time-consuming. While deep learning-based automated gland segmentation can rapidly locate and quantify glandular features, assisting pathologists in clinical diagnosis, therapeutic effect evaluation and prognostic prediction. In recent years, many automated gland segmentation methods (Graham et al. 2019; Wen et al. 2020; Ding et al. 2022; Sun, Huang, and Zheng 2023) require a large amount of elaborate pixel-level annotations to achieve desirable segmentation results. However, these histopathological gland images with dense labels are often difficult to obtain due to the high annotation costs.

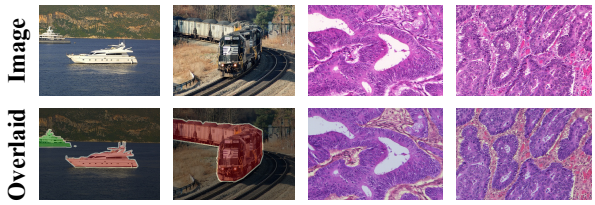

Figure 1: Characteristic difference between natural images and histopathological gland images. The upper row is raw images, and the bottom row is overlaid results with ground truth. It can be seen that the characteristics of gland images are morphological homogeneity and ambiguous boundaries among different tissues.

Is there a solution that can achieve good results without extensive annotation data? The answer is yes. Image-level weakly supervised semantic segmentation (WSSS) can address this issue. According to previous studies (Lin et al. 2014; Han et al. 2022), giving image-level labels for an image can save approximately 60 times the annotation time compared to fine-grained pixel-level labels. Therefore, WSSS holds great promise for gland segmentation tasks. Currently, many WSSS works (Ahn and Kwak 2018; Jiang et al. 2021; Chen et al. 2022b; Yoon et al. 2024; Feng et al. 2024) utilize classification network to generate Class Activation Map (CAM) (Zhou et al. 2016) as pixel-level pseudo masks for subsequent segmentation. Nevertheless, these methods are mostly designed based on natural images and are not suitable for histopathological gland datasets for several reasons. First, as shown in Fig. 1, objects in natural images usually have recognizable boundaries, and different

*Corresponding authors

categories exhibit significant differences in color and morphology. While gland images are often low color contrast, and with similar target shapes, making precise boundary detection extremely difficult, which in turn leads to *class confusion*. Second, natural images typically following some semantic prior rules, e.g., boats generally appear on water, and trains only run on tracks. These rules help model understand and distinguish the semantic relationships between different categories. While in gland images, the distribution of glandular tissues is random and uneven, leading to anomalous semantic correlations like "boats float on tracks", which results in *semantic confusion*.

How to solve the above confusion problems? We think the key idea is to keep inter-class consistency and semantic consistency. From this point of view, we propose two highly-optimized modules called Consistency Correlation Attention (CCA) and Multi-scale Pyramid Fusion Pooling (MPFP) to guide network to focus on class consistency and semantic consistency, respectively. CCA starts from the perspective of category features, it assigns different importance to category pixels based on the correlation of features at different positions to amplify the differences between classes, and thus capturing the long-range semantic dependencies of the same class on a global scale. MPFP constructs a feature fusion pyramid, sequentially aggregating coarse-to-fine semantic features to achieve better class localization. The pyramid's output is used as pixel-level weights to determine the contribution of CAM to the predicted classes, and maximize the reduction of the impact of anomalous semantic correlations on classification. Moreover, we introduce a simple but effective noise suppression strategy named Purified Labels Filtration (PLF) on segmentation stage in order to further eliminate the inevitable noisy pixels caused by limited supervision signals. Our contributions can be summarized as follows:

- We observe that the challenges of gland segmentation arise from class and semantic consistency inherent in histopathological images, which is often overlooked by many current WSSS methods.
- We propose CCA and MPFP to encourage the CAM to keep class and semantic consistency, thereby generating fine-grained pseudo masks. And then we mitigate the damage of inevitable noise via PLF during segmentation phase.
- Our method achieves 87.69%, 80.44% and 77.36% in terms of mIoU, which sets a new state-of-the-art performance with only image-level labels on three publicly avaliable gland segmentation datasets ProG, GlaS, and EBHI, respectively. Experimental results also reveal that our method has great domain adaptation capability, which is beneficial to clinical applications. Our code is available at https://github.com/director87/wsss_gland.

Related Works

Histopathological Gland Segmentation

Recently, many studies have proposed deep learning-based gland segmentation models that balance efficiency and robustness. DCAN (Chen et al. 2016) leveraged multi-level features of glands and employed multi-task learning to segment them and their contours. Xie et al. (Xie et al. 2019) designed two complementary networks to obtain inconsistencies in network predictions and perform online emendation for over-segmented and under-segmented gland regions. GCSBA-Net (Wen et al. 2020) combined Gabor-based encoding and cascade squeeze bi-attention to capture glandular boundary information and designed a hybrid loss to mitigate the impact of imbalanced data. TA-Net (Wang, Xian, and Vakanski 2022) implemented a topology-aware network to describe gland topology and separate densely clustered glands. Wang et al. (Wang et al. 2024) proposed a boundary-enhanced attention via dual-encoder network to restore global features to obtain better segmentation performance. However, these studies heavily rely on plentiful high-quality annotations. Considering that weak labels are easier to obtain in clinical practice compared to pixel-level labels, our goal is to explore WSSS techniques to address the issue of limited annotations in histopathological gland segmentation.

