SAM-Path: A Segment Anything Model for Semantic Segmentation in Digital Pathology

Jingwei Zhang[1]([✉]), Ke Ma[2], Saarthak Kapse[1], Joel Saltz[1], Maria Vakalopoulou[3], Prateek Prasanna[1], and Dimitris Samaras[1]

[1] Stony Brook University, New York, USA
{jingwezhang,samaras}@cs.stonybrook.edu,
{saarthak.kapse,prateek.prasanna}@stonybrook.edu,
Joel.Saltz@stonybrookmedicine.edu
[2] Snap Inc., New York, USA
kemma@cs.stonybrook.edu
[3] CentraleSupélec, University of Paris-Saclay, Gif-sur-Yvette, France
maria.vakalopoulou@centralesupelec.fr

Abstract. Semantic segmentations of pathological entities have crucial clinical value in computational pathology workflows. Foundation models, such as the Segment Anything Model (SAM), have been recently proposed for universal use in segmentation tasks. SAM shows remarkable promise in instance segmentation on natural images. However, the applicability of SAM to computational pathology tasks is limited due to the following factors: **(1)** lack of comprehensive pathology datasets used in SAM training and **(2)** the design of SAM is not inherently optimized for semantic segmentation tasks. In this work, we adapt SAM for semantic segmentation by first introducing trainable class prompts, followed by further enhancements through the incorporation of a pathology encoder, specifically a pathology foundation model. Our framework, `SAM-Path` enhances SAM's ability to conduct semantic segmentation in digital pathology without human input prompts. Through extensive experiments on two public pathology datasets, the BCSS and the CRAG datasets, we demonstrate that the fine-tuning with trainable class prompts outperforms vanilla SAM with manual prompts by 27.52% in Dice score and 71.63% in IOU. On these two datasets, the proposed additional pathology foundation model further achieves a relative improvement of 5.07% to 5.12% in Dice score and 4.50% to 8.48% in IOU.

Keywords: Segment anything · Semantic segmentation · Fine-tuning

1   Introduction

Digital pathology has revolutionized histopathological analysis by leveraging sophisticated computational techniques to augment disease diagnosis and prognosis [6,16]. A critical aspect of digital pathology is semantic segmentation, which entails dividing images into discrete regions corresponding to various tissue structures, cell types, or subcellular components [12,17]. Accurate and efficient semantic segmentation is essential for numerous applications, such as tumor detection, grading, and prognostication, in addition to the examination of tissue architecture and cellular interactions [4,13–15]. As a result, the development and optimization of robust segmentation algorithms hold significant importance for the ongoing advancement of digital pathology [8,10,20].

The AI research community is currently experiencing a significant revolution in the development of large foundation models. Among the latest advancements in computer vision is the Segment Anything Model (SAM), which serves as a universal segmentation model [9]. SAM is pretrained on a dataset containing over 1 billion masks across 11 million images. The model is designed to segment objects using various human input prompts, such as dots, bounding boxes, or text. SAM's evaluation highlights its remarkable zero-shot performance, frequently competing with or even surpassing previous fully supervised models across diverse tasks. Considering these capabilities, SAM has the potential to become a valuable tool for enhancing segmentation in digital pathology.

Although SAM has demonstrated considerable potential in computer vision, its direct applicability to digital pathology has two major limitations: 1) The basic design of SAM involves manually inputting prompts, or densely sampled points, to segment instances while it does not have any component for semantic classification. Consequently, it does not intrinsically facilitate semantic segmentation, a crucial component in digital pathology that enables the identification and differentiation of various tissue structures, cell types, and sub-cellular components. 2) The training set of SAM lacks diverse pathology images. This hinders SAM's capacity to effectively address digital pathology tasks without additional enhancements. Deng et al. confirm that the zero-shot SAM does not achieve satisfactory performance in digital pathology tasks, even with 20 prompts (clicks/boxes) per image [3].

