All-in-SAM: from Weak Annotation to Pixel-wise Nuclei Segmentation with Prompt-based Finetuning

Can Cui
Computer Science
Vanderbilt University
Nashville, USA

Ruining Deng
Computer Science
Vanderbilt University
Nashville, USA

Quan Liu
Computer Science
Vanderbilt University
Nashville, USA

Tianyuan Yao
Computer Science
Vanderbilt University
Nashville, USA

Shunxing Bao
Electrical and Computer Engineering
Vanderbilt University
Nashville, USA

Lucas W. Remedios
Computer Science
Vanderbilt University
Nashville, USA

Bennett A. Landman
Electrical and Computer Engineering
Vanderbilt University
Nashville, USA

Yucheng Tang
NVIDIA Cooperation
Redmond, WA, USA

Yuankai Huo
Computer Science
Vanderbilt University
Nashville, USA
yuankai.huo@vanderbilt.edu

***Abstract*—The Segment Anything Model (SAM) is a recently proposed prompt-based segmentation model in a generic zero-shot segmentation approach. With the zero-shot segmentation capacity, SAM achieved impressive flexibility and precision on various segmentation tasks. However, the current pipeline requires manual prompts during the inference stage, which is still resource intensive for biomedical image segmentation. In this paper, instead of using prompts during the inference stage, we introduce a pipeline that utilizes the SAM, called all-in-SAM, through the entire AI development workflow (from annotation generation to model finetuning) without requiring manual prompts during the inference stage. Specifically, SAM is first employed to generate pixel-level annotations from weak prompts (e.g., points, bounding box). Then, the pixel-level annotations are used to finetune the SAM segmentation model rather than training from scratch. Our experimental results reveal two key findings: 1) the proposed pipeline surpasses the state-of-the-art (SOTA) methods in a nuclei segmentation task on the public Monuseg dataset, and 2) the utilization of weak and few annotations for SAM finetuning achieves competitive performance compared to using strong pixel-wise annotated data.**

***Index Terms*—foundation model, SAM, Segment Anything, annotation, prompt**

I. Introduction

The foundation models have recently been proposed as a powerful segmentation model [1], [2]. Segment Anything Model (SAM), as an example, was trained by millions of images to achieve a generic segmentation capability [3]. SAM can automatically segment a new image, and it also accepts the prompts input of foreground/background points or the box regions for better segmentation [4]–[7]. However, recent studies have revealed SAM's limited performance in specific domain tasks, such as medical image segmentation, particularly when an insufficient number of prompts are available [4]. The main reason is that medical data was rare to see in the training set of SAM while the medical segmentation tasks always in requirement of higher professional knowledge than natural image segmentation [8]. Using the finetuning strategy to adapt the generic segmentation model to downstream tasks provides a promising solution to utilize the power of the generic model in detecting low-level and general image patterns and features but adjust the final segmentation based on the characteristics and high-level understanding of downstream tasks. Previous approaches [6], [9] have proposed finetuning methods to improve SAM's performance in downstream tasks. However, these methods mostly require complete data annotation for finetuning and did not explore the impact of weak annotation and few training data on the finetuning of the pretrained SAM model.

Nuclei segmentation is a crucial task in biomedical research and clinical applications, but manual annotation of nuclei in whole slide images (WSIs) is time-consuming and labor-intensive. Previous works attempted to automatically segment nuclei with supervised learning [10], [11]. More recently, some methods used self-supervised learning to further improve the model performance [12], [13]. SAM has great potential to benefit nuclei segmentation if it can be adapted appropriately. This paper investigates the performance of transferring the SAM to nuclei segmentation. Previous studies have indicated that SAM performed poorly in nuclei segmentation without box/point information as prompts, but achieved promising segmentation when the bounding box of every nuclei was provided as the prompt in the inference stage. However, manually annotating all the boxes during inference remains time-consuming. To address this issue, we propose a pipeline for label-efficient finetuning of SAM, with no requirement for annotation prompts during inference. Also, instead of relying on complete annotations for finetuning, we leverage weak annotations to further reduce annotation costs while achieving comparable segmentation performance to state-of-the-art (SOTA) methods.

