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

Appendix

Pseudo-Mask Refining Module

The structure of the pseudo-mask refining module is shown in Figure 1.

Ablation Studies

In this section, two more ablation studies are conducted on the WSSS4LUAD dataset to validate the effectiveness of the proposed synthesized dataset generation module.

Ablations of Different Cropping Sizes of Spliced Images In single tissue type image splicing, the cropping width and height of each selected image, i.e., H_p and W_p, are essential hyperparameters for image synthesis. In this experiment, we fix the shape of a synthesized image to 224×224 and adjust H_p from 224 to 16 to study the effects of different cropping sizes. For simplicity, W_p is set equal to H_p.

The ablation results are listed in Table 1. The table reveals that the segmentation accuracy achieves the best when $H_p = 32$ (therefore $n_w = n_h = 224/32 = 7$), which is the setting utilized in PistoSeg. On the one hand, larger cropping sizes lead to lower segmentation performance. For example, when $H_p = 112$, the normal IoU drops sharply from 0.7246 to 0.6201, decreasing over 0.1. Other types of IoU also decrease by around 0.02 to 0.03 when changing H_p from 32 to 112. On the other hand, when more images are spliced ($H = 16$), the mean IoU also decreases slightly. This phenomenon gives us an insight that although the bigger cropping size leads to fewer images to be spliced and rarer artifacts existing in the synthesized images, the variety of TME can hardly be simulated with such a small number of images with a single tissue type, due to the complexity of TME. Besides, when the cropping sizes are smaller and more images are spliced, additional artifacts are brought to the synthesized images, which differ from the real histopathology images and lead to degradation of segmentation performance. In conclusion, a moderate cropping size of 32×32 achieves the best performance.

H_p	TUM	STR	NOM	mIoU	fwIoU
224	<u>0.7990</u>	0.6952	0.6676	0.7206	0.7527
112	0.7916	0.6813	0.6201	0.6976	0.7414
32	**0.8119**	**0.7173**	<u>0.7246</u>	**0.7513**	**0.7707**
16	0.7948	<u>0.7000</u>	**0.7532**	<u>0.7493</u>	<u>0.7548</u>

Table 1: Effect on different cropping sizes of spliced images. We bold the highest and underline the second highest results.

Preliminary Segmentation with Different Image Synthesis Strategies In this paper, a synthesized dataset is generated based on the Mosaic transformation to train a preliminary segmentation model, which is utilized to infer the pseudo-masks of the training set, addressing the WSSS problem in a fully-supervised manner. Besides Mosaic transformation, several uncomplicated strategies can also generate synthesized images with pixel-level masks. Three other image synthesis strategies are adopted for comparison to prove that the proposed synthesized image generation module in PistoSeg can achieve the best performance. The generated synthesized images of the three strategies and the PistoSeg are illustrated in Figure 2. The introduction of the three comparison strategies is listed as follows.

- One Label. This strategy directly utilizes images with a single tissue type for preliminary segmentation. Specifically, to expand the dataset size, random resized crop, random flip, and random 90-degree rotation are utilized.
- Cutmix. This strategy composes a synthesized image using two images with a single tissue type by Cutmix (Yun et al. 2019). More specifically, Cutmix first randomly crops a region of a selected image and then pastes it to another to form a synthesized image.
- Gridding. The Gridding strategy generates a synthesized image by randomly cropping $n_h \times n_w$ images with a single tissue type to the shape of $H_p \times W_p$. Then, the cropped images are gridded following a raster order to generate a synthesized image. Here, n_h and n_w are set to 7 as is in PistoSeg for a fair comparison.