In this work, we adpat vanilla SAM for semantic segmentation tasks in computational pathology. Our proposed adaptation involves the incorporation of trainable class prompts, which act as cues for the targeted class of interest. The performance is further enhanced by introducing a pathology foundation model as an additional feature encoder, thereby incorporating domain-specific knowledge. The proposed method enables SAM to perform semantic segmentation without the need for human input prompts. Our primary contributions are summarized as follows:

1. The introduction of a novel trainable prompt approach, enabling SAM to conduct multi-class semantic segmentation.
2. The introduction of a pathology foundation model as an additional pathology encoder to provide domain-specific information.
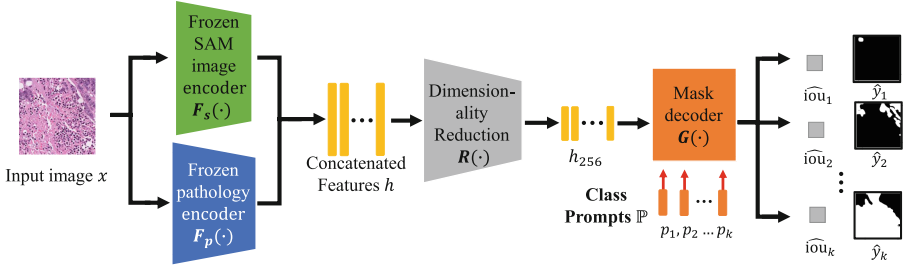
Fig. 1. Overview of `SAM-Path`. A pathology encoder $F_p(\cdot)$ is added in parallel with the vanilla SAM image encoder $F_s(\cdot)$ to provide more domain knowledge. The concatenated features from both the SAM image encoder and the pathology encoder are then passed to a dimensionality reduction module $R(\cdot)$. For mask prediction, we use class prompts $\mathbb{P}$ consisting of k learnable prompt tokens, in which each token prompts the mask decoder to predict the mask $\hat{y}_i$ of class i.

Through experimentation on two public pathology datasets, BCSS and CRAG, we demonstrate the superiority of our method over vanilla SAM. Here vanilla SAM refers to the classic SAM method with manual dot prompts or densely sampled dot prompts and some post-processing. On the CRAG dataset, the proposed trainable prompts achieve a relative improvement of 27.52% in Dice score and 71.63% in IOU compared to the vanilla SAM with manual prompts. We also demonstrate the benefit of the extra pathology foundation model, which leads to a further relative improvement of 5.07% to 5.12% in Dice score and 4.50% to 8.48% in IOU. Note that our goal is not to achieve SOTA performance on these datasets but to adapt SAM to semantic segmentation in digital pathology and boost its performance. To the best of our knowledge, we are the first to adapt SAM for semantic segmentation tasks in digital pathology without the need of manual prompts. By leveraging the power of SAM, pathology foundation models, and our innovative fine-tuning scheme, we aim to advance digital pathology segmentation and contribute to the ongoing development of AI-assisted diagnostic tools. Our code is available at https://github.com/cvlab-stonybrook/SAMPath.

2   Method

As shown in Fig. 1, our method consist of four modules: a SAM image encoder $F_s(\cdot)$ and a SAM mask decoder $G(\cdot)$ inspired from the vanilla SAM, a pathology encoder to extract domain-specific features $F_p(\cdot)$, and a dimensionality reduction module $R(\cdot)$. We discard the prompt encoder in the vanilla SAM because of the manually labeled prompts are not available in our segmentation tasks. Formally, given an input image x, our task is to predict its corresponding segmentation map y with the same resolution as x. Each pixel in y belongs to one of k predefined classes. We convert y into k segmentation masks $\{y_1, y_2, \ldots, y_k\}$, where y_i represents the segmentation mask of class i.

2.1   Pathology Encoder

The vanilla SAM uses a Vision Transformer (ViT) network pretrained on mostly natural images as the image encoder and thus its generated features lack pathology specific information. In our study, we use an extra pathology encoder to provide domain specific information. In this study, we use a pathology foundation model, the first stage ViT-Small of the HIPT model [2] which is pretrained on the TCGA Pan-cancer dataset [18]. As shown in Fig. 1, input image x is fed into both the vanilla SAM image encoder $F_s(\cdot)$ and the pathology encoder $F_p(\cdot)$. The output features are then concatenated as

$$h = [F_s(x), F_p(x)]. \tag{1}$$

The vanilla SAM contains the dimensionality reduction module within its image encoder, but as the dimensionality of output features h is now increased and not capable with decoder, we move this module $R(\cdot)$ after concatenation and adjust its input dimensionality accordingly.