In this work, we proposed the All-in-SAM pipeline, utilizing the pretrained SAM for annotation generation and model finetuning. Instead of using prompts during the inference stage,

This research was supported by NIH R01DK135597 (Huo), NSF CAREER 1452485, NSF 2040462,NCRR Grant UL1-01 (now at NCATS Grant 2 UL1 TR000445-06), NVIDIA hardware grant, resources of ACCRE at Vanderbilt University

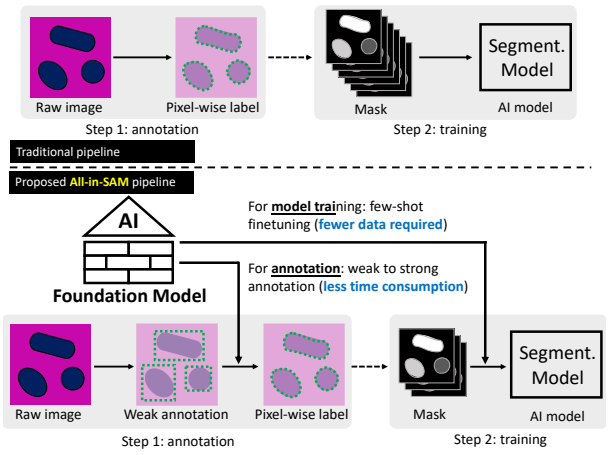

Fig. 1. This figure shows the overall idea of the proposed All-in-SAM pipeline. First, the AI foundation model SAM is used in the annotation phase to convert weak annotations (bounding boxes) to strong annotations (pixel-wise labels), which reduces the time consumption during the labeling process. Then, the SAM model is fine-tuned with fewer strong annotations. The ultimate goal of the All-in-SAM pipeline is to enable efficient few-shot and weak annotation for AI model adaptations.

no manual prompts are required during the inference stage (Fig. 1). The contribution of this work can be summarized into two points:

1) Utilization of weak annotations for cost reduction: Rather than relying exclusively on fully annotated data for finetuning, we demonstrate the effectiveness of leveraging weak annotations and the pretrained SAM. This approach helps to minimize annotation costs while achieving segmentation performance that is comparable to the current state-of-the-art methods.

2) Development of a pipeline for label-efficient finetuning: We propose a method that allows SAM to be finetuned for nuclei segmentation without the requirement of annotation prompts during inference. This significantly reduces the time and effort involved in manual annotation.

Overall, this work aims to enhance the application of SAM in nuclei segmentation by addressing the annotation burden and cost issues through label-efficient finetuning and the utilization of weak annotations.

II. Method

A. Overview

Motivated by the promising performance of the SAM model in interactive segmentation tasks with sparse prompts and the potential for finetuning, we propose a segmentation pipeline that leverages weak and limited annotations and apply this pipeline to the nuclei segmentation task.

The proposed pipeline consists of two main stages: SAM-empowered annotation and SAM finetuning. In the first stage, we utilize the pretrained SAM model to generate high-quality approximate nuclei masks for pathology images. This is achieved by providing the bounding boxes of nuclei as input to the pretrained SAM model. These approximate masks serve as initial segmentation masks for the nuclei. In the second stage, the generated approximate masks are employed to finetune the SAM model, which allows the model to adapt and refine its segmentation capabilities specifically for nuclei segmentation. The proposed pipeline is displayed in Fig. 2. Two stages are introduced in detail in II-B and II-C.

Furthermore, we evaluate the performance of the model when only a small number of annotated data for downstream tasks. By minimizing the number of annotated samples, we aim to reduce annotation labor while still achieving satisfactory segmentation results.

