PathMamba: Weakly Supervised State Space Model for Multi-class Segmentation of Pathology Images

Jiansong Fan[1], Tianxu Lv[1], Yicheng Di[1], Lihua Li[3], Xiang Pan[1,2,4](✉)

[1] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China
[2] Engineering Research Center of Intelligent Technology for Healthcare, Ministry of Education
[3] Institute of Biomedical Engineering and Instrumentation, Hangzhou Dianzi University, Hangzhou, China
[4] Shanghai Key Laboratory of Molecular Imaging, Shanghai University of Medicine and Health Sciences, xiangpan@jiangnan.edu.cn

Abstract. Accurate segmentation of pathology images plays a crucial role in digital pathology workflow. Fully supervised models have achieved excellent performance through dense pixel-level annotation. However, annotation on gigapixel pathology images is extremely expensive and time-consuming. Recently, the state space model with efficient hardware-aware design, known as Mamba, has achieved impressive results. In this paper, we propose a weakly supervised state space model (PathMamba) for multi-class segmentation of pathology images using only image-level labels. Our method integrates the standard features of both pixel-level and patch-level pathology images and can generate more regionally consistent segmentation results. Specifically, we first extract pixel-level feature maps based on Multi-Instance Multi-Label Learning by treating pixels as instances, which are subsequently injected into our designed Contrastive Mamba Block. The Contrastive Mamba Block adopts a state space model and integrates the concept of contrastive learning to extract non-causal dual-granularity features in pathological images. In addition, we suggest a Deep Contrast Supervised Loss to fully utilize the limited annotated information in weakly supervised methods. Our approach facilitates a comprehensive feature learning process and captures complex details and broader global contextual semantics in pathology images. Experiments on two public pathology image datasets show that the proposed method performs better than state-of-the-art weakly supervised methods. The code is available at https://github.com/hemo0826/PathMamba.

Keywords: Weakly supervised · Pathology images · Multi-class segmentation· Visual mamba.

1 Introduction

Pathological images are significant for the clinical diagnosis and prognosis of diseases. In the past decade, with the vigorous development of artificial intel-

ligence technology, automatic analysis of tissue pathology images has achieved performance comparable to that of human pathologists in certain tasks [5, 17]. However, most methods are based on supervised learning, and their performance mainly relies on many training samples with detailed annotations [6, 7]. These annotations often require an experienced pathologist, are expensive to obtain, and are prone to human error.

Compared with supervised and unsupervised learning, weakly supervised learning algorithms only require coarse-grained annotations to perform language semantic segmentation [11,13] automatically. Therefore, weakly supervised learning algorithms that achieve a good balance between cost and accuracy are a promising approach. According to the degree of coarse-grained labeling, the annotation of weakly supervised image segmentation can be divided into image annotation [11], bounding box annotation [15], and point annotation [18]. Among these annotations, image-level label annotation has the lowest cost and the broadest range of applications [20]. In this work, our motivation is to segment cancer regions at the pixel level of histopathology images and rely only on image-level labels.

However, most existing weakly supervised pathological image segmentation methods are explored based on CAM methods [9, 19]. Nevertheless, CAM-based methods face significant challenges because classification networks tend to distinguish objects by their most discriminative features, whereas segmentation tasks aim to find complete objects. Simultaneously, the spatial correlation among different locations was disregarded.

Recently, inspired by the state space model (SSM) [12], researchers developed Mamba [8] to address the bottleneck of lengthy sequence modeling. The main idea is to effectively capture long-range dependencies and improve training and inference efficiency by selecting scanning mechanisms and implementing hardware-aware algorithms. U-Mamba [14] and Vision Mamba [21] based on SSM have been used for fully supervised image classification and semantic segmentation tasks.