Weakly Supervised Semantic Segmentation

Based on the amount of supervision information provided, common weak annotations including image-level labels (Chang et al. 2020), bounding boxes (Lee et al. 2021), scribbles (Zhang et al. 2024), and points (Laradji et al. 2021). Since image-level labels are the easiest to obtain and CAM (Zhou et al. 2016) has opened new directions for image-level annotation methods, many current WSSS works are based on image-level labels. For example, SEAM (Wang et al. 2020) utilized equivariant regularization to constrain CAMs and refined it by pixel correlation. AMR (Qin et al. 2022) proposed two complementary activation branches and used activation modulation recalibration strategy to enhance low activation areas and then obtained accurate CAMs. Kweon et al. (Kweon, Yoon, and Yoon 2023) found that minimizing inferability could enhance inter-class segmentation, then proposed a framework of adversarial learning and image reconstruction to improve the activation capability of CAMs.

The aforementioned methods are based on natural images, and there are also many valuable works in the field of medical image segmentation. C-CAM (Chen et al. 2022a) utilized causal inference and anatomical prior knowledge to alleviate the problem of ambiguous boundaries in MRI images. MLPS (Han et al. 2022) proposed progressive dropout strategy to pay close attention to non-prominent activated regions. OEEM (Li et al. 2022) encouraged the segmentation network to focus on reliable supervision signals from pseudo masks by online easy example mining strategy. Kuang et al. (Kuang et al. 2023) refined CAMs with cluster-re-activation paradigm and inter-modality self-supervised learning. CBFNet (Du et al. 2024) introduced a CAM-guided cycle-consistency network with a complementary branches fusion module to accurately generate lesion masks from weak labels. Although these methods have achieved excellent results, they perform unsatisfactory on histopathological gland images where inter-class morphology is highly similar. Therefore, we aim to explore the potential relationships between ambiguous categories to find

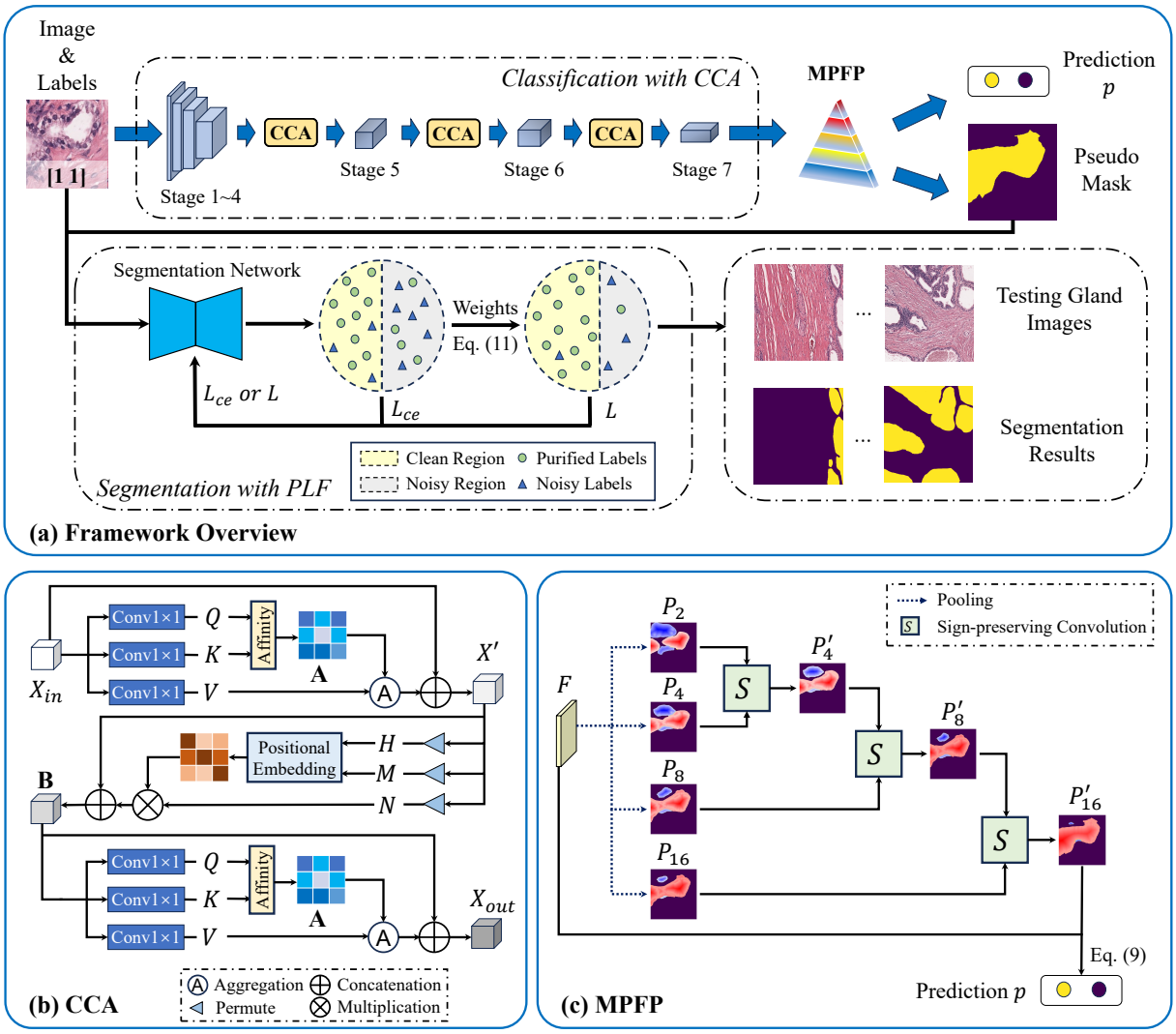

Figure 2: (a) The whole pipeline of our proposed method. (b) The structure of Consistency Correlation Attention (CCA). (c) Illustration of Multi-scale Pyramid Fusion Pooling (MPFP). We visualize the feature maps at each stage of the pyramid, where the red and blue regions represent positive and negative features, respectively. Note that only gland class are demonstrated here.

more effective weakly supervised gland segmentation techniques.

