Mining Gold from the Sand: Weakly Supervised Histological Tissue Segmentation with Activation Relocalization and Mutual Learning

Siyang Feng[1], Jiale Chen[1], Zhenbing Liu[1], Wentao Liu[2], Zimin Wang[1], Rushi Lan[3(✉)], and Xipeng Pan[1,4(✉)]

[1] School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China
pxp201@guet.edu.cn
[2] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, 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
[4] Guangxi Key Laboratory of Image and Graphic Intelligent Processing, Guilin University of Electronic Technology, Guilin 541004, China

Abstract. Class activation maps- (CAMs-) based image-level weakly supervised tissue segmentation has became a popular research topic due to the advantage of its low annotation cost. However, there are still two challenges exist in this task: (1) low-quality pseudo masks generation, and (2) training with noisy label supervision. To address these issues, we propose a novel weakly supervised segmentation framework with Activation Relocalization and Mutual Learning (ARML). First, we integrate an Activation Relocalization Scheme (ARS) into classification phase to more accurately cover the useful areas in initial CAMs. Second, to deal with the inevitably noisy annotations in pseudo masks generated by ARS, we propose a noise-robust mutual learning segmentation model. The model promotes peer networks to capture different characteristics of the outputs, and two noise suppression strategies namely samples weighted voting (SWV) and samples relation mining (SRM) are introduced to excavate the potential credible information from noisy annotations. Extensive experiments on BCSS and LUAD-HistoSeg datasets demonstrate that our proposed ARML exceeds many state-of-the-art weakly supervised semantic segmentation methods, which gives a new insight for tissue segmentation tasks. The code is available at: https://github.com/director87/ARML.

Keywords: Weakly supervised tissue segmentation · Activation relocalization · Noise suppression · Mutual learning

1 Introduction

Histological assessment of Hematoxylin and Eosin- (H&E-) stained tissue specimens remains the gold standard for cancer diagnosis [14]. With the rapid development of deep learning, automatic tissue segmentation has became an essential part in computational pathology. However, it is time-consuming and costly to obtain the dense pixel-level annotations of the whole slide image (WSI) with giga-pixel [6]. Therefore, utilizing a weakly supervised semantic segmentation (WSSS) method with only image-level to reduce the annotation expenses has became a research hotspot.

Currently, most mainstream WSSS studies achieve their goals based on class activation maps (CAMs) [20]. Although CAM can take advantage of localization clues brought by classification network to achieve pseudo masks generation, it still inevitably has the shortage of failing to extract precise target boundaries and makes the generated pseudo masks incomplete, which brings a great challenge to segmentation task. To cope with this issue, several methods devote to refine the quality of CAM to obtain fine-grained pseudo masks [2,6,13,4,19]. HistoSegNet [2] used a series of post-processing steps to correct the target boundaries. MLPS [6] creatively proposed progressive dropout attention to discard the highlighted activations, and push model to focus on non-predominant features. AME-CAM [4] aggregated CAMs with different resolutions and optimized them by contrastive learning. However, the refining process of these methods is complex and heavily relies on the discriminative activation regions. In contrast, we propose the Activation Relocalization Scheme (ARS) to capture contextual information from spatial responses in local feature maps and reactivate the channel-wise minor regions, which can adaptively expand the useful areas of CAMs.

Since there are unavoidable noisy annotations contained in pseudo masks, how to deal with them to obtain more accuracy segmentation results during semantic segmentation stage is another existing challenge. URN [10] reweighted the segmentation loss to suppress noise by uncertainty estimation. OEEM [11] proposed online easy examples mining to filter out the credible predictions. However, performances of these methods are still limited since the single segmentation network cannot deal with noisy labels very well without extra supervision. From this point of view, we introduce Mutual Learning (ML) between three networks during the segmentation phase to mine the useful signals in noisy annotations. A samples weighed voting (SWV) strategy is utilized to determine potential reliable labels and suppress the disagreement ones, and then make one network serve as teacher to supervise the other two networks via samples relation mining (SRM) strategy that further restrains the noisy labels.