B. SAM-empowered annotation

The SAM model consists of three key components: the prompt encoder, the image encoder, and the mask decoder. The image encoder utilizes the Vision Transformer (ViT) as its backbone, employing a 14×14 windowed attention mechanism and four equally spaced global attention blocks to learn image representations effectively. The prompt encoder can take two forms: sparse or dense. In the sparse form, prompts can be in the form of points, boxes, or text, whereas in the dense form, prompts are represented as a grid or mask. The encoded prompts are then added to the image representation for the subsequent mask decoding process. In a previous study [4], it was observed that when only automatically generated dense prompts were used, nuclei segmentation sometimes failed to produce satisfactory results. However, significant improvement was achieved when weak annotations such as points or boxes were provided during the segmentation inference. Notably, when the bounding box of nucleus was available as a weak annotation, the segmentation achieved a dice value of 0.883 in the public Monuseg dataset [14], significantly surpassing the results obtained from supervised learning methods. It indicates that SAM has strong capabilities in edge detection, enabling clear detection of nuclei boundaries within focus regions. This makes it a potential tool to generate precise approximate masks, which can enhance supervised learning approaches with lower annotation costs.

C. SAM-finetuning

SAM has been trained on a large dataset for generic segmentation tasks, giving it the ability to perform well in general segmentation. However, when applied to specific tasks, SAM may exhibit suboptimal performance or even fail. Nonetheless, if the knowledge accumulated by SAM can be transferred to these specific tasks, it holds great potential for achieving better performance compared to training the model from scratch using only downstream task data, especially when the available data for the downstream task is limited.

To optimize the transfer of knowledge, rather than finetuning the entire large pretrained model, a more effective and efficient approach is to selectively unfreeze only the last few layers. However, in our experiments, this approach still yielded inferior results compared to some baselines. Recently, there has been growing attention in the natural language processing community toward the use of adapters as an effective tool for

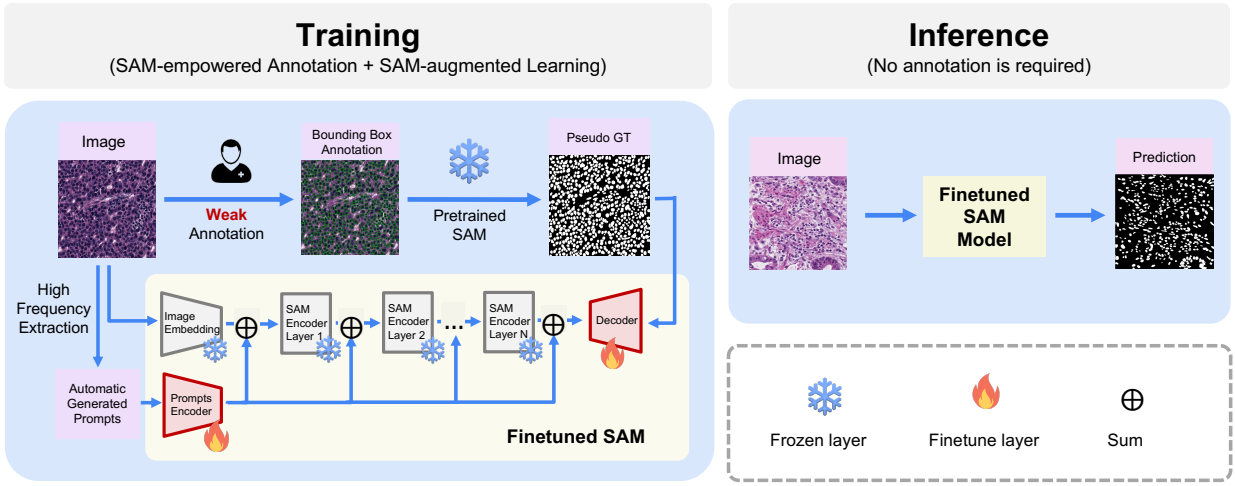

Fig. 2. The proposed pipeline for using weak and few annotations for nuclei segmentation. In the training stage, only the bounding boxes (in green color) of nuclei were provided as the weak annotation label to generate the approximate segmentation masks. Then, with the supervision of the approximate segmentation masks, the prompt-based finetuning was applied on the pretrained SAM model. In the inference stage, nuclei can be segmented directly from images without box prompts.