Inspired by the above observations, we propose a novel framework called PathMamba for weakly supervised multi-class segmentation of pathology images. Unlike methods that only consider pixels or patches, our framework comprehensively considers standard pixel-level and patch-level pathology image features. Specifically, our PathMamba first utilizes Multi-Instance Multi-Label to extract pixel-level feature maps by treating pixels as instances. Subsequently, we design a novel Contrastive Mamba Block (CMB) to study the correlation between different granularities of pathological images. Since the structured state space sequence model with selective scanning (S6) [8] can only capture the information of the scanning part of the data, it cannot handle non-causal data, such as multiple contrast information in images. To this end, we incorporate Dual-granularity Comparative Mamba (DC-Mamba), a structured state-space sequence model with contrast-selective scanning, into a patch-level encoder to achieve efficient visual representation learning. In addition, since weakly supervised methods lack supervision and are difficult to constrain the learning process, we introduce a

Deep Contrast Supervised Loss (DCL) based on deep supervision [11]. It can better utilize image-level annotations to supervise feature learning of each network layer. Finally, we adopt a lightweight decoder head to integrate dual-granularity contrastive feature sequences to predict segmentation masks.

We use two datasets, LUAD-HistoSeg [9] and BCSS-WSSS [1], to verify the effectiveness of our proposed PathMamba in weakly supervised pathology image segmentation tasks. Our method yields superior performance compared to existing state-of-the-art weakly supervised methods. In summary, the main contributions of this work are as follows:

•We present a weakly supervised state space model (PathMamba) for multi-class segmentation of pathology images using only image-level labels. To our best knowledge, this is the first work introducing Mamba to the weakly supervised image segmentation task.

•We combine contrastive learning and Visual Mamba to design a novel Contrastive Mamba Block, which can explore the coherent learning of non-causal dual-granularity features in pathology images.

•We propose a Deep Contrast Supervised Loss that enables the network to fully exploit limited annotation information.

•Experiments on two public pathology image datasets show that our proposed method can achieve better performance than state-of-the-art weakly supervised methods.

2 Methodology

Figure 1 illustrates the basic structure of the proposed weakly supervised state space model (PathMamba) for multi-class segmentation of pathology images. We present the concept of Multi-Instance Multi-Label Learning (MIML) to extract pixel-level characteristics and spatial connections. Our Contrastive Mamba Block (CMB) is designed to use the extracted pixel-level feature and the original image as inputs. The CMB, adopting Visual Mamba and contrastive learning, allocates attentional weights to pixel-level and patch-level features at each stage of learning. Meanwhile, we develop a Dual-granularity Comparative Mamba (DC-Mamba) to capture pixel-level and patches-level contrast selected spatial features, which the original Mamba model is unable to do because of its limitation in capturing non-causal information. In addition, we follow the method proposed by Jia et al. [11] and further develop a Deep Xontrast Supervised Loss to solve the problem of insufficient supervision data in weakly supervised learning. We will provide a detailed description of our components in the upcoming sections.

2.1 Generate pixel-level feature maps

Multi-Instance Learning (MIL) methods assign a set of instances as "negative" or "positive", where the group of instances is called a package. The goal is to predict both package-level and instance-level labels, but it is more commonly used to solve image binary classification problems. In this paper, We consider

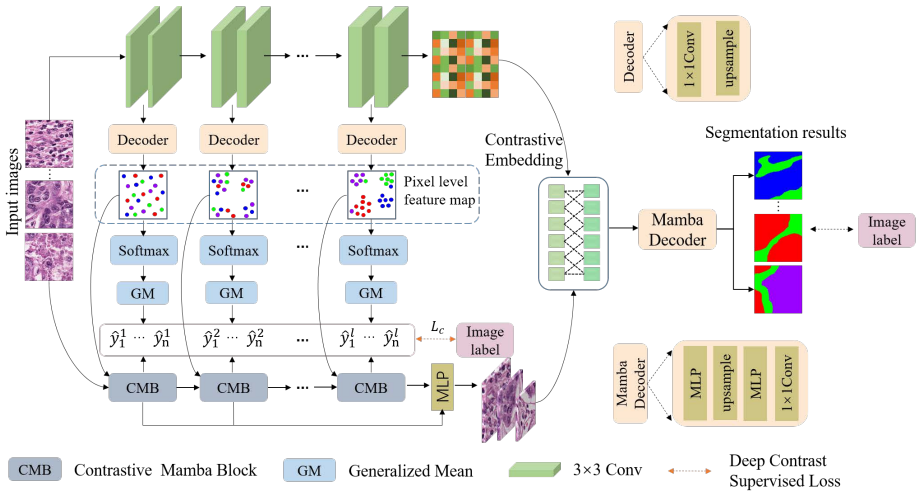

Fig. 1. The overview of the proposed PathMamba. The green block represents the MIML to generate pixel level feature process. Image label is the image-level label.