The major contributions of our work are in three aspects:

- We propose a weakly supervised histological tissue segmentation framework with Activation Relocalization and Mutual Learning (ARML) to overcome the existing challenges.
- ARS is proposed to obtain more accuracy pseudo masks, and Mutual Learning between three networks with SWV and SRM is proposed to effectively

excavate the reliable samples from noisy annotations for precise segmentation.
- Our method obtains the state-of-the-art WSSS performance on the BCSS and LUAD-HistoSeg benchmarks.

2 Methodology

2.1 Framework Overview

The framework of our proposed ARML is shown in Fig. 1. For pseudo masks generation phase, we first put the input image I into the classification network F_{cls} supervised by image-level labels Y to extract the initial CAMs. Then, we introduce the ARS on feature maps to refine it and finally obtain the pixel-level pseudo mask M. The multi-label soft margin loss $\mathcal{L}_{cls}$ is leveraged to optimize the classification prediction. For segmentation phase, we utilize three networks F_a, F_b and F_c with the same structure and training them concurrently. Due to the unavoidable noisy signals in M, we propose two denoising strategies to guide network to select reliable information in the confusing regions. Since the strategies exerted on the three networks are similar, we take F_a as an example to introduce in the following sections of paper for the sake of simplicity. First, we modify the standard cross entropy loss by giving a different weight coefficient for different predicted samples according to the "opinions" of F_b and F_c. Second, we design a new loss function to mine the relationship of output logits between F_a and other two networks, thus further make each network noise-robust.

2.2 Activation Relocalization Scheme

Conventional CAMs generation method suffers from incomplete object regions, which will sacrifice the performance of semantic segmentation. To address this issue, we propose the Activation Relocalization Scheme (ARS) to refine initial CAMs to be more task-friendly. As shown in Fig. 1(a), given a local feature map $f \in \mathbb{R}^{C\times H\times W}$, the ARS first generate a query map $Q \in \mathbb{R}^{C'\times H\times W}$, a key map $K \in \mathbb{R}^{C'\times H\times W}$, and a value map $V \in \mathbb{R}^{C\times H\times W}$ by three 1×1 convolutional layers. Here, the channel number C' is less than C for dimension reduction. Then, attention maps $A \in \mathbb{R}^{(H+W-1)\times H\times W}$ are obtained from Q and K via *similarity* operation. This operation aims to calculate the similarity between different pixels in the spatial dimension, which is defined as:

$$s_{i,j} = Q_j \mathbf{\Phi}_{i,j}^{\top}, \tag{1}$$

where $s_{i,j} \in \mathbb{R}^{(H+W-1)\times H\times W}$ denotes the similarity correlation, $Q_j \in \mathbb{R}^{C'}$ is the feature vector of Q at position j, and $\mathbf{\Phi}_{i,j} \in \mathbb{R}^{C'}$ is the i-th element of $\mathbf{\Phi}_j \in \mathbb{R}^{(H+W-1)\times C'}$ extracted from K, therein $i = [1, ..., |\mathbf{\Phi}_j|]$. Next, we combine A and feature vectors $\mathbf{\Psi} \in \mathbb{R}^{(H+W-1)\times C}$ in V by *fusion* operation defined as