Methodology

Overview of Framework

We depict the whole process of our method as Fig. 2(a). For a training image with image-level labels, we first feed it into the classification network embedded with CCA to generate initial CAM. And then we refine it through MPFP to obtain the classification network's prediction results and fine-grained pseudo mask that contain rich class semantic consistency features. Next, we train the segmentation model based on pseudo mask and PLF strategy in fully supervised manner. Finally, we perform inference with the trained segmentation model on testing images to obtain the final gland segmentation results.

Consistency Correlation Attention

In weakly supervised medical image segmentation tasks, most mainstream methods utilized ResNet and its variants (e.g., ResNet38 (Wu, Shen, and Van Den Hengel 2019)) as classification network to generate CAM. However, these methods overlooked the issue of class consistency in semantic segmentation tasks, namely, that pixels of the same class should have similar features, and pixels of different classes should have distinct features. This oversight makes it difficult for classification network to learn the long-distance dependencies between pixels as decision cues for class prediction, resulting in poor discrimination ability for ambiguous classes and negative impact on the subsequent generation of CAM. To solve these problems, we introduce a novel strat-

egy called CCA. CCA enables a single feature to perceive the features of all other locations, and allows the assignment of different attention weights to the extracted features based on their relevance. Additionally, it remodulates channel feature values to enhance the distinction between object regions and edge features, and amplify the differences between pixels of different categories. This provides significant guidance for the prediction of classification network, which helps improve semantic segmentation.

Fig. 2(b) shows the structure of CCA. First, the input feature $\mathcal{X}_{in} \in \mathbb{R}^{c\times w\times h}$ with the width w and the height h is fed into a criss-cross attention block (Huang et al. 2019) to obtain the feature map $\mathcal{X}'$ with abundant contextual information.

$$\mathcal{X}' = \mathcal{X}_{in} + \sum_{i=0}^{w+h-1} \mathbf{A}_i \boldsymbol{\xi}_i \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{(w+h-1)\times w\times h}$ denotes the attention map calculated by feature maps $\mathcal{Q}$ and $\mathcal{K}$ via affinity operation (Shi et al. 2023b), and $\boldsymbol{\xi} \in \mathbb{R}^{(w+h-1)\times c}$ denotes the feature vector set in feature map $\mathcal{V}$. Second, to explore the spatial positional relationships and long-range dependencies of pixels in each class, we applied three different convolution layers on $\mathcal{X}'$ to generate three feature maps: $\mathcal{H} \in \mathbb{R}^{(w\times h)\times c'}$, $\mathcal{M} \in \mathbb{R}^{c'\times(w\times h)}$, and $\mathcal{N} \in \mathbb{R}^{c\times(w\times h)}$, where $c' = c/4$. Then, the position attention map $\mathbf{B}$ can be aggregated with $\mathcal{X}'$ as follows.

$$\mathbf{B}_j = \sum_{k=1}^{w\times h} \frac{\mathcal{N}_k \cdot \exp(\mathcal{H}_j \cdot \mathcal{M}_k)}{\sum_{j=1}^{w\times h} \exp(\mathcal{H}_j \cdot \mathcal{M}_k)} \tag{2}$$

where j, k denote the pixel coordinates. Finally, we applied the second criss-cross attention operation on $\mathbf{B}$ to further harvest global context information of the whole image. After that, the resulting feature is reshaped to obtain the final output $\mathcal{X}_{out} \in \mathbb{R}^{c\times w\times h}$. To maximize the advantages of class consistency and integrate feature dependencies at various resolutions, we deploy CCA at the last layer from stage 4 to stage 6 of classification network, as shown in Fig. 2(a).

Multi-scale Pyramid Fusion Pooling

Retrospection of Conventional CAM Generation. Before introducing our proposed MPFP, let's start with a brief review of the general process of obtaining CAM. Assume $\mathcal{F} \in \mathbb{R}^{c\times h\times w}$ with c classes and a spatial size of $h \times w$ as the feature map generated by the last convolutional layer of the classification network. Then, we can obtain the image-level class prediction p_c by applying global average pooling (GAP) on $\mathcal{F}$.

$$p_c = \sigma(\frac{1}{h\times w}\sum_{i}^{h}\sum_{j}^{w}\mathcal{F}(i,j)) \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function and $\mathcal{F}(i,j)$ represents the feature vector at the location (i,j). Next, taking the ReLU activation function to $\mathcal{F}$ and then normalizing it in $[0,1]$ for each class, an activation map $\mathbf{C}_n$ of n-th class will be generated as follows.

$$\mathbf{C}_n = \frac{\delta(\mathcal{F}_n)}{\max(\delta(\mathcal{F}_n))} \tag{4}$$

where $\delta(\cdot)$ is the ReLU function.

If we take a step back and carefully consider this process, we can discover that the GAP operation makes the feature contribute equally regardless of their spatial location. And GAP also can mislead the classification network to learn erroneous correlations between image-level and pixel-wise labels due to its dependency on contextual information. Therefore, the generated CAM is tend to capture almost the most discriminative features and activate them, while ignoring the less-correlated regions. Since these regions also contain valuable semantic information, the pixel-level pseudo masks often result in incomplete object representations. This poses a critical disadvantage for accurate gland segmentation.

