ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888097

PDSeg: Patch-Wise Distillation and Controllable Image Generation for Weakly-Supervised Histopathology Tissue Segmentation

Wei-Hua Li[⋆] *Yu-Hsing Hsieh*[⋆] *Huei-Fang Yang*[†] *Chu-Song Chen*[⋆]

[⋆] National Taiwan University, Taipei, Taiwan
[†] National Sun Yat-sen University, Kaohsiung, Taiwan

ABSTRACT

Weakly-supervised semantic segmentation, which achieves pixel-wise segmentation using image-level labels, has emerged as an alternative to fully supervised methods by reducing the need for detailed annotations. Inspired by the recent success of the teacher-student strategy in various vision tasks, we present a transformer-based weakly supervised framework that distills knowledge from a CNN teacher. Specifically, we incorporate a sequence of patch-wise distillation tokens into the transformer student, with each token focused on learning a specific patch under the teacher's guidance. This design enables the teacher to provide more reliable supervision to the student. On the other hand, in pathology images, it is often observed that certain tissue types are less represented than others. This class imbalance poses a significant challenge for many WSSS algorithms. To address this issue, we further introduce a data synthesis pipeline using a diffusion model conditioned on semantic label maps to mitigate the effects of class imbalance in histopathology images. Unlike previous methods that rely on full annotations to construct semantic label maps, our approach leverages the intrinsic characteristics of histopathology images. This leads to an approach that does not require full annotations and is well-suited for weakly-supervised scenarios. Through extensive experiments on the LUAD-HistoSeg and BCSS-WSSS datasets, we demonstrate that our approach outperforms state-of-the-art methods.

Index Terms— Weakly-supervised histopathology semantic segmentation, data augmentation

1. INTRODUCTION

Histopathology tissue segmentation is a crucial step for identifying the presence of cancer within tissues, which assists doctors in evaluating cancer and enhances the efficiency of diagnosis [1]. However, training a fully supervised segmentation requires obtaining pixel-level annotations of images. This annotation process not only demands expert knowledge but also is labor-intensive. Hence, weakly-supervised semantic segmentation (WSSS), which achieves pixel-wise segmentation using image-level labels, has received much attention and emerged as an alternative to fully supervised methods by reducing the need for detailed annotations.

The majority of existing studies on weakly-supervised semantic segmentation have relied on classification activation mapping (CAM) [2–6] to generate pixel-wise pseudo-labels for supervising a segmentation model. However, CAM is known to produce coarse segmentations due to being trained using only image-level class labels, and considerable research efforts have been dedicated to improving the quality of CAM. Han *et al.* [7] employed a two-stage network with Progressive Dropout Attention to maximize the value of the sparse annotation. This design enables the model to focus not only on the most discriminative regions but also to identify less dominant areas, thereby enhancing the quality of CAM. Li *et al.* [8] proposed a two-stage network with online easy example mining (OEEM). They address that CAM might yield classification errors, particularly at the edges of structures. Therefore, they employed weighted cross-entropy loss, allowing the model to be influenced more by credible labels rather than noisy labels. Zhang *et al.* [9] proposed Swin-MIL [9], which employs Swin Transformer [10] for extracting global features. Additionally, they incorporate deep supervision to leverage multi-scale features. With the rapid development of the vision-language model, TPRO [11] computes a similarity map for the histopathology image by a vision and label encoder. Besides, they also enrich the vision feature by integrating relevant pathological knowledge. Despite significant research efforts, achieving pixel-wise reliable pseudo supervision remains a challenging task.