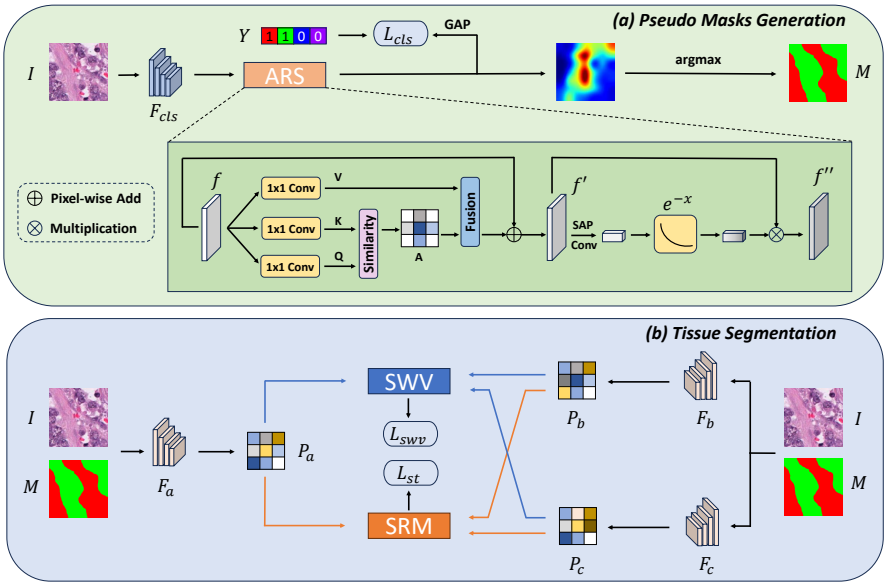

Fig. 1. Overall structure of our proposed ARML. (a) Pseudo masks generation phase with Activation Relocalization Scheme. (b) Tissue segmentation phase with Mutual Learning and two denoising strategies. For convenience, we regard F_a as the final segmentation model and only show the noise suppress loss functions for illustration.

Eq. (2) to obtain the enhanced contextual features.

$$f' = f \oplus \sum_{k=0}^{H+W-1} A_k \mathbf{\Psi}_k. \tag{2}$$

After fusion, we exploit a reactivation attention to restrain the most sensitive features and expand the non-discrimination features in f'.

$$f'' = f' \otimes \mathcal{E}(Conv(P^s_{avg}(f'))), \tag{3}$$

where $\mathcal{E}$ is the exponential function e^{-x} to redistribute the features to highlight the minor responses in the channel dimension, $Conv$ is a 7×7 convolutional layer, and P^s_{avg} is the spatial average pooling (SAP) operation. In the end, the output relocalized feature maps f'' aggregates the pixel-wise context representations in spatial dimension and recalibrate activation values of each category in channel dimension, which are more credible for semantic segmentation.

2.3 Samples Weighted Voting

To deal with noisy annotations, some previous denoising works leveraged the loss functions to filter the hard samples within confusing areas for segmentation

stage [12,7]. However, these methods ignored that potential clean annotations also exist in low quality regions. The accuracy of segmentation model will be limited if simply discarding these useful information. To address this issue, we propose the $\mathcal{L}_{swv}$ to reweight predicted pixels based on the outputs of two networks by utilizing a voting approach. In details, we denote the logits map generated by three networks as P_a, P_b, and P_c. Then, we assign a weight for pixels in P_a according to the vote results. If a predict value in P_b is identical to the one in P_c, then this pixel in P_a will be seen as a credible one and will be given a higher weight. Conversely, the pixel will be given a lower weight if there is a divergence between P_b and P_c. Finally, the $\mathcal{L}_{swv} \in \mathbb{R}^{H \times W}$ can be formulated as follows:

$$\mathcal{L}_{swv}(P_a, M) = \begin{cases} -\omega \cdot \sum_{i}^{H} \sum_{j}^{W} \log \frac{e^{P_a(M_{i,j},i,j)}}{\sum_{c=0}^{C} e^{P_a(c,i,j)}} & P_b(i,j) = P_c(i,j) \\ -\frac{1}{\omega} \cdot \sum_{i}^{H} \sum_{j}^{W} \log \frac{e^{P_a(M_{i,j},i,j)}}{\sum_{c=0}^{C} e^{P_a(c,i,j)}} & P_b(i,j) \neq P_c(i,j) \end{cases}, \quad (4)$$

where M represents the pseudo mask, (i, j) means the coordinate of a pixel in logits map, $\omega \in [1, +\infty)$ denotes the weight coefficient, and C is the total number of categories.

2.4 Samples Relation Mining

To further explore the relationship among the three logits maps and suppress more incredible responses, we embed the samples relation mining strategy into our mutual learning framework as well, where one network serves as a teacher to supervise the other two students. Here, two types of loss functions are proposed to mine the relations between these models.