In the experiment, the number of generated synthesized images with different image synthesis strategies is all set

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

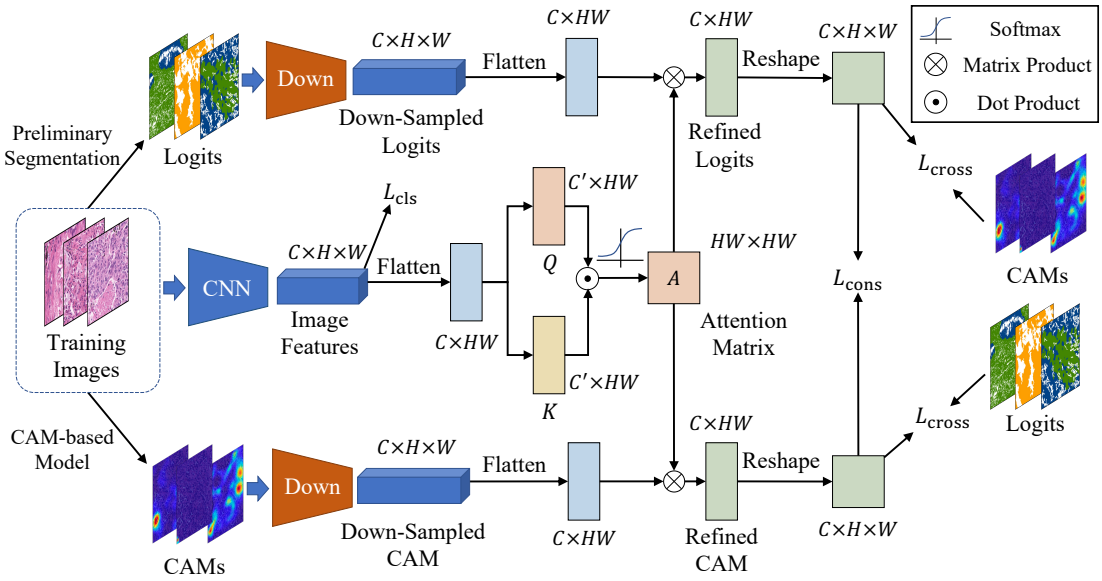

Figure 1: The pseudo-mask refining module. As the pseudo-mask and CAM are different transfers of the same image, their potential features should be closed to each other, bringing the regularization for the training of the module.

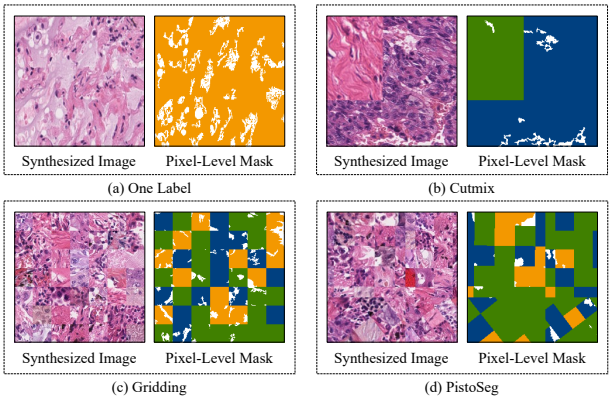

Figure 2: The generated synthesized images and pixel-level masks of the three comparison synthesis strategies along with the proposed PistoSeg. Blue, green, and yellow represent tumor, stroma, and normal tissue, respectively.

to 20,000, as well as PistoSeg. The segmentation performances of the three comparison strategies and PistoSeg are listed in Table 2. The table shows that PistoSeg outperforms all comparison strategies by a large margin except the normal IoU, which achieves the second best as well. The One Label strategy performs incredibly inferior in segmenting the stroma tissue. The performance of PistoSeg for stroma is around 0.25 higher than that of One Label. The performance of Cutmix is also not satisfactory, especially for normal IoU, which is over 0.15 lower than that of PistoSeg. The tumor and stroma IoU of Cutmix are also around 0.02 to 0.03 lower than PistoSeg. In the aspect of the Gridding strategy, although the normal IoU of Gridding is slightly higher

Strategy	TUM	STR	NOM	mIoU	fwIoU
One Label	0.7130	0.4762	0.7176	0.6356	0.6164
Cutmix	0.7937	0.6816	0.5723	0.6825	0.7413
Gridding	0.7936	0.6899	**0.7351**	0.7395	0.7495
PistoSeg	**0.8119**	**0.7173**	0.7246	**0.7513**	**0.7707**

Table 2: Effects of different image synthesis strategies.

than that of PistoSeg, the tumor and stroma IoU of Gridding are remarkably lower, making the mIoU and fwIoU inferior than PistoSeg as well.