Multi-scale Pyramid Fusion Pooling (MPFP) Layer. In order to solve the problems of GAP mentioned above, we propose a MPFP layer to generate fine-grained CAM, as shown in Fig. 2(c). Different from the previous studies simply utilizing single-scale operations like softmax on the CAM (Araslanov and Roth 2020; Zhong et al. 2023), our main idea is employing multi-scale to pool CAM and capture a pyramid fusion feature. Inspired by Zhao et al. (Zhao et al. 2017), we successively refine the pooled feature map from a low scale to a high scale by applying multiple sign-preserving convolution layers, to achieve a coarse-to-fine feature fusion.

In multi-label classification, positive predictions indicate the importance of "existence of a class", while the negative predictions also are equally crucial for the model to determine the "non-existence of a class". Therefore, we devise the sign-preserving convolution $\mathcal{S}$ to enhance the feature in both positive and negative directions. The process is formulated as follows.

$$\mathcal{S}(a,b) = \sigma(Conv_{3\times3}(\delta(a) \parallel \delta(b))) \tag{5}$$

where $\parallel$ means the concatenate operation. Then, a feature pyramid $\mathcal{P}_\tau$ is generated by applying $\tau\times\tau$ pooling with different scales τ to the feature map $\mathcal{F}$. Here, the values of τ are 2, 4, 8 and 16. After upsampling the feature pyramid to the resolution of $\mathcal{F}$, we can obtain the feature maps of positive (Pos) and negative (Neg), where $Pos, Neg \in \mathbb{R}^{2\times h\times w}$.

$$Pos_k = \mathcal{S}(\mathcal{P}'_{2^k}, \mathcal{P}_{2^{k+1}}) \tag{6}$$

$$Neg_k = \mathcal{S}(-\mathcal{P}'_{2^k}, -\mathcal{P}_{2^{k+1}}) \tag{7}$$

where $k \in \{1,2,3\}$. And the aggregated feature $\mathcal{P}'_{2^{k+1}}$ is calculated as follows.

$$\begin{aligned}\mathcal{P}'_{2^{k+1}} = &\frac{\delta(\mathcal{P}'_{2^k})\cdot Pos^1_k + \delta(\mathcal{P}_{2^{k+1}})\cdot Pos^2_k}{2} \\ &- \frac{\delta(-\mathcal{P}'_{2^k})\cdot Neg^1_k + \delta(-\mathcal{P}_{2^{k+1}})\cdot Neg^2_k}{2}\end{aligned} \tag{8}$$

where Pos^1_k/Neg^1_k and Pos^2_k/Neg^2_k represent the first and the second channel of Pos_k/Neg_k, respectively. Note that $\mathcal{P}'_{2^k}$ equals to $\mathcal{P}_{2^k}$ when $k = 1$. Finally, we select $\mathcal{P}'_{16} \in$

$\mathbb{R}^{c\times h\times w}$ as the final output of the MPFP and thus obtain the prediction p.

$$p = \sigma(\frac{\sum(\delta(\mathcal{P}'_{16})\cdot\delta(\mathcal{F})) - (\delta(-\mathcal{P}'_{16})\cdot\delta(-\mathcal{F}))}{h\times w}) \quad (9)$$

After decoupling the feature $\mathcal{F}$ with the multi-scale feature pyramid and fusing it from coarse-level to fine-level with MPFP, the regions in $\mathcal{F}$ that unrelated to segment objects will be suppressed, while the highly related regions will be enhanced. Therefore, the generated CAM through the MPFP not only maximizes the localization of the entire target tissue region but also effectively alleviates the issue of semantic confusion caused by morphological homogeneity in histopathological images.

Gland Segmentation with Purified Labels Filtration

Through the exertion of CCA and MPFP, we improve the quality of pseudo masks to a level close to that of ground truth. However, due to the low contrast among different regions and the weak supervisory signal of image-level labels, the pseudo mask inevitably contain a significant amount of noise, which affects segmentation accuracy. Discarding all noise from the pseudo mask is not a wise choice, as the noise region may contain many clean labels. To select the signals that are beneficial for segmentation classes, we propose a denoising strategy called PLF.

During the training process, PLF computes the reliability matrix of each pixel by multiplying a weight map that combines confidence metric and loss value metric on the standard cross-entropy loss $\mathcal{L}_{ce}$. This reliability matrix is then used to filter out potentially clean samples based on their reliability scores. To cover a wider confidence region and better excavate potential objects in difficult samples, we apply the softmax operation ω along the class dimension and use the difference between the maximum and minimum values as the confidence metric. Unlike confidence, the loss value can exhibit the bias of noise towards different classes. Based on this characteristic, we apply the softmax operation ω along the spatial dimension of the loss map and perform mean normalization, assigning higher weights to clean samples within the noise. By combining these two metrics, the segmentation network can be guided to gradually filter out noisy pixels during continuous iterative training and remains potential purified labels. The whole process can be formulated as follows.

$$\mathcal{L}_{ce} = -\frac{1}{\sum_C |\mathcal{G}_C|}\sum_C\sum_{q\in\mathcal{G}_C}\log\mathbf{H}_C(q) \quad (10)$$

$$\mathcal{L} = \begin{cases} \mathcal{L}_{ce}\cdot\frac{\omega(-\mathcal{L}_{ce})\cdot[\max(\omega(\mathbf{H}_{h,w}))-\min(\omega(\mathbf{H}_{h,w}))]}{mean(\omega(-\mathcal{L}_{ce}))} & C>1 \\ \mathcal{L}_{ce}\cdot 1 & C=1 \end{cases} \quad (11)$$

where q is the location of pixel, $\mathbf{H}_C$ is the probability map for the category C, $\mathcal{G}_C$ is a set of locations labeled as the category C in ground truth, and $\mathcal{L}$ is the final segmentation loss. Note that we only applied PLF on images which contain multiple classes, as yielding additional noise in images with only one class is inappropriate.