The first loss $\mathcal{L}_{pt}$ based on *point-wise* can be realize as Eq. (5). The motivation of this loss is that convolutional neural network is tend to fall into overfitting for easy learning samples. Thus we apply the softmax operation sm to prediction maps to make results of student models more stable and robust to against noisy signals.

$$\mathcal{L}_{pt} = \frac{1}{2}\Big(\frac{1}{N}\sum_{i=1}^{N}\big(\|sm(P_{ai}) - sm(P_{bi})\|_2 + \|sm(P_{ai}) - sm(P_{ci})\|_2\big)\Big). \quad (5)$$

Different from *point-wise* loss $\mathcal{L}_{pt}$ only transfers the intra-relation of samples across different outputs, the second *structure-wise* loss $\mathcal{L}_{st}$ also considers the inter-relation of different samples in the same output, which are compounded of binary distance and ternary angle relations. For binary distance relations, given a pair of samples $\{(\theta_i, \theta_j)|i \neq j\}$ from 2-tuples sample set $\mathbf{\Theta}^2$, we first calculate the Euclidean distance $\mathcal{D}$ of these two samples. Next, we use $\mathcal{D}$ to measure in both teacher model and student models in our framework, then we get the distance loss $\mathcal{L}_{dist}$ expressed as Eq. (6). This loss transfers the relationship of

samples by penalizing distance differences between output representations, which encourages the student models to excavate the clean samples rather than noisy ones.

$$\mathcal{L}_{dist} = \frac{1}{2} \sum_{(\theta_i,\theta_j)\in\mathbf{\Theta}^2} \Big(\mathcal{H}\big(\mathcal{D}(P_{ai}, P_{aj}), \mathcal{D}(P_{bi}, P_{bj})\big) + \mathcal{H}\big(\mathcal{D}(P_{ai}, P_{aj}), \mathcal{D}(P_{ci}, P_{cj})\big)\Big), \tag{6}$$

where $\mathcal{H}$ is Huber loss (see [9] for detailed definition). Similar to distance relations, given a triplet of samples $\{(\theta_i, \theta_j, \theta_k)|i \neq j \neq k\}$ from 3-tuples sample set $\mathbf{\Theta}^3$, the ternary angle loss $\mathcal{L}_{ang}$ can be calculated as follows:

$$\begin{aligned} \mathcal{L}_{ang} = \frac{1}{2} \sum_{(\theta_i,\theta_j,\theta_k)\in\mathbf{\Theta}^3} \Big(&\mathcal{H}\big(\mathcal{A}(P_{ai}, P_{aj}, P_{ak}), \mathcal{A}(P_{bi}, P_{bj}, P_{bk})\big) \\ &+ \mathcal{H}\big(\mathcal{A}(P_{ai}, P_{aj}, P_{ak}), \mathcal{A}(P_{ci}, P_{cj}, P_{ck})\big)\Big), \end{aligned} \tag{7}$$

where $\mathcal{A} = \cos\left\langle \frac{\theta_i - \theta_j}{\|\theta_i - \theta_j\|_2}, \frac{\theta_k - \theta_j}{\|\theta_k - \theta_j\|_2} \right\rangle$ denotes the angular differences between samples. Since angle contains more high-order features than distance, it gives more guidance to models in learning confusing pixels. Finally, we combine the distance-wise loss and angle-wise loss together to get the second loss $\mathcal{L}_{st}$. For an image, the primary information lies in the structure of the data embedding space which consists of numerous samples, thus the context meaning of individual sample is limited. Therefore, we suggest that $\mathcal{L}_{st}$ will get a better performance and choose it as our final loss for samples relation mining strategy.

$$\mathcal{L}_{st} = \alpha\mathcal{L}_{dist} + \beta\mathcal{L}_{ang}. \tag{8}$$