Multi-Instance Multi-Label Learning (MIML), which can perform the task of multi-label segmentation relying only on image-level labels. In our case, we refer to the image as a bag and each pixel in the image as an instance. Each bag can be associated with multiple labels. The initial three convolution stages of ResNet-50 [10], depicted in Fig. 1 as the green block, serve as the foundational component for capturing pixel-level features. In our studies, three convolution blocks in a trunk are adequate for feature extraction. The output channel size is reduced to 1 using a 1×1 convolution. A pixel-level feature map is created by restoring the image to its original size following a bilinear upsampling procedure.

2.2 Contrastive Mamba Block

Exploring non-causal dual-granularity information and global linkages is essential for pathological picture segmentation. The Transformer design can efficiently capture global information but faces significant computational challenges when dealing with excessively long feature sequences. We develop a Contrastive Mamba Block (CMB) to overcome this limitation, which can efficiently model dual-granularity information and global information.

Figure 2a demonstrates that CMB utilizes the Patch Embedding layer to partition the input pathologic image into non-overlapping patches and carry out the mapping process. The flattened sequence is normalized by Layer Normalization before being fed into the Dual-granularity Comparative Mamba module (DC-Mamba) and the deep convolutional layer. The DC-Mamba module creates contrast distinctions among several levels of diseased pictures, while the deep convolution function is designed to preserve intricate details. The original shape

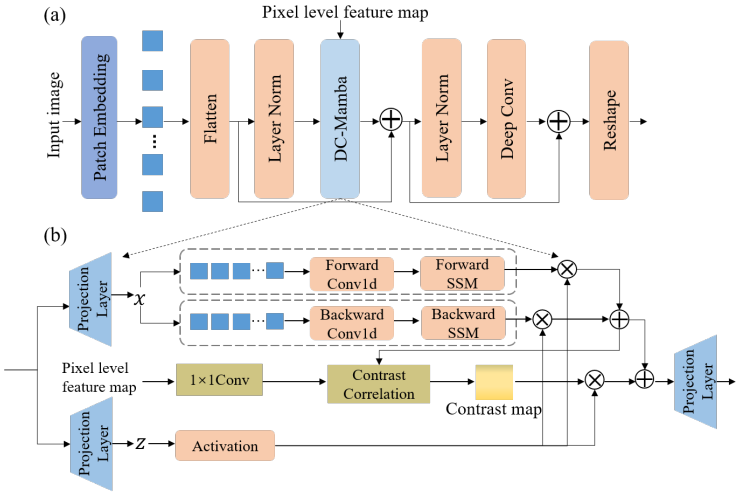

Fig. 2. Detailed Implementation of Contrastive Mamba Block.

is acquired by performing the inverse process. The process in the CMB layer can be described as:

$$\begin{aligned} \hat{h}^l &= \text{DCMamba}\left(\text{LayerNorm}\left(\phi(h^{l-1})\right)\right) + \phi(h^{l-1}), \\ h^l &= \phi^{-1}(\text{DeepConv}\left(\text{LayerNorm}\left(\hat{h}^l\right)\right) + \hat{h}^l), \end{aligned} \tag{1}$$

Where ϕ denotes the transpose and flattening operations, ϕ^{-1} represents its inverse operation, $l \in [1, N_m]$.