2.2   Class Prompts

To enable the mask decoder $G(\cdot)$ to conduct semantic segmentation without manually inputting prompts, we use the trainable prompt token [7,19]. As shown in Fig. 1, for a segmentation task with k classes, we provide a set of class prompts. It consists of k trainable tokens $\mathbb{P} = \{p_i | i = 1, 2, \ldots, k\}$, where p_i is the class prompt of class i. Each of these class prompts p_i serve as the prompt to the mask decoder that it should segment class i. Different from the manually annotated dot prompts in the vanilla SAM, our class prompts are trainable and thus do not require human labelling.

For a class prompt p_i, the mask decoder, like that in the vanilla SAM, produces a predicted segmentation map $\hat{y}_i$ of class i and a IOU (Intersection over Union) prediction $i\hat{o}u_i$ that predicts the IOU of the predicted segmentation map and the ground truth y_i. The prediction is formulated as follows:

$$G(h_{256}, \mathbb{P}) = \{<i\hat{o}u_i, \hat{y}_i> | i = 1, 2, \ldots, k\} \tag{2}$$

Note that we conduct an extra softmax on all y_i for better performance.

2.3   Optimization

The vanilla SAM uses a combination of Dice loss, focal loss and the IOU loss (MSE loss on IOU predictions). We adapt their loss as follows:

$$\mathcal{L} = \sum_{i=1}^{k} [(1 - \alpha)\mathcal{L}_{dice}(\hat{y}_i, y_i) + \alpha\mathcal{L}_{focal}(\hat{y}_i, y_i) + \beta\mathcal{L}_{mse}(i\hat{o}u_i, IOU(\hat{y}_i, y_i))] \tag{3}$$

where $\alpha \in [0, 1]$ and β are weight hyper-parameters. $\mathcal{L}_{dice}$ represents the Dice loss function, $\mathcal{L}_{focal}$ represents the focal loss function and $\mathcal{L}_{mse}$ represents the Mean Squared Error (MSE) loss function. We update parameters in the mask decoder $G(\cdot)$, class prompts $\mathbb{P}$ and the dimensionality reduction module $R(\cdot)$ and keep the SAM image encoder $F_s(\cdot)$ and the pathology encoder $F_p(\cdot)$ frozen.

Table 1. Quantitative results of segmentation on the BCSS and CRAG datasests.

Dataset	BCSS		CRAG	
Metric	Dice	IOU	Dice	IOU
Vanilla SAM	/[a]	/	0.5245^b	0.3555^b
Vanilla SAM with post-processing	/[a]	/	0.6598	0.4924
Fine-tuned SAM (w.o. F_p)	0.7562	0.6080	0.8414	0.8451
SAM-Path w.o F_s	0.7813	0.6411	0.8191	0.8252
SAM-Path	**0.7949**	**0.6596**	**0.8841**	**0.8831**

[a] The vanilla SAM does not work on the BCSS dataset, as it cannot assign semantic labels to the multi-class segmented objects in this dataset.
[b] We assume all the objects that the vanilla SAM segmented are glands.

3   Experiments

3.1   Dataset

In our experiments, we use the BCSS [1] and CRAG [5] datasets for model evaluation. For both datsets, we use their official training and test splits and further split 20% of the training data into an explicit validation set.

BCSS: The Breast Cancer Semantic Segmentation (BCSS) dataset [1] has over 20,000 semantic segmentation annotations of tissue regions sampled from 151 H&E stained breast cancer images at $40\times$ magnification from TCGA-BRCA [11]. The annotations include 21 classes, we use the major 4 classes: Tumor, Stroma, Inflammatory and Necrosis. The rest are grouped into the 'others' class.

CRAG: The Colorectal adenocarcinoma gland (CRAG) dataset [5] has 213 images of the size $\approx 1536 \times 1536$ sampled from 38 H&E whole slide images (WSIs) at $20\times$ magnification. The annotations include the instance-level segmentation masks of the adenocarcinoma and benign glands in colon cancer. In our experiments, we convert the instance-level masks to semantics masks.

3.2   Results

For both datasets, we use the Dice score and Inter-section Over Union (IOU) as the main evaluation metrics. Implementation details and hyper-parameters are provided in the supplementary material. We also show the comparison of average prediction time in supplementary Table 1.