Experiments

Datasets and Evaluation Metrics

Datasets. We evaluate our method on three publicly available histopathological gland datasets. **(1) Prostate Gland (ProG) Dataset:** ProG (Salvi et al. 2021) consists of 1500 Hematoxylin and Eosin-stained (H&E-stained) histopathological images of 1500×1500 pixels obtained from 150 patients with prostate cancer. We crop all images into patches of 224×224 pixels and split them into a training set (36,000 patches), a validation set (8,244 patches), and a testing set (9,756 patches). For the training set, we generate a one-hot label for each patch to indicate the existence of glandular and non-glandular tissue. **(2) The Gland Segmentation (GlaS) Challenge Dataset:** GlaS (Sirinukunwattana et al. 2017) consists of 165 H&E-stained histopathological images of colorectal adenocarcinoma patients at stage T3 or T4 collected by the University of Warwick, UK. Each image has a resolution of 567-775×430-522 pixels. We divide data into 70 training images, 15 validation images, and 80 testing images. Following the previous work (Li et al. 2022), we use a sliding window method with a side 112 and stride 56 to split each training image into patches. Note that those patches with a large proportion of white background are discarded. **(3) Enteroscope Biopsy Histopathological Image (EBHI) Dataset:** EBHI (Shi et al. 2023a) is a colorectal cancer enteroscope biopsy dataset contains multiple cancer tissue types collected by the Cancer Hospital of China Medical University. All images are with a resolution of 224×224 pixels. We select 1,169 images containing glands and divide them into training, validation, and testing sets in a 7:1:2 ratio.

Evaluation Metrics. We employ mIoU and F1-score, which are commonly used in WSSS and gland segmentation, to evaluate the model's performance. Additionally, we perform a statistical test on the evaluation results using the Independent Samples T-test. When p-value < 0.05, it is considered to indicate a statistical significance.

Experimental Settings

During training, we use data augmentation methods like random flip, random distortion, and Gaussian blur. The classification backbone ResNet38 is pre-trained on ImageNet and its parameters are initialized by Xavier initialization. The initial learning rate is set to 0.01, and the model is trained for 20 epochs under a polynomial decay strategy. For the segmentation phase, we use PSPNet with ResNeSt200 (Zhang et al. 2022), and train the network for 30 epochs with the SGD optimizer. The learning rate is set to 0.005, momentum to 0.9, weight decay to 0.0005, and batch size to 20. All components and networks are written by PyTorch 2.0.1 and trained on one NVIDIA 4090 GPU.

Comparisons with Existing Methods

We compare our method with eight state-of-the-art image-level weakly supervised semantic segmentation methods, including SEAM (Wang et al. 2020), ReCAM (Chen et al. 2022b), AMR (Qin et al. 2022), MLPS (Han et al. 2022),

Method	ProG			GlaS			EBHI		
	mIoU (%)	F1-score (%)	p-value	mIoU (%)	F1-score (%)	p-value	mIoU (%)	F1-score (%)	p-value
Fully Supervised	89.59±0.44	94.51±0.25	-	80.62±0.09	89.27±0.05	-	78.43±0.32	87.89±0.18	-
SEAM (2020)	81.34±0.54	89.71±0.31	***	71.36±0.49	83.28±0.33	***	68.35±0.66	77.81±0.40	***
ReCAM (2022)	82.40±0.08	90.35±0.05	***	56.31±2.53	71.51±2.21	***	47.75±1.26	65.48±0.86	***
AMR (2022)	84.06±0.51	91.34±0.30	***	72.83±0.37	84.26±0.25	***	73.94±0.51	84.90±0.30	***
MLPS (2022)	83.13±0.07	90.79±0.05	***	73.60±0.16	84.79±0.11	***	73.81±0.21	84.85±0.08	***
OEEM (2022)	84.56±0.13	91.63±0.07	**	76.48±0.10	86.67±0.06	***	75.79±0.31	86.26±0.17	*
AME-CAM (2023)	83.29±0.61	90.88±0.39	***	74.09±0.13	85.11±0.08	***	73.98±0.22	84.91±0.13	***
HAMIL (2023)	83.74±0.24	91.14±0.14	***	77.37±0.73	87.24±0.46	***	74.45±0.40	85.34±0.28	**
CBFNet (2024)	83.60±0.50	91.06±0.30	***	76.30±0.26	86.55±0.17	***	74.13±0.30	85.12±0.22	***
Ours	**87.69±0.12**	**93.44±0.09**	-	**80.44±0.05**	**89.16±0.03**	-	**77.36±0.23**	**87.24±0.11**	-

Table 1: Segmentation performance on the ProG, GlaS, and EBHI. The results are reported in "mean±std" format. **Bold** denotes the best results and underline represents the second best. ***, **, and * represent the p-value$<$0.001, p-value$<$0.005, and p-value$<$0.05, respectively.