Dual-granularity Comparative Mamba Although S6 has causal properties for temporal data, it cannot handle multi-input non-causal information. To solve this problem, we design the DC-Mamba, which combines different granularity information of pathological images for comparative modeling, as shown in Figure 2b. Specifically, to explicitly explore the relationship between patch-level and pixel-level features, we first expand each pathology image into a sequence along four different directions through a scan expansion operation. These sequences are then processed by the S6 block for feature extraction, ensuring that information from all directions is scanned thoroughly to capture different features. Then, we design the Contrast Correlation operation to get the contrast map. Given a patch-level feature map P, the Contrast Correlation operation based on the idea of contrastive learning is defined as follows:

$$\mathcal{C}_{\mathbf{s},w} = \mathcal{N}(\left(ReLU(\text{CSM}(P, A_s) - \eta\text{CSM}(P, A_w))\right)^2), \tag{2}$$

where the CSM represents cosine similarity. The A_s and A_w denote the pixel-level feature map's vital and weak attention area vectors. The operation $\mathcal{N}$ is a two-dimensional normalization operation, $\eta \in [0, 1]$ is a positive constant.

DC-Mamba considers pathological image patch-level and pixel-level feature coherence and utilizes parallel SSM to establish long-range dependencies, thereby achieving more effective visual representation learning.

2.3 Deep contrast supervised loss

The training set is denoted by $S = \{(X_n, Y_n), n = 1, 2, 3, \ldots, t\}$, where X_n is the $n - th$ input picture and $Y_n \in \{0, 1, \ldots, m\}$ represents the label of the $n - th$ input image. $Y'_n(i, j)$ represents the probability of the pixel at location $p_{\mathrm{i,j}}$ in the prediction of the $n - th$ image. The output of the Contrastive Mamba Block is denoted as Y''_n. Thus, the image-level prediction can be defined as:

$$\hat{Y_n} = \xi \left(\frac{1}{|X_n|} \sum_{i,j} \left[\hat{Y'_n}(i, j) \right]^r \right)^{\frac{1}{r}} + (1 - \xi) \left(\frac{1}{|X_n|} \sum_{i,j} \left[\hat{Y''_n}(i, j) \right]^r \right)^{\frac{1}{r}}, \quad (3)$$

The parameter r controls clarity and proximity to the hard function: $Y \to \max_{i,i}$, $Y_n(i, j)$ as $r \to \infty$, which is used to control loss weights. The model is trained by minimizing the loss between the output prediction and the true situation. Deep Contrast Supervised Loss designed in the form of a cross-entropy loss function:

$$L_c = \sum_{u=1}^{m} \sum_{v=1}^{m} \left(- \sum \left(\mathbf{I}\left(Y_n = u\right) \log \hat{Y}_n + \mathbf{I}\left(Y_n = \nu\right) log(1 - \hat{Y}_n) \right) \right), \quad (4)$$

where $\mathbf{I}(\cdot)$ is the indicator function.

3 Experiments

Dataset: To demonstrate the effectiveness of our proposed PathMamba, we evaluate our weakly supervised segmentation method for multi-class pathology images using two datasets, LUAD-HistoSeg [9] and BCSS-WSSS [1]. The LUAD-HistoSeg dataset has 31,826 pathology pictures sized at 224 x 224. The sample consists of four tissue classes: tumor epithelium (TE), tumor-associated mesenchyme (TAS), necrosis (NEC), and lymphocyte (LYM). The BCSS-WSSS collection has 17,286 pathologic pictures sized at 224 x 224. The dataset contains detailed annotations at the pixel level for each Region of Interest (ROI) in five distinct categories: tumor (TUM), stroma (STR), lymphocytic infiltration (LYM), necrosis (NEC), and other (OTR).

Competing Methods and Evaluation Metrics: We assess the suggested method by comparing it with weakly supervised segmentation methods such as HistoSegNet [2], SC-CAM [3], C-CAM [4] and WSSS-Tissue [9]. Simultaneously, it is compared with the fully supervised UNet [16] method. All methods are evaluated using four metrics: Dice, IoU, Recall, and Precision.

Implementation Details: Our experiments are performed on an NVIDIA GeForce RTX A6000 GPU with 48G memory and repeated five times to calculate the average. The model is trained using the Adam optimizer with a weight decay $5e-4$ and a fixed learning rate of $1e-5$. We set the batch size to 16 and the training epochs to 100. The parameter r of the generalized mean function is set to 4, and the parameter ξ controlling the loss weight is set to 0.5.