finetuning models for different downstream tasks by leveraging task-specific knowledge [15]. In line with this, Chen et al. [9] successfully adapted prompt adapters [16] in the finetuning process of SAM. Specifically, they automatically extracted and encoded the texture information of each image as handcrafted features, which were then added to multiple layers in the encoder. Additionally, the proposed prompt encoder, along with the unfrozen lightweight decoder, became learnable during the finetuning process. Following their work, we implement such finetuning strategy in the nucleus segmentation task, but we explore more about its performance in different numbers of training data scenarios.

III. DATA AND EXPERIMENTS

A. Data and Task

For evaluating the model performance, we employ the MICCAI 2018 Monuseg dataset [17]. It consists of 30 training images and 14 testing images, all with dimensions of 1000×1000 pixels. Each image is accompanied by corresponding masks of nuclei. To ensure a fair and comparable evaluation, we use the same data split as a recent study [18]. The 30 training images are divided into two subsets, with 24 images assigned to the training set and the remaining 6 images forming the validation set. To evaluate the model performance of nucleus segmentation, Dice, AUC, Recall, Precision, best F1 (maximized F1 score at the optimal threshold), IoU (Intersection over Union) and ADJ (Adjusted Rand Index) are calculated.

B. Experiment Setting

In this work, we designed 3 sets of experiments to explore the performance of finetuned SAM on the nucleus segmentation task.

1) Finetuned by complete annotation or weak annotation. For complete annotation, the pixel-wise complete masks were provided for training data to finetune the pretrained SAM model. As for the weak annotation, only the bounding boxes of nuclei were provided. In this work, the bounding boxes were automatically prepared by using the complete masks. And then, these bounding boxes were used as the prompts in the pretrained SAM to generate pixel-level pseudo labels for finetuning.

2) Finetuned by different numbers of annotated data. To evaluate the performance of the proposed pipeline finetuned with different numbers of data, we adjusted the number of annotated images and the area of annotated regions. The complete training set contains 24 1000×1000 image patches with corresponding annotations. In the 4% training data set, only a 200×200 random patch was selected from each large patch for annotation. To keep the parameters of the model unchanged for a fair comparison, the rest area without annotation was set to intensity 0. In the extreme cases, only 3 patches (1 from each image) in the size of 200×200, taking up 0.5% of the original complete dataset, were randomly selected.

3) Comparison with other SOTA. In this study, we conducted a performance comparison between our proposed pipeline and other state-of-the-art (SOTA) methods. LViT [18] is a recently proposed model integrating language information and images for annotation and achieved SOTA performance in Monuseg dataset. We followed their data splits in our experiments. BEDs [19] integrated the self-ensemble and testing stage stain augmentation mechanism in UNet models for nuclei segmentation. Although more data were used for training, the model was evaluated in the same testing set. So, the comparable performance results of LViT [18] and BEDs [19] are from their paper. For the task of learning from a small annotated dataset, Xie et al. [12] proposed to use self-supervised learning to utilize the unlabeled data and achieved better results when only a small number of