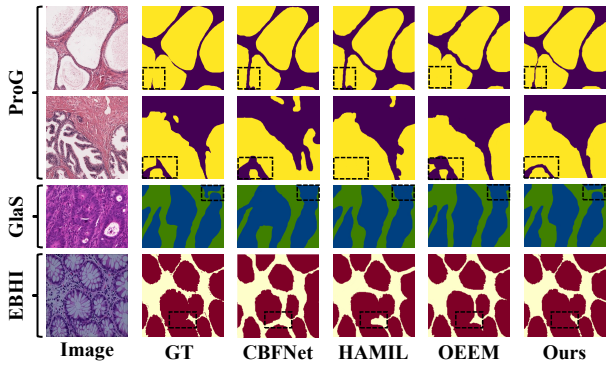

Figure 3: Visualization results of gland segmentation. The gland/non-glandular regions in ProG, GlaS, and EBHI datasets are marked as yellow/indigo, blue/green, and ochre/beige, respectively.

OEEM (Li et al. 2022), AME-CAM (Chen et al. 2023), HAMIL (Zhong et al. 2023), and CBFNet (Du et al. 2024). The results are listed in Tab. 1. On the ProG dataset, our method achieves an mIoU of 87.69% and an F1-score of 93.44%, surpassing the second-best results by 3.13% and 1.81%, respectively, and also with statistical significance. Compared to the ProG, the glands in GlaS and EBHI are less distinguishable from the non-glandular tissues, causing some methods to perform not well on these datasets. However, we are pleased to see that our method overcomes these challenges to establish a new state-of-the-art performance of 80.44% and 77.36% mIoU on GlaS and EBHI, respectively. Furthermore, the comparison results with fully supervised method are encouraging, as they indicate that our method can achieve performance close to or comparable with fully supervised without requiring a large amount of high-quality annotations. Fig. 3 illustrates the visual comparisons on three datasets. From the areas marked with black rectangles, it can be seen that our method achieves more precise segmentation results compared to other methods, even when dealing with clustered or ambiguous gland contours.

Baseline	CCA	MPFP	PLF	mIoU (%)
✓				80.01±0.46
✓	✓			82.90±0.10
✓		✓		86.11±0.15
✓	✓	✓		86.82±0.08
✓	✓	✓	✓	**87.69±0.12**

Table 2: Performance comparisons of each components of our method on the ProG dataset.

Method	mIoU (%)
Criss-cross Attention Block (×1)	84.88±0.21
Criss-cross Attention Block (×2)	86.34±0.23
CCA	**87.69±0.12**
Stage 4 only	85.25±0.14
Stage 5 only	86.43±0.14
Stage 6 only	86.91±0.15
Stage 4 + Stage 5 + Stage 6	**87.69±0.12**

Table 3: Performance comparisons of different settings in CCA on the ProG dataset.

Ablation Studies

We conduct ablation experiments on the ProG dataset. More ablation results and discussions are available in **Supplementary Material**.

To measure the impact of each core component, we conduct a series of ablation experiments. Tab. 2 reports the results. The model incorporated CCA or MPFP achieves 82.90% or 86.11% mIoU, respectively, which is represented improvements of 2.89% and 6.10% compared to the single classification network (Baseline). After combining these two components, the mIoU further increases to 86.82%, sur-

Method	Fusion	mIoU (%)
$\mathcal{P}_2$	-	83.52±0.17
$\mathcal{P}_4$	-	84.79±0.15
$\mathcal{P}_8$	-	85.86±0.15
$\mathcal{P}_{16}$	-	86.77±0.18
$\mathcal{P}_2 + \mathcal{P}_4$	$\mathcal{S}$	85.41±0.13
$\mathcal{P}_2 + \mathcal{P}_4 + \mathcal{P}_8$	$\mathcal{S}$	86.95±0.10
$\mathcal{P}_2 + \mathcal{P}_4 + \mathcal{P}_8 + \mathcal{P}_{16}$	$Average$	87.10±0.10
$\mathcal{P}_2 + \mathcal{P}_4 + \mathcal{P}_8 + \mathcal{P}_{16}$	$\mathcal{S}$	**87.69±0.12**

Table 4: Performance comparisons of different sizes of pooling feature maps in MPFP on the ProG dataset. The "$Average$" denotes simply averaging the feature pyramid. The "$\mathcal{S}$" denotes aggregating the feature pyramid by proposed sign-preserving convolution in Eq. 5.

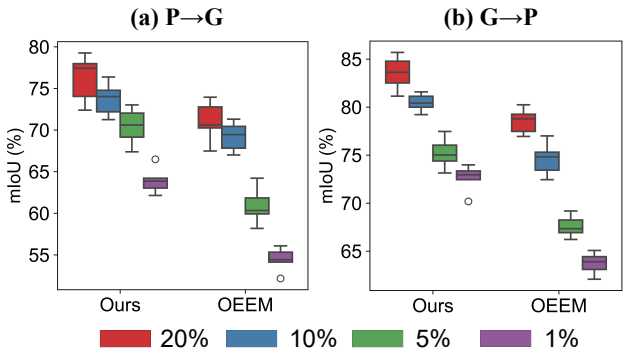

Figure 4: Box-plot of domain adaptation results. (a) P→G denotes the model is first trained on the ProG and then fine-tune it on the GlaS. (b) G→P is the opposite.

passing all existing methods. The whole framework reaches its highest performance of 87.69% mIoU after adding PLF. These results demonstrate the superiority of each core component and how they complement each other to enhance segmentation performance.