Table 1. Quantitative comparison with state-of-the-art methods on the BCSS-WSSS and LUAD-HistoSeg datasets. mP: mPrecision; mR: mRecall.

Model	Supervision	BCSS-WSSS				LUAD-HistoSeg			
		mDice	mIoU	mP	mR	mDice	mIoU	mP	mR
HistoSegNet [2]	Weakly supervision	0.505	0.276	0.582	0.571	0.641	0.478	0.654	0.662
SC-CAM [3]	Weakly supervision	0.729	0.663	0.742	0.733	0.715	0.641	0.756	0.761
C-CAM [4]	Weakly supervision	0.645	0.559	0.666	0.657	0.682	0.572	0.703	0.708
WSSS-Tissue [9]	Weakly supervision	0.767	0.697	0.786	0.758	0.818	0.756	0.826	0.822
UNet [16]	Fully supervision	0.785	0.706	**0.810**	0.788	0.840	**0.770**	0.862	**0.855**
Our Model	Weakly supervision	**0.789**	**0.712**	0.804	**0.798**	**0.842**	0.767	**0.870**	0.853

Comparison with SOTA Methods: Table 1 reports a quantitative comparison of the proposed method with recent state-of-the-art methods. Experiment results show that our method outperforms other models in testing (Figure 3). We attribute this to the global modeling capabilities of Contrastive Mamba Block and its ability to characterize pathological images through contrastive learning at two granularities of patches and pixel levels. Specifically, Our method surpasses the current leading weakly supervised segmentation method WSSS-Tissue by 2.2% and 2.4% in Dice score on both datasets. Additionally, it may be noted that the performance of our PathMamba is nearly equivalent to that of a fully monitored U-Net. This could be attributed to two factors: 1) PathMamba considers two different granularities of information from pathological images simultaneously, which is helpful for pathological image segmentation. 2) The suggested Deep Contrast Supervised Loss effectively captures the sufficient semantic information of each layer within the model's middle section and enhances segmentation performance. In summary, the proposed framework can produce accurate prediction masks using only image-level labels of pathology images. There is no need for dense pixel-level annotation on pathology images.

Ablation Study: We conduct ablation studies to determine the efficiency of individual components and identify the best settings. Table 2 displays the impact of including several components on segmentation performance, such as Multi-Instance Multi-Label Learning (MIML), Contrastive Mamba Block (CMB), and

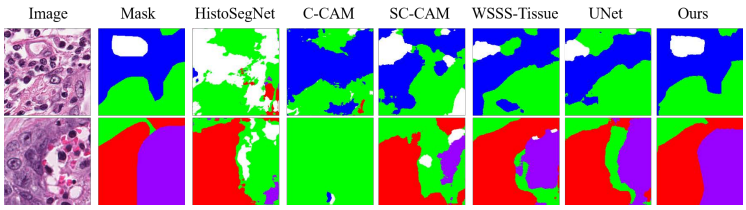

Fig. 3. Visualization of the prediction results of our proposed PathMamba compared with the State of the Art.

Deep Contrast Supervised Loss (DCL). Introducing pixel-level feature maps created by MIML (model b) improves Dice, demonstrating the benefit of pixel-level granularity. Models c and d demonstrate that incorporating CMB enhances segmentation accuracy, showcasing the effectiveness of global modeling and dual-granularity contrastive learning for feature representation. Moreover, the model f's performance is optimized with the use of DCL. We suppose it is because DCL enables the network to extract target information effectively without relying on pixel-level labeling. The results indicate that the weakly supervised segmentation approach for multi-class pathology images outperforms the single-granularity baseline when paired with Contrastive Visual Mamba.

Table 2. Ablation analysis of different components in the proposed PathMamba on LUAD-HistoSeg. mP: mPrecision; mR: mRecall. CMB: Contrastive Mamba Block. DCL: Deep Contrast Supervised Loss. MIML is the operation that generates pixel-level feature maps.