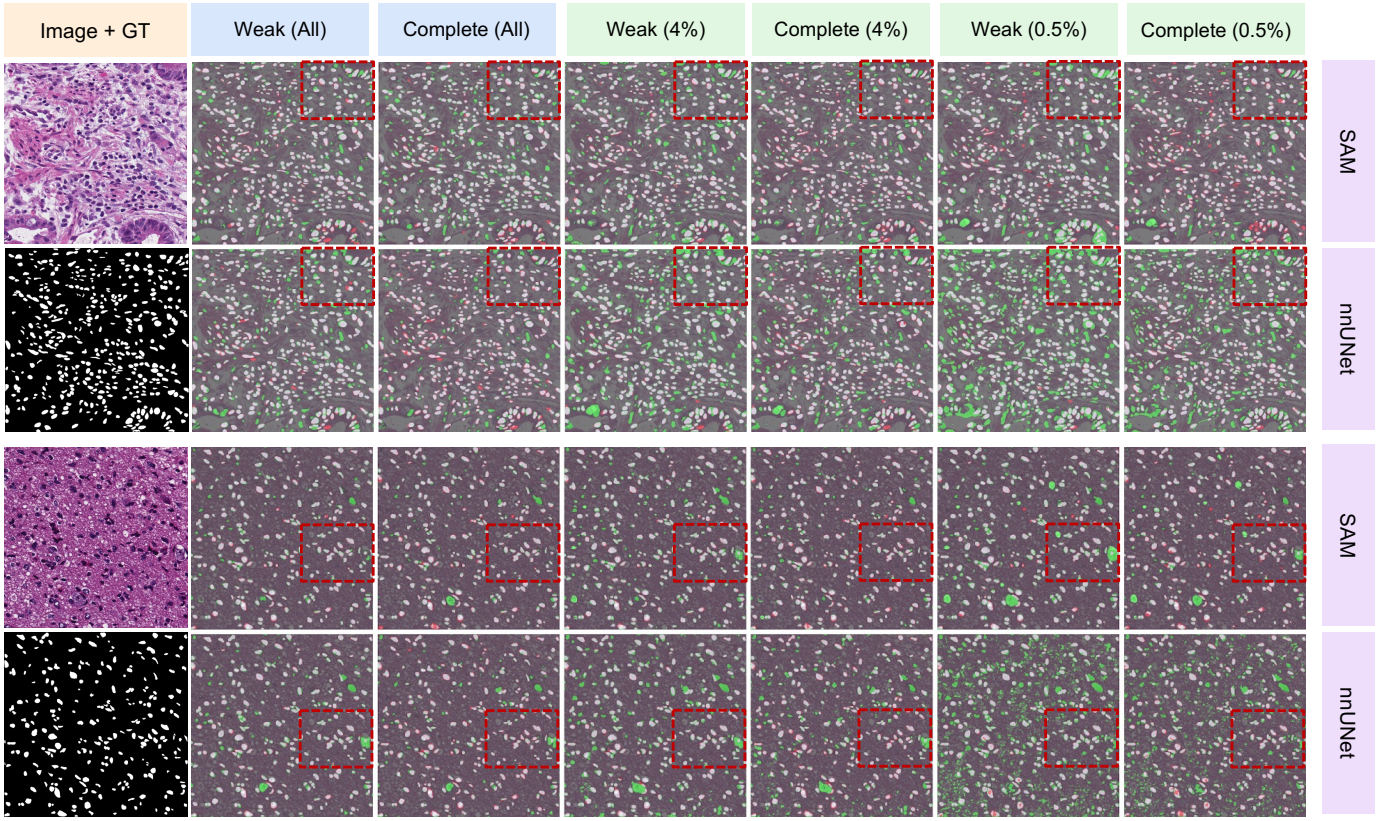

Fig. 3. Qualtitative results of nuclei segmentation by different methods with different numbers of annotated data. Upper row: proposed method; lower row: nnUNet. The mask region in green color indicates the false positive of prediction, the white color indicates the true positive, and the red color indicates the false negative.

TABLE I

COMPARISON WITH OTHER SOTA METHODS WHEN USING DIFFERENT NUMBERS OF TRAINING DATA AND COMPLETE ANNOTATION OR WEAK ANNOTATION. THE BEST PERFORMANCE IS HIGHLIGHTED IN RED COLOR, WHILE THE SECOND-BEST PERFORMANCE IS HIGHLIGHTED IN BLUE COLOR.