Tab. 3 shows the results of CCA under different settings of structure (the first three rows) and applied position (the last four rows). It can be seen that CCA is more effective than criss-cross attention in maintaining global class consistency, which achieves higher mIoU. Moreover, applying CCA to multiple stages of the classification network aggregates more complementary semantic features of different resolutions compared to single-stage deployment, thereby enhancing the network's ability to discern class consistency and improve segmentation results.

To evaluate the performance of different sizes of pooling feature maps and fusion strategies on the feature pyramid, we conduct ablation experiments on MPFP. As shown in Tab. 4, when the size factor τ increases, the receptive field of the features becomes larger, allowing the model to learn more positive and negative samples, thereby improving the mIoU score. Additionally, the more pooling feature maps are fused as the output of the feature pyramid (Eq. 9), the better the model performance. The results also indicate that the performance of fusion using coarse-to-fine sign-preserving convolution ($\mathcal{S}$) is superior to average fusion ($Average$), as the former can effectively retain both global context and feature importance information.

Discussion on Module Compatibility

To verify the compatibility of the proposed CCA and MPFP, we incorporate these two modules with two state-of-the-art methods, i.e., OEEM (Li et al. 2022) and HAMIL (Zhong et al. 2023). The results are listed in Tab. A5 of **Supplementary Material**. For OEEM, the CCA and MPFP improve the segmentation performance by 2.49% and 3.34% mIoU on ProG and GlaS, respectively. For HAMIL, it also achieves 2.62% and 1.11% mIoU improvement after adding our proposed modules. The results indicate that CCA and MPFP are plug-and-play and can improve performance of other WSSS baselines.

Discussion on Domain Adaptation

Unlike other modalities of medical imaging, histopathological images are scarce and more difficult to obtain. Therefore, the domain adaptation ability of the model is definitely important. We conduct experiments on the domain adaptability of the proposed model. Specifically, we first train the model using all data from one dataset (ProG/GlaS), and then fine-tune it using a small portion of data (i.e., 20%, 10%, 5%, and 1%) from another dataset (GlaS/ProG). Fig. 4 shows the performance comparisons of our model from ProG to GlaS (P→G) and from GlaS to ProG (G→P) with another weakly supervised gland segmentation method OEEM (Li et al. 2022). We can observe that our model achieves good performance when fine-tuning on a small amount of training samples. When the sample size decreases to 1%, the performance of OEEM drops substantially, while our model still maintains stable results. This demonstrates that our model has strong adaptability to data from unseen domain, and only requires a few data to achieve satisfactory results.

Conclusion

In this paper, we argue that the key issue in gland segmentation is the confusion between categories and semantic relationships caused by low color contrast among different tissues. Therefore, we propose CCA and MPFP to guide the network in making correct classifications for CAM from the perspectives of class consistency and semantic consistency. Additionally, we introduce PLF to further reduce the impact of residual noisy signals in pseudo masks. Experiments demonstrate the superiority of our method compared with previous state-of-the-art methods with image-level supervision only. Besides, the strong domain adaptation capability of our model also offers new prospects for clinical applications.

Acknowledgments

This work was supported in part by the Guangxi Natural Science Foundation (No. 2024GXNSFFA010014) and the National Natural Science Foundation of China (Nos. 62172120, 82360356 and 62362014).

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.

Araslanov, N.; and Roth, S. 2020. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4253–4262.

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.

Chen, H.; Qi, X.; Yu, L.; and Heng, P.-A. 2016. DCAN: deep contour-aware networks for accurate gland segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2487–2496.

Chen, Y.-J.; Hu, X.; Shi, Y.; and Ho, T.-Y. 2023. Ame-cam: Attentive multiple-exit cam for weakly supervised segmentation on mri brain tumor. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 173–182. Springer.

Chen, Z.; Tian, Z.; Zhu, J.; Li, C.; and Du, S. 2022a. 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*, 11676–11685.

Chen, Z.; Wang, T.; Wu, X.; Hua, X.-S.; Zhang, H.; and Sun, Q. 2022b. Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 969–978.

Ding, S.; Wang, H.; Lu, H.; Nappi, M.; and Wan, S. 2022. Two path gland segmentation algorithm of colon pathological image based on local semantic guidance. *IEEE Journal of Biomedical and Health Informatics*, 27(4): 1701–1708.

Du, W.; Huo, Y.; Zhou, R.; Sun, Y.; Tang, S.; Zhao, X.; Li, Y.; and Li, G. 2024. Consistency label-activated region generating network for weakly supervised medical image segmentation. *Computers in Biology and Medicine*, 173: 108380.

Epstein, J. I.; Zelefsky, M. J.; Sjoberg, D. D.; Nelson, J. B.; Egevad, L.; Magi-Galluzzi, C.; Vickers, A. J.; Parwani, A. V.; Reuter, V. E.; Fine, S. W.; et al. 2016. A contemporary prostate cancer grading system: a validated alternative to the Gleason score. *European Urology*, 69(3): 428–435.

Feng, S.; Chen, J.; Liu, Z.; Liu, W.; Wang, Z.; Lan, R.; and Pan, X. 2024. Mining gold from the sand: Weakly supervised histological tissue segmentation with activation relocalization and mutual learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 414–423. Springer.

Graham, S.; Chen, H.; Gamper, J.; Dou, Q.; Heng, P.-A.; Snead, D.; Tsang, Y. W.; and Rajpoot, N. 2019. MILD-Net: Minimal information loss dilated network for gland instance segmentation in colon histology images. *Medical Image Analysis*, 52: 199–211.

Han, C.; Lin, J.; Mai, J.; Wang, Y.; Zhang, Q.; Zhao, B.; Chen, X.; Pan, X.; Shi, Z.; Xu, Z.; et al. 2022. Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. *Medical Image Analysis*, 80: 102487.

Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; and Liu, W. 2019. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 603–612.

Jiang, P.-T.; Han, L.-H.; Hou, Q.; Cheng, M.-M.; and Wei, Y. 2021. Online attention accumulation for weakly supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 7062–7077.

Kim, B.-h.; Kim, J. M.; Kang, G. H.; Chang, H. J.; Kang, D. W.; Kim, J. H.; Bae, J. M.; Seo, A. N.; Park, H. S.; Kang, Y. K.; et al. 2020. Standardized pathology report for colorectal cancer. *Journal of Pathology and Translational Medicine*, 54(1): 1–19.

Kuang, Z.; Yan, Z.; Zhou, H.; and Yu, L. 2023. Cluster-re-supervision: bridging the gap between image-level and pixel-wise labels for weakly supervised medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 27(10): 4890–4901.

Kweon, H.; Yoon, S.-H.; and Yoon, K.-J. 2023. Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11329–11339.

Laradji, I.; Rodriguez, P.; Manas, O.; Lensink, K.; Law, M.; Kurzman, L.; Parker, W.; Vazquez, D.; and Nowrouzezahrai, D. 2021. A weakly supervised consistency-based learning method for covid-19 segmentation in ct images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2453–2462.

Lee, J.; Yi, J.; Shin, C.; and Yoon, S. 2021. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2643–2652.

Li, Y.; Yu, Y.; Zou, Y.; Xiang, T.; and Li, X. 2022. Online easy example mining for weakly-supervised gland segmentation from histology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 578–587. Springer.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 740–755. Springer.

Qin, J.; Wu, J.; Xiao, X.; Li, L.; and Wang, X. 2022. Activation modulation and recalibration scheme for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2117–2125.

Rechsteiner, A.; Dietrich, D.; and Varga, Z. 2023. Prognostic relevance of mixed histological subtypes in invasive breast

carcinoma: a retrospective analysis. *Journal of Cancer Research and Clinical Oncology*, 149(8): 4967–4978.

Salvi, M.; Bosco, M.; Molinaro, L.; Gambella, A.; Papotti, M.; Acharya, U. R.; and Molinari, F. 2021. A hybrid deep learning approach for gland segmentation in prostate histopathological images. *Artificial Intelligence in Medicine*, 115: 102076.

Shi, L.; Li, X.; Hu, W.; Chen, H.; Chen, J.; Fan, Z.; Gao, M.; Jing, Y.; Lu, G.; Ma, D.; et al. 2023a. EBHI-Seg: A novel enteroscope biopsy histopathological hematoxylin and eosin image dataset for image segmentation tasks. *Frontiers in Medicine*, 10: 1114673.

Shi, T.; Ding, X.; Zhou, W.; Pan, F.; Yan, Z.; Bai, X.; and Yang, X. 2023b. Affinity feature strengthening for accurate, complete and robust vessel segmentation. *IEEE Journal of Biomedical and Health Informatics*, 27(8): 4006–4017.

Sirinukunwattana, K.; Pluim, J. P.; Chen, H.; Qi, X.; Heng, P.-A.; Guo, Y. B.; Wang, L. Y.; Matuszewski, B. J.; Bruni, E.; Sanchez, U.; et al. 2017. Gland segmentation in colon histology images: The glas challenge contest. *Medical Image Analysis*, 35: 489–502.

Stolnicu, S.; Park, K. J.; Kiyokawa, T.; Oliva, E.; McCluggage, W. G.; and Soslow, R. A. 2021. Tumor typing of endocervical adenocarcinoma: contemporary review and recommendations from the International Society of Gynecological Pathologists. *International Journal of Gynecological Pathology*, 40: S75–S91.

Sun, M.; Huang, W.; and Zheng, Y. 2023. Instance-aware diffusion model for gland segmentation in colon histology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 662–672. Springer.

Wang, H.; Xian, M.; and Vakanski, A. 2022. Ta-net: Topology-aware network for gland segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1556–1564.

Wang, H.; Yu, J.; Li, B.; Pan, X.; Liu, Z.; Lan, R.; and Luo, X. 2024. Gland segmentation via dual encoders and boundary-enhanced attention. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2345–2349. IEEE.

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.

Wen, Z.; Feng, R.; Liu, J.; Li, Y.; and Ying, S. 2020. Gcsba-net: Gabor-based and cascade squeeze bi-attention network for gland segmentation. *IEEE Journal of Biomedical and Health Informatics*, 25(4): 1185–1196.

Wu, Z.; Shen, C.; and Van Den Hengel, A. 2019. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90: 119–133.

Xie, Y.; Lu, H.; Zhang, J.; Shen, C.; and Xia, Y. 2019. Deep segmentation-emendation model for gland instance segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 469–477. Springer.

Yoon, S.-H.; Kwon, H.; Kim, H.; and Yoon, K.-J. 2024. Class tokens infusion for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3595–3605.

Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Lin, H.; Zhang, Z.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; et al. 2022. Resnest: Split-attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2736–2746.

Zhang, X.; Zhu, L.; He, H.; Jin, L.; and Lu, Y. 2024. Scribble hides class: Promoting scribble-based weakly-supervised semantic segmentation with its class label. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7332–7340.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2881–2890.

Zhong, L.; Wang, G.; Liao, X.; and Zhang, S. 2023. HAMIL: High-resolution activation maps and interleaved learning for weakly supervised segmentation of histopathological images. *IEEE Transactions on Medical Imaging*, 42(10): 2912–2923.

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.