Segmentation of Pathological Primitive Conditions Based on the Visual Foundation Model

Prashnatita Pal
Dept. of Electronics & Communication Engineering
St Thomas College of Engineering & Technology
Kolkata
prashnatitap@gmail.com

Jayanta Poray
Dept. Computer Science & Engineering
Techno India University
Kolkata
jayanta.p@technoindiaeducation.com

Rituparna Bhattacharya
Dept. Computer Science & Engineering
Techno India University
Kolkata
rituparna.b@technoindiaeducation.

Abstract - Due to distinct picture properties and domain-specific problems, notably in pathology, medical data image processing needs a model carefully trained and prepared datasets. Objective and automated cancer diagnosis and prognosis need primitive detection and segmentation in digital tissue samples. Recent technology called SAM (Segment Anything Model) can accurately segment generic things from natural photos, but it needs human cues to build masks. We describe a unique method for detection-based region suggestions using pre-trained SAM natural image encoders. Cascaded feature propagation layers project pre-trained encoder-proposed regions. Local semantic and global context from multi-scale is used to localize and classify bounding boxes. Finally, the SAM decoder generates a complete basic segmentation map using bounding boxes as prompts. SAM, the underlying framework, can complete two pathological segmentation tasks without training or fine-tuning. While being efficient, our strategy competes with models in F_1 score for nucleus identification and binary or multiclass panoptic quality for segmentation quality on the PanNuke dataset. Our model outperforms Faster RCNN on the secondary dataset with Average Precision.

Keywords - Nuclei detection, glomeruli detection, primitive segmentation, bounding box detection.

1. INTRODUCTION

The evolving landscape of medical image processing calls for automation to mitigate workforce shortages and escalating analysis costs. Machine learning, particularly deep learning, offers promise for efficient disease detection and diagnosis. However, acquiring an adequate, well-annotated mask for supervised training remains a challenge.

In our study, we propose a novel approach, an adaptation of the SAM [1], specifically tailored for segmentation tasks in digital pathology. SAM revolutionizes image segmentation by generating high- quality object masks from prompts, handling diverse inputs, and ex- celling in zero-shot scenarios. Our methodology involves generating bounding boxes from SAM's encoder layers and using them as prompts for SAM's pretrained lightweight mask decoder. This approach mitigates annotation complexity and time constraints. Our innovative technique aims to streamline the annotation process and ease the burden on pathologists by requiring significantly less time to draw bounding boxes around nuclei compared to the exhaustive annotation of nuclear boundaries for training while offering fine- grained segmentation masks during inference time.

Our contributions extend to three distinct areas:

- An innovative feature extraction methodology is introduced to extract salient features from every layer of the transformer encoder, thus elevating overall performance.
- A distinctive architectural design is adopted that amalgamates the transformer encoder with a bottom-up configured convolutional neural network (CNN) decoder, enhancing the robustness of both detection and classification processes.
- An end-to-end network using SAM is presented to directly out- put classified objects in the form of bounding boxes and segmentation masks, eliminating the need for post-processing steps.

Related works included in Section 2. Sections 3-5 describe the methodology, experiments, results and discussions. A conclusion, a prognosis for the future, and recommendations for further effort in section 6 rounded out our work.

2. RELATED WORKS

The nuclei detection and segmentation domain primarily rely on U- Net-like [2] architectures featuring encoders and decoders. Encoders often adopt established structures like ResNet [3], transformers [3], or as seen in Hover-Net [4] and M-RCNN [5]. Feature extraction strategies encompass direct or bottom-up approaches. Recently, state-of-the-art methods [4] have improved segmentation by combining binary, distance, and nuclei-type maps, but are computationally expensive with multiple decoders. For natural images, DETR [6] uses a transformer-encoder for bounding box pre- diction and utilizes hybrid encoders and transformer decoders for bounding box prediction. Nonetheless, DETR's 100 bounding box limitation poses issues with medical images, which can easily contain numerous nuclei instances. Adapting DETR to these images demands intricate and resource-intensive quantification. While CNN methods like [5] display promise, they rely on both bounding boxes and segmentation masks for training, which requires a vast amount of manual annotation. We present a novel method that can directly use domain-agnostic encoder features for all tasks while reducing fine-tuning overhead. This novel approach taps into diverse domains to alleviate the challenge of scarce annotated medical data, revealing a new dimension in pathological primitive detection.