Label	Method	Training Data	Dice	IoU	ADJ
Complete	Xie et al [12]	All	-	-	**70.63%**
		30%			60.31%
		10%	-	-	55.01%
	LViT [18]	All	80.33%	67.24%	-
		25%	79.94%	66.80%	-
	BEDs [19]	All+More	81.77%	-	-
	nnUNet [20]	All	82.44%	**69.76%**	70.28%
		4%	79.20%	65.40%	65.80%
	Proposed	All	**82.54%**	**69.74%**	**70.36%**
		4%	81.34%	68.10%	68.67%
Weak	nnUNet [20]	All	82.12%	69.35%	69.74%
		4%	79.13%	65.27%	65.62%
	Proposed	All	**82.46%**	69.73%	70.24%
		4%	80.99%	67.60%	68.14%

TABLE II
COMPARISON OF DIFFERENT METHODS TRAINED BY DIFFERENT NUMBERS OF WEAKLY/COMPLETELY ANNOTATED DATA

Label	Method	Training Data	Dice	AUC	Recall	Precision	bestF1	IoU	ADJ
Complete	nnUNet	All	82.44%	96.04%	82.82%	82.55%	83.21%	69.76%	70.28%
		4%	79.20%	94.10%	84.90%	74.60%	79.70%	65.40%	65.80%
		0.50%	76.23%	92.49%	86.79%	68.61%	77.97%	61.35%	61.86%
	Proposed	All	82.54%	97.17%	84.74%	80.81%	83.04%	69.74%	70.36%
		4%	81.34%	95.50%	84.92%	78.53%	81.90%	68.10%	68.67%
		0.50%	78.16%	94.40%	84.57%	73.51%	79.17%	63.79%	64.30%
Weak	nnUNet	All	82.12%	95.65%	91.26%	74.98%	83.68%	69.35%	69.74%
		4%	79.13%	94.74%	92.76%	69.22%	81.54%	65.27%	65.62%
		0.50%	75.00%	92.64%	93.36%	61.99%	77.57%	58.81%	59.18%
	Proposed	All	82.46%	97.32%	89.47%	76.78%	83.39%	69.73%	70.24%
		4%	80.99%	95.22%	88.45%	75.09%	81.87%	67.60%	68.14%
		0.50%	78.73%	94.30%	87.04%	72.48%	79.82%	64.57%	65.02%

annotated training data were available. Their proposed method was implemented on the same Monuseg training and testing dataset, so their results were reported here for comparison with ours. In addition, nnUNet, as a popular benchmark in the medical image segment, was run by ourselves on the Monuseg with the same dataset setting as our proposed pipeline. To ensure a fair comparison, we used the default settings of nnUNet and trained it for the default 1000 epochs.

About other settings in our proposed pipeline, the ViT-H backbone was used in both the annotation generation and fine-tuning stages. The training would early stop if the validation loss did not decrease for consecutive 40 epochs. Without specific mention, other default parameters and settings in SAM-adapter [9] were kept. All experiments were repeated three times for average evaluation values. An RTX A6000 was used to run these experiments.

IV. RESULTS AND DISCUSSION

Table 1 shows the comparison of our results with other SOTA methods. The best performance was observed when training with the whole training set and complete annotation. Notably, the nnUNet [20], our proposed method, and Xie's method [12] demonstrated similar performance in this scenario. However, when training with a reduced number of annotated data, our proposed method exhibited a smaller drop in performance compared to other methods and achieved superior results. Additionally, when employing weak labels for training, the proposed method consistently outperformed other methods, maintaining the highest performance. Table 2 and Fig. 3 provide a comprehensive view of the evaluation metrics and show the performance under extreme cases where only 0.5% of the training set data is available. In various evaluation metrics such as Dice, AUC, Precision, bestF1, IoU, and ADJ, the proposed method consistently outperformed nnUNet across different settings. This was particularly evident when utilizing limited and poorly annotated data. However, when training with fewer and weakly annotated data, the proposed method exhibited a lower Recall value compared to nnUNet. Despite this, nnUNet displayed a more aggressive approach to segmenting nuclei, resulting in significantly lower precision and other metrics.