In this paper, we take a different approach that leverages patch-level knowledge extracted from a model trained on image-level labels for more reliable guidance. Our approach is built upon the Segmenter [12], a transformer-based encoder-decoder structure for segmentation. To achieve our goal, we incorporate a sequence of distillation tokens into the model. Each token, representing a patch in the image, distills aggregated region knowledge from a teacher in a teacher-student manner. This design shares the same spirit as DEiT [13], which reveals that distilling knowledge from

This work was supported in part by the National Science and Technology Council, Taiwan under grants NSTC 113-2634-F-006-002, 112-2221-E-110-047-MY3, 113-2640-E-110-002, 113-2326-B-002-009-MY3, and by NTU under grant 114L893104, 113L900902.

a teacher enhances the performance of a transformer model. However, DEiT focuses on classification and uses a single distillation token to capture global knowledge, while our method employs multiple patch-wise distillation tokens to capture finer, localized knowledge from the teacher for segmentation. Besides, we introduce an aggregator to enable training the model with image-level labels, hence termed Patch-wise Distillation for Weakly-supervised Segmenter.

Moreover, in histopathology images, certain tissue types are less represented than others, presenting a significant challenge for segmentation models due to class imbalance. Hence, we further introduce an image synthesis approach to address this issue by generating additional histopathology images for underrepresented tissue types. Our method is built upon ControlNet [14], a controllable diffusion model conditioned on textual inputs and image semantic labels. However, obtaining image semantic labels for histopathology images can be challenging due to their complex tissue structures. To address this challenge, we leverage the inherent characteristics of histopathology images, where certain images contain only one tissue type. This enables us to apply a simple technique to generate semantic label maps, which in turn guide the diffusion model for image synthesis. Experimental results demonstrate that our approach produces realistic images.

Our contributions are summarized as follows. (1) We present a transformer-based weakly-supervised segmentation model featuring a sequence of patch-level distillation tokens. These tokens distill patch-wise knowledge from a teacher model, effectively enhancing performance. (2) We leverage prior knowledge within histopathology images to create semantic label maps for controlling a diffusion model for image synthesis. Our approach constructs the semantic label maps solely based on image-level labels, making it well-suited for weakly-supervised learning scenarios. (3) Extensive experiments on LUAD-HistoSeg and BCSS-WSSS demonstrate that our approach to histopathology image segmentation performs favorably against state-of-the-art approaches.

2. METHOD

This section introduces our Weakly-Supervised Segmenter, which incorporates learnable distillation tokens to receive patch-wise guidance from a teacher model. It also uses an aggregator to facilitate training with image-level labels. We then present data synthesis approach to address class imbalance.

2.1. Patch-wise Distillation Segmenter (PDSeg)

Our weakly-supervised Segmenter, as illustrated in Figure 1, is a transformer-based encoder-decoder structure trained with image-level labels. An image $\boldsymbol{I} \in \mathbb{R}^{H \times W \times 3}$ is divided into patches $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N] \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where (H, W) denotes the image size, N the number of patches, P the patch size, and C the number of channels. Each patch is flattened and projected, producing a sequence of patch embeddings

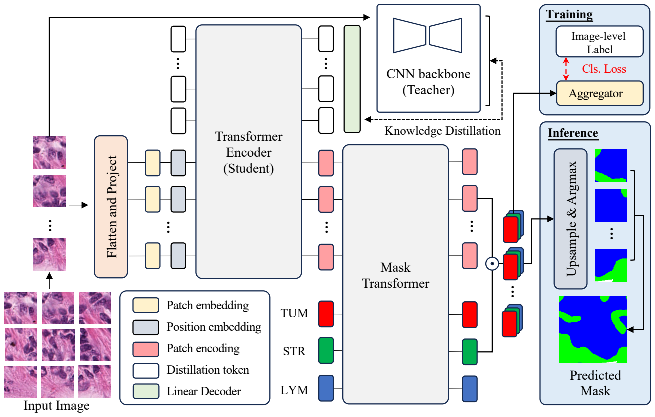

Fig. 1. Overview of patch-wise distillation for weakly-supervised segmenter (PDSeg). The model is equipped with a sequence of learnable distillation tokens, each dedicated to acquiring aggregated region knowledge from a CNN teacher. Additionally, an aggregator is introduced to facilitate training with image-level class labels.