In the end, the final segmentation loss $\mathcal{L}$ for each network can be formulated as follows, where λ is the hyperparameter to balance two loss functions:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{swv} + \lambda\mathcal{L}_{st}. \tag{9}$$

3 Experiments and Results

3.1 Dataset

BCSS contains 23,422 training images, 3,418 validation images and 4,986 testing images cropped from 151 H&E-stained WSIs of breast cancer [1,6]. There are four foreground tissue categories in dataset, i.e., tumor (TUM), stroma (STR), lymphocytic infiltrate (LYM) and necrosis (NEC). To perform weakly supervised method, each training image only have image-level annotations using one-hot encoding vectors. Note that background masks are only provided for validation and testing set and unavailable for training set.

LUAD-HistoSeg includes 17,285 H&E-stained lung adenocarcinoma images with four tissue categories namely tumor epithelial (TE), necrosis (NEC), lymphocyte (LYM) and tumor-associated stroma (TAS) [6]. Following the original partition, we use 16,678 images with image-level annotations for training, 300 and 307 images with pixel-level annotations for validation and testing, respectively.

Table 1. Comparison with state-of-the-art weakly supervised methods. The results are reported in mean ± std. † means statistical significant with the best performance of existed methods ($p < 0.05$ under two-tailed T-test).

BCSS Method	Backbone	TUM	STR	LYM	NEC	mIoU	Dice	ACC
HistoSegNet [2]	VGG16	33.14 ± 1.61	46.46 ± 1.37	29.05 ± 2.93	28.83 ± 3.42	34.37 ± 0.96	50.04 ± 0.78	56.41 ± 0.77
ReCAM [5]	ResNet50	74.34 ± 0.61	68.37 ± 0.89	47.17 ± 2.07	56.91 ± 3.35	61.70 ± 0.42	75.76 ± 0.29	81.15 ± 0.45
GradCAM++ [3]	ResNet38	74.12 ± 0.59	67.61 ± 0.81	54.68 ± 1.43	57.18 ± 1.94	63.39 ± 0.35	77.31 ± 0.24	81.30 ± 0.66
AME-CAM [4]	ResNet50	77.29 ± 0.55	72.20 ± 0.31	56.98 ± 1.33	60.44 ± 0.71	66.73 ± 0.51	79.75 ± 0.38	83.90 ± 0.30
AMR [13]	ResNet50	78.75 ± 0.34	72.12 ± 0.15	59.31 ± 0.84	57.80 ± 1.30	66.99 ± 0.44	79.91 ± 0.29	84.44 ± 0.18