V. CONCLUSION

In summary, we introduce an efficient and effective pipeline that leverages a pretrained self-attention mechanism (SAM) for nuclei segmentation with limited annotation. The experiments show the capability of the pretrained SAM model to generate pseudo labels from weak annotations and subsequently fine-tune with these pseudo labels during the inference phase. This approach achieves competitive performance when compared to state-of-the-art (SOTA) methods while significantly reducing the burden of manual annotation. This pipeline holds great significance in real-world applications of nuclei segmentation, as it offers a practical solution that minimizes annotation efforts without compromising on segmentation accuracy.

REFERENCES

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] OpenAI, "Gpt-4 technical report," 2023.

[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[4] R. Deng, C. Cui, Q. Liu, T. Yao, L. W. Remedios, S. Bao, B. A. Landman, L. E. Wheless, L. A. Coburn, K. T. Wilson *et al.*, "Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging," *arXiv preprint arXiv:2304.04155*, 2023.

[5] J. Ma and B. Wang, "Segment anything in medical images," *arXiv preprint arXiv:2304.12306*, 2023.

[6] J. Wu, R. Fu, H. Fang, Y. Liu, Z. Wang, Y. Xu, Y. Jin, and T. Arbel, "Medical sam adapter: Adapting segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.12620*, 2023.

[7] Y. Zhang, T. Zhou, P. Liang, and D. Z. Chen, "Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model," *arXiv preprint arXiv:2304.11332*, 2023.

[8] Y. Huo, R. Deng, Q. Liu, A. B. Fogo, and H. Yang, "Ai applications in renal pathology," *Kidney international*, vol. 99, no. 6, pp. 1309–1320, 2021.

[9] T. Chen, L. Zhu, C. Ding, R. Cao, S. Zhang, Y. Wang, Z. Li, L. Sun, P. Mao, and Y. Zang, "Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more," *arXiv preprint arXiv:2304.09148*, 2023.

[10] Y. Kong, G. Z. Genchev, X. Wang, H. Zhao, and H. Lu, "Nuclear segmentation in histopathological images using two-stage stacked u-nets with attention mechanism," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 573866, 2020.

[11] H. Hu, Y. Zheng, Q. Zhou, J. Xiao, S. Chen, and Q. Guan, "Mc-unet: Multi-scale convolution unet for bladder cancer cell segmentation in phase-contrast microscopy images," in *2019 IEEE International*

Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2019, pp. 1197–1199.

[12] X. Xie, J. Chen, Y. Li, L. Shen, K. Ma, and Y. Zheng, "Instance-aware self-supervised learning for nuclei segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*. Springer, 2020, pp. 341–350.

[13] M. Sahasrabudhe, S. Christodoulidis, R. Salgado, S. Michiels, S. Loi, F. André, N. Paragios, and M. Vakalopoulou, "Self-supervised nuclei segmentation in histopathological images using attention," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*. Springer, 2020, pp. 393–402.

[14] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE transactions on medical imaging*, vol. 36, no. 7, pp. 1550–1560, 2017.

[15] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.

[16] W. Liu, X. Shen, C.-M. Pun, and X. Cun, "Explicit visual prompting for low-level structure segmentations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 434–19 445.

[17] N. Kumar, R. Verma, D. Anand, Y. Zhou, O. F. Onder, E. Tsougenis, H. Chen, P.-A. Heng, J. Li, Z. Hu *et al.*, "A multi-organ nucleus segmentation challenge," *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1380–1391, 2019.

[18] Z. Li, Y. Li, Q. Li, Y. Zhang, P. Wang, D. Guo, L. Lu, D. Jin, and Q. Hong, "Lvit: language meets vision transformer in medical image segmentation," *arXiv preprint arXiv:2206.14718*, 2022.

[19] X. Li, H. Yang, J. He, A. Jha, A. B. Fogo, L. E. Wheless, S. Zhao, and Y. Huo, "Beds: Bagging ensemble deep segmentation for nucleus segmentation with testing stage stain augmentation," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 659–662.

[20] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.