$\mathbf{E}_{patch} \in \mathbb{R}^{N \times D}$, with D representing the embedding dimension. These patch embeddings, combined with positional embeddings, are processed through the encoder to yield a sequence of patch encodings $\mathbf{E}_{patch}^{enc} \in \mathbb{R}^{N \times D}$, which encapsulate rich semantic information. The encoder-generated patch encodings and the K learnable class embeddings, one for each class, are processed jointly in the decoder. Let $\mathbf{E}_{patch}^{dec}$ denotes the decoder outputted patch embeddings and $\mathbf{E}_{cls} = [\mathbf{e}_{cls}^1, ..., \mathbf{e}_{cls}^K] \in \mathbb{R}^{K \times D}$ denote the outputted class embeddings. A 2D feature map $\mathbf{M} \in \mathbb{R}^{H/P \times W/P \times K}$ containing semantic scores of each class is obtained by performing a scalar product between $\mathbf{E}_{patch}^{dec}$ and $\mathbf{E}_{cls}$, followed by reshaping.

Aggregator We apply an aggregator to all patches to produce class logits for training with BCE loss for classification. A common aggregator used is global average pooling (GAP) as in CAM [15]. However, GAP treats each patch equally and may result in coarse semantic maps. Instead, we employ normalised Global Weighted Pooling (nGWP) combined with focal penalty [16]. nGWP aggregates patches according to their contributions to the relevant class, helping generate finer maps. During inference, the 2D feature map $\mathbf{M}$ is upsampled to the original image size, followed by argmax to obtain the final segmentation mask.

Distillation Compared to the abundance of natural images, histopathology images are often less common. As demonstrated in [13], employing a distillation procedure enables training a transformer in a data-efficient manner. This finding motivates us to incorporate distillation via a teacher-student strategy in our weakly-supervised Segmenter. Unlike [13], which relies on a single distillation token for distilling global information from a teacher for classification, our approach leverages N learnable distillation tokens $\mathbf{E}_{dis} \in \mathbb{R}^{N \times D}$, with each token focused on learning patch-wise local information for segmentation. We augment the encoder input with N distillation tokens. These distillation tokens are jointly processed with the patch embeddings $\mathbf{E}_{patch}$ via attention, allowing the outputted sequence of distillation embeddings $\mathbf{E}_{dis}^{enc} = [\mathbf{e}_{dis}^{1}, ..., \mathbf{e}_{dis}^{N}] \in \mathbb{R}^{N \times D}$ to learn patch-wise semantic information. Each distillation embedding is projected into a semantic score vector of size K via a linear decoder. The semantic score vector can be interpreted as an aggregation of class logits for a patch in the original image, which indicates whether certain classes are present or not.

The target objective for the semantic scores of distillation tokens is provided by a CNN teacher. The CNN teacher consists of a ResNet101 encoder, followed by three layers of decoder, and is trained with global image class labels for classification. Let $\mathbf{M}_{cnn} \in \mathbb{R}^{H/P \times W/P \times K}$ be the output of the teacher. The sequence of class logits of distillation token is reshaped into $\mathbf{M}_{dis}$, which has the same size as the teacher's output. The distillation tokens learn from the teacher through a binary cross-entropy (BCE) loss:

$$L_d = \text{BCELoss}(\mathbf{M}_{dis}, \alpha * (\mathbb{H}(\mathbf{M}_{cnn})) + (1-\alpha) * \mathbb{S}(\mathbf{M}_{cnn})), \quad (1)$$

where α is the weight factor for label smoothing, and $\mathbb{H}$ and $\mathbb{S}$ denote the operations of hard label conversion and softmax.

It is worth noting that our distillation is performed at the patch level via transferring the knowledge of whether certain classes are present within a patch, rather than through pixel-wise distillation of the segmentation maps in the original size. This is because the teacher is trained with image-level classification, and its segmentation predictions could be noisy. The patch-wise distillation, on the other hand, provides more reliable supervision from the teacher.

2.2. Data Augmentation via Image Generation

With the rapid development of generative models, many studies have employed these models [17–19] to generate synthetic data, expanding training sets to enhance their diversity. On the other hand, augmenting the data can also mitigate the effects of class imbalance in histopathology images.