Model name	Method	mDice	mIoU	mP	mR
a	B	0.765	0.701	0.780	0.774
b	B+MIML	0.788	0.719	0.801	0.794
c	B+CMB	0.802	0.746	0.817	0.808
d	B+MIML+CMB	0.816	0.758	0.838	0.825
f	**B+MIML+CMB+DCL(Ours)**	**0.842**	**0.767**	**0.870**	**0.853**

B: baseline.

4 Conclusion

In this paper, we propose a novel weakly supervised learning method using only image-level labels, which explores pixel-level and patch-level standard features of histopathology images via Multi-Instance Multi-Label Learning (MIML) and Contrastive Mamba Block (CMB). MIML adaptively captures pixel-level features from the images to capture pixel-level features, and CMB learns granular contrast features and global relationships with less computational overhead. In

addition, we use Deep Contrast Supervised Loss to utilize under-annotated information better. Experiments show that our proposed framework has the potential to be an effective means of annotating pathology images in clinical applications.

Acknowledgments. This work is supported in part by the National Key R&D Program of China under Grants 2018YFA0701700 and 2021YFE0203700, the Postgraduate Research & Practice Innovation Program of Jiangsu Province SJCX22_1106, and is supported by National Natural Science Foundation of China grants U21A20521 and 62271178, Zhejiang Provincial Natural Science Foundation of China (LR23F010002), Jiangsu Provincial Maternal and Child Health Research Project (F202034), Wuxi Health Commission Precision Medicine Project (J202106), Jiangsu Provincial Six Talent Peaks Project (YY-124), and the construction project of Shanghai Key Laboratory of Molecular Imaging (18DZ2260400).

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. Chang, Y.T., Wang, Q., Hung, W.C., Piramuthu, R., Tsai, Y.H., Yang, M.H.: Weakly-supervised semantic segmentation via sub-category exploration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8991–9000 (2020)
4. Chen, Z., Tian, Z., Zhu, J., Li, C., Du, S.: C-cam: Causal cam for weakly supervised semantic segmentation on medical image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11676–11685 (2022)
5. Cheng, H.T., Yeh, C.F., Kuo, P.C., Wei, A., Liu, K.C., Ko, M.C., Chao, K.H., Peng, Y.C., Liu, T.L.: Self-similarity student for partial label histopathology image segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. pp. 117–132. Springer (2020)
6. Das, R., Bose, S., Chowdhury, R.S., Maulik, U.: Dense dilated multi-scale supervised attention-guided network for histopathology image segmentation. Computers in Biology and Medicine p. 107182 (2023)
7. Graham, S., Vu, Q.D., Jahanifar, M., Raza, S.E.A., Minhas, F., Snead, D., Rajpoot, N.: One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification. Medical Image Analysis **83**, 102685 (2023)
8. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)

9. 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)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Jia, Z., Huang, X., Eric, I., Chang, C., Xu, Y.: Constrained deep weak supervision for histopathology image segmentation. IEEE transactions on medical imaging **36**(11), 2376–2388 (2017)
12. Kalman, R.E.: A new approach to linear filtering and prediction problems (1960)
13. Li, K., Qian, Z., Han, Y., Eric, I., Chang, C., Wei, B., Lai, M., Liao, J., Fan, Y., Xu, Y.: Weakly supervised histopathology image segmentation with self-attention. Medical Image Analysis **86**, 102791 (2023)
14. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
15. Rajchl, M., Lee, M.C., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M.A., Hajnal, J.V., Kainz, B., et al.: Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. IEEE transactions on medical imaging **36**(2), 674–683 (2016)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
17. Skrede, O.J., De Raedt, S., Kleppe, A., Hveem, T.S., Liestøl, K., Maddison, J., Askautrud, H.A., Pradhan, M., Nesheim, J.A., Albregtsen, F., et al.: Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. The Lancet **395**(10221), 350–360 (2020)
18. Zhao, T., Yin, Z.: Weakly supervised cell segmentation by point annotation. IEEE Transactions on Medical Imaging **40**(10), 2736–2747 (2020)
19. 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)
20. Zhou, Y., Zhu, Y., Ye, Q., Qiu, Q., Jiao, J.: Weakly supervised instance segmentation using class peak response. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3791–3800 (2018)
21. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)