OEEM [11]	ResNet38	79.11 ± 0.37	72.88 ± 0.79	54.34 ± 1.61	63.07 ± 2.43	67.35 ± 0.58	80.11 ± 0.42	84.41 ± 0.24
MLPS [6]	ResNet101	78.53 ± 0.60	71.74 ± 0.69	**60.71 ± 0.52**	60.51 ± 1.11	67.87 ± 0.45	80.62 ± 0.31	84.26 ± 0.38
TPRO [19]	MixTransformer	**80.10 ± 0.59**	73.34 ± 0.21	56.26 ± 1.06	64.26 ± 1.39	68.49 ± 0.24	80.95 ± 0.19	85.03 ± 0.16
ARML (Ours)	MixTransformer	79.97 ± 0.41	73.09 ± 0.38	57.72 ± 0.95	65.02 ± 0.33†	68.95 ± 0.39†	81.33 ± 0.22†	85.08 ± 0.44
ARML (Ours)	ResNet38	78.87 ± 0.55	73.18 ± 0.73	58.85 ± 1.04	66.55 ± 0.36†	69.36 ± 0.41†	81.68 ± 0.25†	84.88 ± 0.67
ARML (Ours)	ResNeSt101	79.20 ± 0.30	**73.83 ± 0.17**†	60.25 ± 0.14	**68.96 ± 0.44**†	**70.56 ± 0.19**†	**82.48 ± 0.10**†	**85.63 ± 0.17**†
LUAD-HistoSeg Method	Backbone	TE	NEC	LYM	TAS	mIoU	Dice	ACC
HistoSegNet [2]	VGG16	45.59 ± 0.83	46.30 ± 1.43	58.28 ± 1.95	50.82 ± 0.87	50.25 ± 0.66	63.43 ± 0.51	66.82 ± 0.46
ReCAM [5]	ResNet50	73.81 ± 0.46	58.72 ± 0.77	67.24 ± 1.61	66.77 ± 1.04	66.63 ± 0.53	79.85 ± 0.36	81.80 ± 0.57
GradCAM++ [3]	ResNet38	73.48 ± 0.41	60.02 ± 0.68	68.97 ± 0.92	66.53 ± 1.13	67.25 ± 0.48	80.32 ± 0.35	81.83 ± 0.62
AME-CAM [4]	ResNet50	74.64 ± 0.32	67.58 ± 0.53	69.77 ± 1.18	68.47 ± 0.97	70.11 ± 0.66	82.40 ± 0.54	83.07 ± 0.23
AMR [13]	ResNet50	74.89 ± 0.18	67.47 ± 0.37	70.55 ± 0.75	68.67 ± 0.81	71.45 ± 0.39	83.59 ± 0.28	83.29 ± 0.41
OEEM [11]	ResNet38	76.96 ± 0.27	74.71 ± 0.37	72.30 ± 2.20	71.30 ± 1.52	73.92 ± 0.44	84.99 ± 0.30	85.13 ± 0.89
MLPS [6]	ResNet101	77.68 ± 0.51	76.95 ± 0.58	72.40 ± 0.84	71.81 ± 0.77	74.71 ± 0.41	85.50 ± 0.30	85.45 ± 0.81
TPRO [19]	MixTransformer	75.85 ± 0.47	**81.94 ± 0.95**	74.66 ± 1.25	71.27 ± 0.46	75.90 ± 0.44	86.16 ± 0.28	85.56 ± 0.19
ARML (Ours)	MixTransformer	77.23 ± 0.43	80.40 ± 0.72	**75.43 ± 1.53**†	71.43 ± 0.66	76.13 ± 0.47	86.38 ± 0.29	85.62 ± 0.37
ARML (Ours)	ResNet38	77.87 ± 0.41	81.89 ± 0.62	75.21 ± 0.42†	72.31 ± 0.13†	76.82 ± 0.37†	86.84 ± 0.22†	86.13 ± 0.28†
ARML (Ours)	ResNeSt101	**78.54 ± 0.28**†	80.34 ± 0.27	75.29 ± 0.45†	**74.29 ± 0.48**†	**77.11 ± 0.33**†	**87.01 ± 0.22**†	**86.42 ± 0.24**†