To better understand the merits and demerits of different synthesis strategies, the pseudo-masks generated by preliminary segmentation with different synthesis strategies are shown in Figure 3. The figure shows that One Label tends to predict all pixels of an image in one category. For example, there should be both normal and tumor tissue in the first image. However, One Label predicts all pixels as normal tissue. This issue can be explained by the synthesized images generated by One Label contains only one category, which misleads the preliminary segmentation model to learn that two different tissue types should not exist simultaneously in one image. For Cutmix, since it generates synthesized images by pasting a rectangle region of one image to another, the straight border line between different tissue types in synthesized images makes the predicted borders between different tissues for real histopathology images also be straight, as pointed by the arrows. For Gridding, the synthesized images often have checkerboard artifacts, making the inferred pseudo-masks have checkerboard-shaped artifacts as well, as is illustrated by the arrows of the fourth row, resulting in training noises. Overall, the PistoSeg has the minimum artifacts compared with other synthesis strategies, validating the

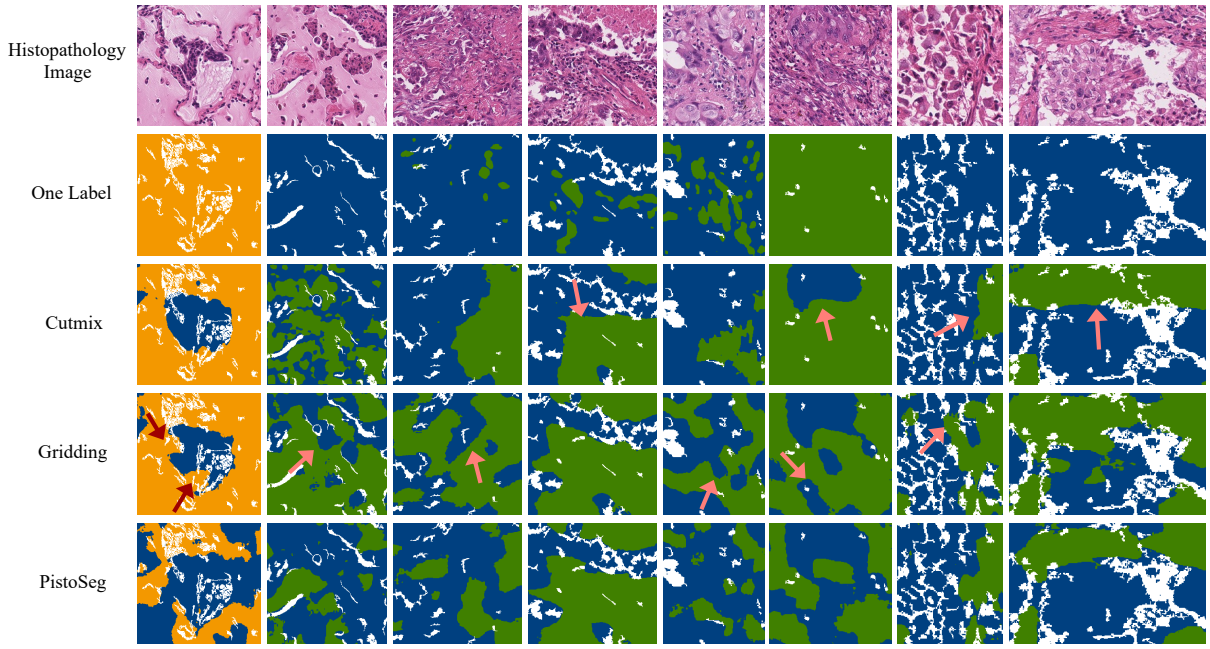

Figure 3: The predicted pseudo-masks of the training set using the three comparison synthesis strategies and the PistoSeg. Arrows point to the artifacts in the predicted pseudo-masks.

reasonability of the proposed synthesized dataset generation module.

References

Yun, S.; Han, D.; Chun, S.; Oh, S. J.; Yoo, Y.; and Choe, J. 2019. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6022–6031.