Evaluation of the Overall Performance. We mainly compare the proposed method with four baselines: 1) the vanilla SAM, i.e., SAM provided with manual dot prompts of each instance, 2) the vanilla SAM with post-processing, i.e., filtering out from the vanilla SAM output any instance occupying more than half of the image; this is because SAM occasionally erroneously segments the entire image as a single instance, 3) Fine-tuned SAM utilizing our class prompts, equivalent to SAM-Path without the pathology encoder F_p, and 4) SAM-Path without
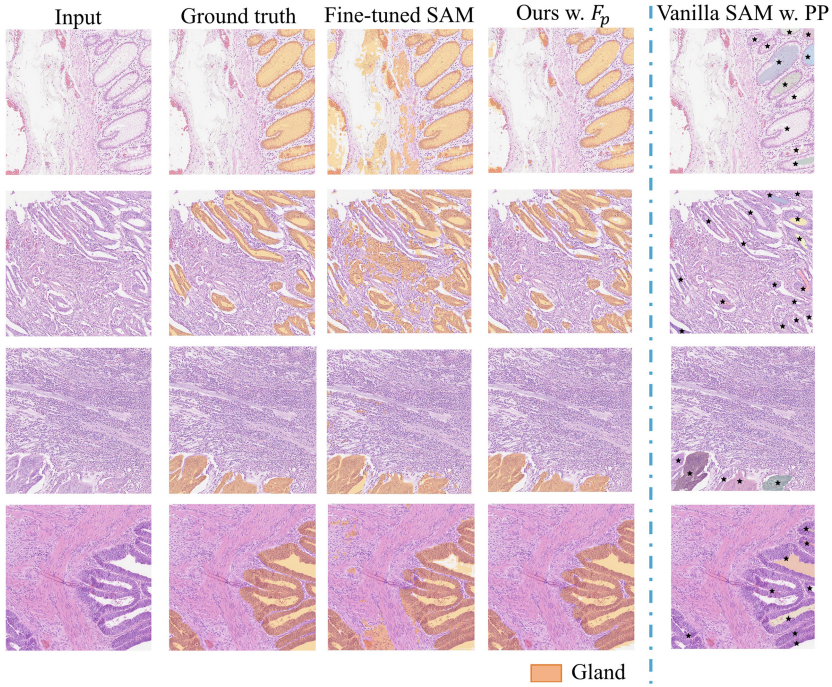
Fig. 2. Qualitative analysis on the CRAG dataset. PP represents post-processing that filters out instances occupying more than half of the image. For vanilla SAM, we provide a dot prompt (black asterisks) for each gland instance and assume all the segmented instances are glands. Our method performs better than the baselines.

the SAM image encoder F_s. Note that the original SAM lacks the capacity to predict semantics; we treat all segmented instances as glands within the context of the CRAG dataset.

As indicated in Table 1, the post-processing step enhances the performance of the original SAM, though the performance remains suboptimal. Compared with the vanilla SAM with post-processing, the fine-tuned SAM on the CRAG dataset achieves a relative improvement of 27.52% in Dice score and 71.63% in IOU, demonstrating the significant enhancement resulting from our fine-tuning scheme. The addition of the pathology encoder F_p (resulting in our proposed `SAM-Path`) leads to further improvements. Compared with the fine-tuned SAM without F_p, our method achieves a relative improvement of 5.12% in Dice score and 8.48% in IOU on the BCSS dataset, and 5.07% in Dice score and 4.50% in IOU on the CRAG dataset. These results underscore the value of incorporating domain-specific information from the pathology encoder to boost the performance of SAM in digital pathology tasks.
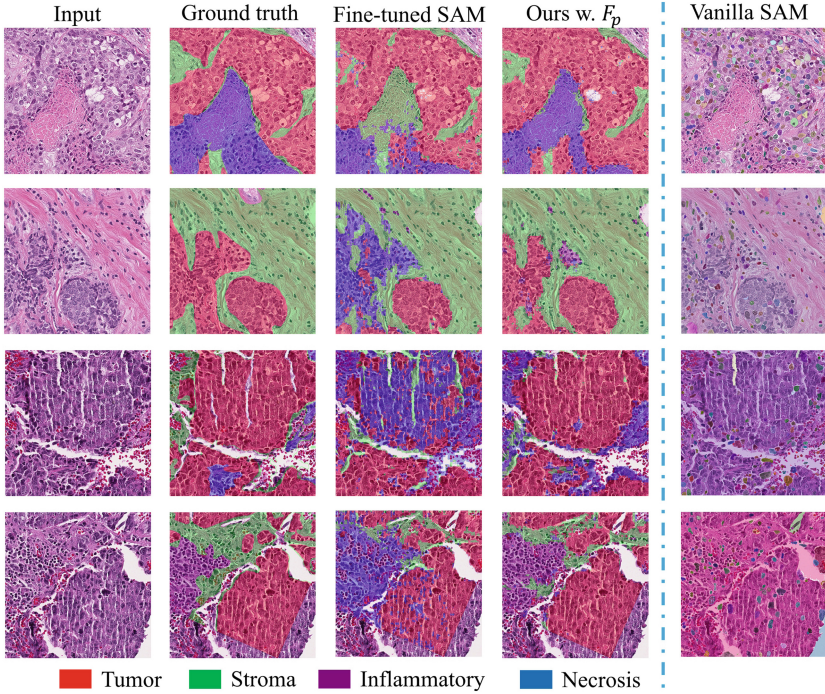
Fig. 3. Qualitative analysis on the BCSS dataset. The other class "others" and unlabeled regions are not colored. For the vanilla SAM, different colors represent different instances without any semantic meaning. Our method performs better than the baselines.