Table 2. Ablation study of different components in ARML.

Baseline	ARS	SWV	SRM	TUM	STR	LYM	NEC	mIoU	Dice	ACC
✓				76.76 ± 0.94	72.17 ± 0.49	58.18 ± 2.04	62.24 ± 1.60	67.33 ± 0.29	80.24 ± 0.20	83.80 ± 0.38
✓	✓			78.17 ± 0.57	72.83 ± 0.57	58.53 ± 0.47	62.35 ± 1.61	67.97 ± 0.28	80.67 ± 0.22	84.46 ± 0.33
✓	✓	✓		78.45 ± 0.61	73.10 ± 0.48	58.75 ± 1.76	66.34 ± 0.17	69.16 ± 0.27	81.54 ± 0.17	84.75 ± 0.39
✓	✓		✓	**79.45 ± 0.28**	71.89 ± 0.28	59.96 ± 0.66	63.73 ± 0.41	68.76 ± 0.55	81.25 ± 0.33	84.62 ± 0.10
✓	✓	✓	✓	79.20 ± 0.30	**73.83 ± 0.17**	**60.25 ± 0.14**	**68.96 ± 0.44**	**70.56 ± 0.19**	**82.48 ± 0.10**	**85.63 ± 0.17**

3.2 Implementation Details

All experiments are implemented in PyTorch 1.11 framework on Ubuntu 20.04 server with an NVIDIA RTX A5000 GPU. For pseudo masks generation phase, ResNet38 [15] is selected for classification backbone, with the learning rate of 1×10^{-2} under a polynomial decay policy in 20 epochs. For segmentation phase, we utilize three different backbones (MixTransformer [16], ResNet38 and ResNeSt101 [18]) to verify our method. During training, we use the SGD optimizer for each network with the learning rate of 5×10^{-3}, momentum of 0.9 and weight decay of 5×10^{-4} in 10 epochs. For SWV, weight hyperparameter ω is set to 2. To balance the distance and angle loss terms in SRM, we set α to 30 and β to 60. For $\mathcal{L}$, we set λ to 0.2 in practise. During the model inference phase, category-wise intersection over union (IoU), mean IoU (mIoU), Dice score, and pixel-level accuracy (ACC) are adopted as the metrics.

Table 3. Effectiveness of different values of ω in SWV. N/A denotes not utilizing SWV for training.

ω	mIoU	Dice
N/A	67.97 ± 0.28	80.67 ± 0.22
1	69.22 ± 0.22	81.54 ± 0.15
2	**70.56 ± 0.19**	**82.48 ± 0.10**
3	70.00 ± 0.24	82.13 ± 0.14
5	69.52 ± 0.29	81.78 ± 0.20
$+\infty$	69.69 ± 0.33	81.90 ± 0.19

Table 4. Performance of using different relation mining strategies. CE means the baseline with standard cross entropy loss.

	mIoU	Dice
CE	67.97 ± 0.28	80.67 ± 0.22
KL [8]	68.22 ± 0.84	80.92 ± 0.57
CRL [17]	68.68 ± 0.58	81.20 ± 0.41
$\mathcal{L}_{pt}$	69.83 ± 0.33	82.00 ± 0.17
$\mathcal{L}_{st}$	**70.56 ± 0.19**	**82.48 ± 0.10**

3.3 Comparison Study

We compare our ARML with other eight state-of-the-art weakly supervised segmentation methods under the same settings mentioned in Section 3.2, and the results are listed in Table 1. It is evident that our method exceeds other models on both two datasets, even with three different backbones. With the ResNeSt101 backbone, our method achieves the highest mIoU at 70.56% and 77.11% on BCSS and LUAD-HistoSeg respectively, which is 2.07% and 1.21% higher than the best existing method TPRO [19] ($p < 0.05$). The results are proved that our ARML is superior to WSSS for histological tissues. Qualitative comparisons on two datasets are available at Supplementary Material.

3.4 Ablation Study

We conduct several ablation studies on **BCSS** dataset as described in the following parts. More ablation studies can be found in Supplementary Material. Table 2 shows the specific results for each module of our ARML. When adding the ARS to conventional CAM baseline, the segmentation metrics have increased. Compared with the model without denoising strategies, the SWV and SRM has improved the mIoU by 1.19% and 0.79%, respectively. By combining these two strategies, the mIoU is further improved. In a word, each core component in our method plays a crucial role to enhance the segmentation capacity of model.

To analysis the effectiveness of weight coefficient ω in SWV that reweights the predicted pixels, we set different values of ω to evaluate the segmentation performance of our model. The results are shown in Table 3. When $\omega \to +\infty$, it means the pixels with inverse voting results between two networks are discarded in supervising the third network. We can observe that when the settings of ω is 2, the model achieves the best mIoU and Dice score than other values.

To verify the practicability of relation mining loss introduced in Section 2.4, we compare two common relation mining strategies KL [8] and CRL [17] with our proposed $\mathcal{L}_{pt}$ and $\mathcal{L}_{st}$. As shown in Table 4, our relation mining strategies are significantly better than KL and CRL, which is benefitting from the credible noise-robust designment. Among them, structure-wise relation mining loss $\mathcal{L}_{st}$ performs the best, which surpasses 2.59% and 1.81% in terms of mIoU and Dice compared to the cross entropy loss. This is due to the essential guidance role of binary distance and ternary angle inter-relation features.