ControlNet for Histopathology Image Synthesis Our image synthesis builds upon ControlNet [14], which enhances text-to-image diffusion models with additional spatial conditions for a more precise expression of user intent. In specific, ControlNet generates images guided by both the text description and the control image. While semantic label maps can be used as control images for synthesizing plausible histopathology images, obtaining them using image-level labels is challenging due to the complex nature of tissue structures. To address this challenge, we exploit the intrinsic characteristics observed in histopathology images. Given that pathology whole slide images are often gigapixel-sized, a feasible approach for segmentation models to process is to generate smaller images from these whole slide images. Consequently, histopathology image datasets often contain patches of only one tissue type, which means they are associated with a single image-level label. We utilize these patches to construct the training set. That is because once backgrounds are identified in such patches, the remaining content can be treated as belonging to a specific tissue type. We obtain control images by applying a simple binary thresholding technique to the grayscale images. Despite its simplicity, our approach relies solely on image-level labels, making it suitable for weakly-supervised scenarios. Since we utilize patches associated with a single image label, we construct the text prompt by specifying the specific tissue type that we want to generate, following the format: "**[Tissue Type]** in the human **[Organ]**." For example, if our goal is to generate tissue containing Lymphocytic infiltrate, as observed in the BCSS-WSSS [7] dataset for breast cancer, the text prompt would be "Lymphocytic infiltrate in the human breast".

3. EXPERIMENTS

In this section, we begin by introducing the datasets and experimental setups. We then present experimental comparisons with other approaches, followed by ablation studies to provide insight into the effect of the introduced components.

3.1. Dataset

LUAD-HistoSeg [7] is a weakly-supervised tissue semantic segmentation dataset for lung adenocarcinoma, created with the aim of achieving pixel-level segmentations solely from patch-level labels. These labels include four tissue categories: tumor epithelial (TE), tumor-associated stroma (TAS), lymphocyte (LYM), and necrosis (NEC). The dataset comprises 17,285 patches cropped from whole slide images (WSI), divided into three subsets: 16,678 patches for training, 300 for validation, and 307 for testing. Each patch is of 224 × 224. The training patches are annotated with only patch-level labels while pixel-level annotations are provided for the validation and test patches. **BCSS-WSSS** comprises 31,826 patches generated by [7] from the H&E stained whole slide images of breast cancer. The training set contains 23,422 patches, the validation set has 3,418 patches, and the test set comprises 4,986 patches. There are four tissue category labels: Tumor (TUM), Stroma (STR), Lymphocytic infiltrate (LYM), and Necrosis (NEC).

3.2. Implementation Details

We train the model with Adam optimizer with an initial learning rate set to 1×10^{-2}. Our model is implemented in PyTorch and trained on a computer equipped with a single RTX 3090Ti GPU. We set the patch size P to 16, the embedding dimension D to 384, the number of layers in the encoder to 12, and the number of layers in the mask decoder to 2. To train ControlNet, we use stable diffusion v1.5 [21] as the backbone generation model, with a learning rate set to 1×10^{-5}. In the datasets used, the LYM and NEC categories are relatively

Table 1. Comparison between PDSeg and previous works on LUAD-HistoSeg and BCSS-WSSS.

Method	LUAD-HistoSeg					BCSS-WSSS				
	TE	NEC	LYM	TAS	mIoU	TUM	STR	LYM	NEC	mIoU
HistoSegNet [20] (ICCV'19)	45.59	36.30	58.28	50.82	47.75	33.14	46.46	29.05	1.91	27.64
TransWS [9] (MICCAI'22)	57.04	49.98	59.46	58.59	56.27	44.71	36.49	41.72	38.08	40.25
OEEM [8] (MICCAI'22)	73.81	70.49	71.89	69.48	71.42	74.86	64.68	48.91	61.03	62.37
WSSS-Tissue [7] (MIA'22)	73.90	77.48	73.61	69.53	73.63	74.54	64.45	52.54	58.67	62.55
TPRO [11] (MICCAI'23)	75.80	80.56	**78.14**	72.69	76.80	77.95	65.10	54.55	64.96	65.64
PDSeg (Ours)	78.53	81.35	74.65	**73.75**	77.07	79.33	73.08	60.45	65.71	69.64
PDSeg+Augment (Ours)	**78.95**	**81.79**	75.56	73.55	**77.46**	**80.09**	**74.13**	**61.56**	**67.31**	**70.77**

Table 2. Comparison between different numbers of distillation tokens.