Also, when the SAM image encoder F_s is excluded, the BCSS dataset shows a relative decrease in performance by 1.71% in Dice score and 2.80% in IOU. For the CRAG dataset, the performance decline is more substantial, with a relative drop by 7.35% in Dice score and 6.56% in IOU. This suggests that the pathology segmentation can benefit from pre-taining of millions of natural images. Intriguingly, Table 1 reveals that `SAM-Path` without the pathology encoder (line 3) outperforms `SAM-Path` without the SAM encoder (line 4) on the CRAG dataset. However, the inverse is true for the BCSS dataset. This discrepancy is likely attributed to the fact that BCSS dataset segmentation involves multiclass semantic segmentation and hence benefits more from a domain-specific encoder, in contrast to the single semantic class of the CRAG dataset.

Qualitative Analysis. To qualitatively compare the performance of our method against others, we visualize the segmentation masks. In Fig. 2, we compare our method with vanilla SAM in which the dot prompts for each gland are provided (shown in black asterisks). Without fine-tuning, SAM lacks significant knowledge about the semantics in the pathology images. It frequently

segments the entire image as a single object (these instances are filtered out in the figure), or segments the white region within the gland as an object. However, our class prompts allow us to fine-tune SAM, thereby enabling the learning of semantic information from the training data. This leads to substantial improvement in performance. Also, the visualizations of vanilla SAM and vanilla SAM with post-processing are illustrated in Supplementary Fig. 1. Figure 3 further illustrates that in the BCSS dataset, our method with the pathology encoder outperforms its counterpart that lacks the pathology encoder. This is particularly evident in distinguishing between semantic classes like stroma and necrosis. For the vanilla SAM shown in Fig. 3, since the BCSS dataset is a semantic segmentation dataset without instance labels, we deploy the "segment everything" function of SAM. This function densely samples dots within the image to create segment instances.

Ablation Study. We conduct an ablation study to evaluate the influence of two loss weight values, α and β, on our model's performance, where α is the loss weight controlling the dice loss and focal loss and β is the loss weight controlling the IOU loss. Figure 4 presents the results, indicating the optimal values of α and β for the two datasets. Specifically, Fig. 4 (left) reveals that an α value of 0.25 yields the best performance for the BCSS dataset and an α value of 0.125 yields the best performance for the CRAG dataset. Similarly, Fig. 4 (right) shows that a β value of 0.0625 leads to optimal results for the BCSS dataset and the best β value for the CRAG dataset is 0.
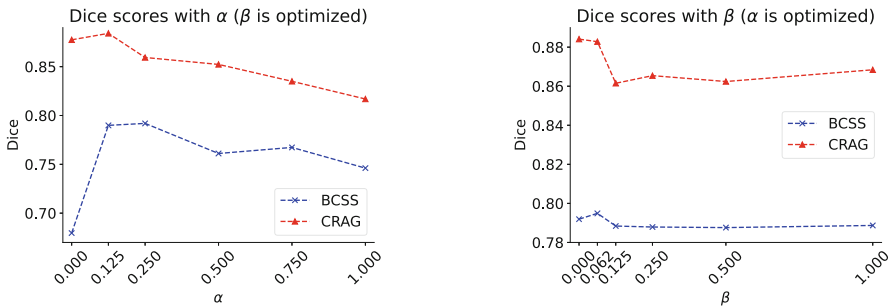