4 Conclusion

In weakly supervised histological tissue segmentation, there are two existing challenges about low-quality pseudo masks generation and training with noisy labels. To solve these problems, we propose the Activation Relocalization Scheme to obtain more accurate pseudo masks. Then, we utilize Mutual Learning with two denoising strategies to mitigate the impact of noisy annotations. Experimental results show that our method addresses these matters effectively and achieves new state-of-the-art performance on two histological datasets, which make a contribution to the research community of computational pathology.

Acknowledgments. This work was supported in part by Guangxi Natural Science Foundation (No. 2024GXNSFFA010014 and 2019GXNSFFA245014) and National Natural Science Foundation of China (No. 82360356, 62172120, and 82272075).

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

References

1. Amgad, M., Elfandy, H., Hussein, H., Atteya, L.A., Elsebaie, M.A., Abo Elnasr, L.S., Sakr, R.A., Salem, H.S., Ismail, A.F., Saad, A.M., et al.: Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics **35**(18), 3461–3467 (2019)
2. Chan, L., Hosseini, M.S., Rowsell, C., Plataniotis, K.N., Damaskinos, S.: Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10662–10671 (2019)
3. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 839–847 (2018). https://doi.org/10.1109/WACV.2018.00097
4. Chen, Y.J., Hu, X., Shi, Y., Ho, T.Y.: Ame-cam: Attentive multiple-exit cam for weakly supervised segmentation on mri brain tumor. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 173–182. Springer Nature Switzerland, Cham (2023)
5. Chen, Z., Wang, T., Wu, X., Hua, X.S., Zhang, H., Sun, Q.: Class re-activation maps for weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 969–978 (2022)
6. Han, C., Lin, J., Mai, J., Wang, Y., Zhang, Q., Zhao, B., Chen, X., Pan, X., Shi, Z., Xu, Z., et al.: Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. Medical Image Analysis **80**, 102487 (2022)
7. He, J., Zhou, G., Zhou, S., Chen, Y.: Online hard patch mining using shape models and bandit algorithm for multi-organ segmentation. IEEE Journal of Biomedical and Health Informatics **26**(6), 2648–2659 (2022). https://doi.org/10.1109/JBHI.2021.3136597

8. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
9. Huber, P.J.: Robust estimation of a location parameter. In: Breakthroughs in statistics: Methodology and distribution, pp. 492–518. Springer (1992)
10. Li, Y., Duan, Y., Kuang, Z., Chen, Y., Zhang, W., Li, X.: Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1447–1455 (2022)
11. Li, Y., Yu, Y., Zou, Y., Xiang, T., Li, X.: Online easy example mining for weakly-supervised gland segmentation from histology images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 578–587. Springer (2022)
12. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988 (2017)
13. Qin, J., Wu, J., Xiao, X., Li, L., Wang, X.: Activation modulation and recalibration scheme for weakly supervised semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2117–2125 (2022)
14. Shen, B., Saito, A., Ueda, A., Fujita, K., Nagamatsu, Y., Hashimoto, M., Kobayashi, M., Mirza, A.H., Graf, H.P., Cosatto, E., et al.: Development of multiple ai pipelines that predict neoadjuvant chemotherapy response of breast cancer using h&e-stained tissues. The Journal of Pathology: Clinical Research **9**(3), 182–194 (2023)
15. Wu, Z., Shen, C., Van Den Hengel, A.: Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recognition **90**, 119–133 (2019)
16. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)
17. Xue, C., Deng, Q., Li, X., Dou, Q., Heng, P.A.: Cascaded robust learning at imperfect labels for chest x-ray segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23. pp. 579–588. Springer (2020)
18. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., et al.: Resnest: Split-attention networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2736–2746 (2022)
19. Zhang, S., Zhang, J., Xie, Y., Xia, Y.: Tpro: Text-prompting-based weakly supervised histopathology tissue segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 109–118. Springer (2023)
20. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2921–2929 (2016)