Method	LUAD-HistoSeg					BCSS-WSSS				
	TE	NEC	LYM	TAS	mIoU	TUM	STR	LYM	NEC	mIoU
w/o distillation token	76.58	75.08	71.67	71.48	73.70	78.33	71.88	57.57	64.91	68.15
global token (class logits)	75.57	67.16	71.69	69.60	71.01	77.30	71.32	58.21	61.78	67.15
local patch-wise tokens (Ours)	**78.53**	**81.35**	**74.65**	**73.75**	**77.07**	**79.00**	**73.16**	**61.13**	**65.00**	**69.57**

less represented compared to the other two tissue types. To address the class imbalance, we add 1,000 synthesized images to each underrepresented type.

3.3. Comparison with previous works

Table 1 compares the performance of PDSeg with past state-of-the-art approaches [7–9, 11, 20] on LUAD-HistoSeg and BCSS-WSSS datasets. In the table, we compare IoU values for different categories as well as the mIoU (mean Intersection over Union) scores. As demonstrated, other approaches based on CAM yield inferior performance primarily because CAM tends to highlight only discriminative regions of an object, thus failing to generate pseudo-labels encompassing the entire object. In contrast, our approach that leverages distilling patch-wise knowledge from a CNN teacher already achieves the best overall performance on both datasets. Furthermore, its performance is enhanced when synthetic data is incorporated; specifically, our method outperforms the previous best approach by 5.13% on BCSS-WSSS.

3.4. Ablation Study

Table 2 compares the performance of using different numbers of distillation tokens: (1) no distillation token, (2) a single global distillation token learning from the CNN teacher's classification logits, as in [13], and (3) the proposed multiple local patch-wise tokens. The global token does not improve the results due to its limited capability to capture finer details. In contrast, our multiple distillation tokens allow the teacher to provide more detailed patch-wise guidance, boosting mIoU by 3.37% on LUAD-HistoSeg and 1.42% on BCSS-WSSS.

In Table 3, we investigate the impact of adding synthetic images to different tissue categories. In this experiment, we chose to augment each category of data with 1000 images. As shown in the table, when augmenting the classes with abundant training data in the original dataset, such as TUM and STR, the performance does not exhibit improvement. However, substantial performance gains are observed when augmenting the classes that were initially sparse, indicating the potential importance of balancing the dataset.

Table 3. Ablation study on the impact of adding synthetic images to different tissue categories.

Augmented Class				BCSS-WSSS				
TUM	STR	LYM	NEC	TUM	STR	LYM	NEC	mIoU
-	-	-	-	79.00	73.16	61.13	65.00	69.57
✓	-	-	-	79.37	73.61	59.69	65.73	69.60
-	✓	-	-	78.69	73.26	60.30	65.39	69.41
-	-	✓	-	78.98	73.06	**62.00**	65.78	69.95
-	-	-	✓	78.24	73.03	60.36	**68.61**	70.06
-	-	✓	✓	**80.09**	**74.13**	61.56	67.31	**70.77**
✓	✓	✓	✓	79.90	73.93	60.98	65.94	70.19

4. CONCLUSION

In this paper, we introduced a weakly-supervised transformer-based model that utilizes a sequence of learnable distillation tokens to learn from a CNN teacher. Each token receives aggregated patch-wise information from the teacher. To tackle the class imbalance in histopathology images, we proposed a data generation pipeline based on a controllable diffusion model. This pipeline leverages the inherent characteristics of histopathology images to synthesize images for enriching less-represented tissue types. Experimental results on two datasets demonstrate the effectiveness of our approach.