Fig. 4. Ablation study on the choice of two loss weights: α and β

4   Conclusion

In this paper, we introduced a novel fine-tuning approach using trainable class prompts to identify classes in segmentation tasks using SAM. Furthermore, we proposed the integration of a pathology encoder to incorporate more domain-specific knowledge. We evaluated our approach on two pathology segmentation

datasets, demonstrating that our method facilitates semantic segmentation without the need for manually inputted prompts and the pathology encoder consistently yielded improvements in Dice and IOU scores. Our approach indicates the promising potential of SAM for pathology semantic segmentation tasks. In future research, we plan to explore its potential in pathology panoptic segmentations.

References

1. Amgad, M., et al.: Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics **35**(18), 3461–3467 (2019)
2. Chen, R.J., et al.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, pp. 16144–16155 (2022)
3. Deng, R., et al.: Segment anything model (SAM) for digital pathology: assess zero-shot segmentation on whole slide imaging. arXiv preprint arXiv:2304.04155 (2023)
4. Ding, R., et al.: Image analysis reveals molecularly distinct patterns of TILs in NSCL associated with treatment outcome. npj Precis. Oncol. **6**(1), 33 (2022)
5. Graham, S., et al.: MILD-Net: minimal information loss dilated network for gland instance segmentation in colon histology images. Med. Image Anal. **52**, 199–211 (2019)
6. Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B.: Histopathological image analysis: a review. IEEE Rev. Biomed. Eng. **2**, 147–171 (2009)
7. Jia, M., et al.: Visual prompt tuning. In: Avidan, S., Brostow, G., Cisse, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision, ECCV 2022. LNCS, vol. 13693. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19827-4_41
8. Kapse, S., Torre-Healy, L., Moffitt, R.A., Gupta, R., Prasanna, P.: Subtype-specific spatial descriptors of tumor-immune microenvironment are prognostic of survival in lung adenocarcinoma. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2022)
9. Kirillov, A., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
10. Komura, D., Ishikawa, S.: Machine learning methods for histopathological image analysis. Comput. Struct. Biotechnol. J. **16**, 34–42 (2018)
11. Lingle, W., et al.: Radiology data from the cancer genome atlas breast invasive carcinoma (TCGA-BRCA) collection. Cancer Imaging Arch. **10**, K9 (2016)
12. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
13. Lu, C., et al.: Feature-driven local cell graph (FLocK): new computational pathology-based descriptors for prognosis of lung cancer and HPV status of oropharyngeal cancers. Med. Image Anal. **68**, 101903 (2021)
14. Madabhushi, A., Lee, G.: Image analysis and machine learning in digital pathology: challenges and opportunities. Med. Image Anal. **33**, 170–175 (2016)
15. Niazi, M.K.K., Parwani, A.V., Gurcan, M.N.: Digital pathology and artificial intelligence. Lancet Oncol. **20**(5), e253–e261 (2019)
16. Pantanowitz, L., et al.: Validating whole slide imaging for diagnostic purposes in pathology: guideline from the college of American pathologists pathology and laboratory quality center. Arch. Pathol. Lab. Med. **137**(12), 1710–1722 (2013)
17. Tizhoosh, H.R., Pantanowitz, L.: Artificial intelligence and digital pathology: challenges and opportunities. J. Pathol. Inf. **9**(1), 38 (2018)

18. Weinstein, J.N., et al.: The cancer genome atlas Pan-Cancer analysis project. Nat. Genet. **45**(10), 1113–1120 (2013)
19. Zhang, J., et al.: Prompt-MIL: boosting multi-instance learning schemes via task-specific prompt tuning. arXiv preprint arXiv:2303.12214 (2023)
20. Zhang, J., et al.: Precise location matching improves dense contrastive learning in digital pathology. In: Frangi, A., de Bruijne, M., Wassermann, D., Navab, N. (eds.) Information Processing in Medical Imaging, IPMI 2023. LNCS, vol. 13939, pp. 783–794. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-34048-2_60