5. REFERENCES

[1] Bingchao Zhao, Xin Chen, Zhi Li, Zhiwen Yu, Su Yao, Lixu Yan, Yuqian Wang, Zaiyi Liu, Changhong Liang, and Chu Han, "Triple U-net: Hematoxylin-aware nuclei segmentation with progressive dense feature aggregation," *Med. Image Anal.*, 2020.

[2] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[3] Lian Xu, Mohammed Bennamoun, Farid Boussaid, Senjian An, and Ferdous Sohel, "An improved approach to weakly supervised semantic segmentation," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing.*, 2019.

[4] Hyeokjun Kweon and Kuk-Jin Yoon, "From sam to cams: Exploring segment anything model for weakly supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

[5] Sung-Hoon Yoon, Hoyong Kwon, Hyeonseong Kim, and Kuk-Jin Yoon, "Class tokens infusion for weakly supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

[6] Bingfeng Zhang, Siyue Yu, Yunchao Wei, Yao Zhao, and Jimin Xiao, "Frozen clip: A strong backbone for weakly supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

[7] Chu Han, Jiatai Lin, Jinhai Mai, Yi Wang, Qingling Zhang, Bingchao Zhao, Xin Chen, Xipeng Pan, Zhenwei Shi, Zeyan Xu, Su Yao, Lixu Yan, Huan Lin, Xiaomei Huang, Changhong Liang, Guoqiang Han, and Zaiyi Liu, "Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels," *Med. Image Anal.*, 2022.

[8] Yi Li, Yiduo Yu, Yiwen Zou, Tianqi Xiang, and Xiaomeng Li, "Online easy example mining for weakly-supervised gland segmentation from histology images," in *Int. Conf. Medical Image Computing and Computer-Assisted Intervention*, 2022.

[9] Shaoteng Zhang, Jianpeng Zhang, and Yong Xia, "Transws: Transformer-based weakly supervised histology image segmentation," in *Int. Conf. Medical Image Computing and Computer-Assisted Intervention*, 2022.

[10] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Int. Conf. Comput. Vis.*, 2021.

[11] Shaoteng Zhang, Jianpeng Zhang, Yutong Xie, and Yong Xia, "Tpro: Text-prompting-based weakly supervised histopathology tissue segmentation," in *Int. Conf. Medical Image Computing and Computer-Assisted Intervention*, 2023.

[12] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid, "Segmenter: Transformer for semantic segmentation," in *Int. Conf. Comput. Vis.*, 2021.

[13] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou, "Training data-efficient image transformers & distillation through attention," in *Int. Conf. Machine learning*, 2021.

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, "Adding conditional control to text-to-image diffusion models," in *Int. Conf. Comput. Vis.*, 2023.

[15] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[16] Nikita Araslanov and Stefan Roth, "Single-stage semantic segmentation from image labels," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[17] Hyun-Jic Oh and Won-Ki Jeong, "Diffmix: Diffusion model-based data synthesis for nuclei segmentation and classification in imbalanced pathology image datasets," in *Int. Conf. Medical Image Computing and Computer-Assisted Intervention*, 2023.

[18] Puria Azadi Moghadam, Sanne Van Dalen, Karina C Martin, Jochen Lennerz, Stephen Yip, Hossein Farahani, and Ali Bashashati, "A morphology focused diffusion probabilistic model for synthesis of histopathology images," in *IEEE Winter Conference on applications of computer vision*, 2023.

[19] Kyungho Kim, Junseo Lee, and Jihwa Lee, "Image generation is may all you need for vqa," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing.* IEEE, 2023.

[20] Lyndon Chan, Mahdi S Hosseini, Corwyn Rowsell, Konstantinos N Plataniotis, and Savvas Damaskinos, "Histosegnet: Semantic segmentation of histological tissue type in whole slide images," in *Int. Conf. Comput. Vis.*, 2019.

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.