3. METHODOLOGY

Our foundational network, SAM [7], employs bounding boxes, points, text, or masks as prompts to generate segmentation masks, necessitating human intervention throughout the process. Moreover, its architecture lacks end-to-end capabilities, hindering streamlined analysis. We propose an extension to SAM that automatically generates segmentation prompts in the form of bounding boxes within the network and feeds them into the mask decoder. Our choice of bounding boxes as prompts stems from their superior

The performance observed in our initial experiments. In our approach, we completely freeze SAM's encoder and decoder, significantly reducing the trainable parameters of the entire framework. In prac- tice, only the bounding box predictor component is trained. This approach offers several key advantages: 1) it leverages the high- quality feature representation learned from natural images, reducing training time; 2) it seamlessly integrates bounding box detection and instance segmentation models, eliminating the need for pre- or post-processing.

3.1. Frozen Transformer Encoder

For our feature extraction backbone, we use the encoder from SAM [7], namely the lightweight SAM-B model. SAM has been trained on natural images and used for prompted generation. We propose that rich feature representation learned from natural images is sufficient for detection in the medical domain. As a result, we freeze the encoder entirely during training and use the embeddings from each layer for our decoder network.

3.2. Bounding Box Decoder Network

The decoder network in our proposed model exhibits a feature propagation process, which involves dividing the encoder into four dis- tinct blocks, denoted as $\boldsymbol{Bi}$ where i 1, 2, 3, 4. The selection of these blocks is informed by the location of global attention applied in the encoder blocks. Furthermore, we introduce projection layers denoted as $\boldsymbol{p_j} \in \{1,2,3,4,5,6\}$ in our decoder, corresponding to each individual layer's feature from $\boldsymbol{B_i} \in \{1,2,3,4\}$ We utilize 6 projection layers because empirically it stabilizes our decoder network. In total, our decoder consists of six projection layers, where we perform up-sampling on the first projection layer $\boldsymbol{p_1}$ and down- sampling in the projection layer $\boldsymbol{p_3}$ and $\boldsymbol{p_4}$, while keeping $\boldsymbol{p_2}$ un- changed. Each projection layer comprises a set of three convolutional layers with their corresponding up and down sampling and batch normalization, accompanied by a dense layer to increase the receptive field.

In the projection layer, feature aggregation occurs bottom-up, wherein the features from the three individual layers $\boldsymbol{l}$ are combined into one feature map. The dense layer includes a single convolutional operation, followed by a rectified linear unit (ReLU) activation function and batch normalization. After projection, we get the regression and classification logits $\boldsymbol{z_j} \in \{1,2,3,4,5,6\}$. coder architecture outlined above in Fig. 1 contributes significantly to our model's ability to capture and refine hierarchical features from the encoder. By aggregating features from multiple blocks and employing dense layers, we facilitate the integration of multi-scale in- formation. The skip connections and feature maps also enhance the network's capacity to handle varying object sizes and scales, resulting in more accurate and context-aware detection.

In summary, we can define the whole process as follows:

$$z'\, l = f_{enc}(x_i) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

$$z_j \;\; = p_j\,(z\,(i',\, i'+3) \in l) \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

where $\boldsymbol{f_{enc}(.)}$ is the transformer encoder, $\boldsymbol{x_i}$ is the input image, $\boldsymbol{z'}$ is the extracted features from encoders, $\boldsymbol{i'}$ is the global attention index, $\boldsymbol{z_j}$ is the output from the projection layer $\boldsymbol{p_j}$, these projections are then fed into the regression and classification network.

$$y_{b,c} = f_{det}\,(z_j), \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

$$y_{mask} = f_{seg}\,(zj',\, y_{b,c}), \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

where $\boldsymbol{f_{det}(.)}$ is the bounding box regression and classification head, $\boldsymbol{f_{seg}}$ **(.)** is the mask segmentation head, $\boldsymbol{z_j'}$ is the last layer's output from the encoder, $\boldsymbol{y_{b,c}}$ is the classified bounding box. The bounding box and classification decoder consist of three core com- ponents: anchor generation, bounding box regression, and classification. We adapt Retina Net's [8] anchor box generation technique, utilizing six labels instead of five, generated with aspect ratios [1:2, 1:1, 2:1].

3.3. Frozen Mask Decoder

The decoder component of the SAM [7] is meticulously designed to deliver high-quality segmentation results, mainly focusing on producing refined segmentation masks near object boundaries. The de- coder is trained with the encoder on natural images to achieve this. The decoder prioritizes high-quality segmentation, emphasizing re- fined masks at object boundaries. Building upon the encoder's al- ready learned feature representation, we apply a frozen approach for the decoder. This strategic decision further limits the number of trainable parameters in our network.

3.4. Loss Function

Finally, we utilize the focal loss for training our network. Focal loss is a specialized loss function specifically designed for handling class imbalance in object detection and segmentation tasks. It addresses the

common issue where the majority of the pixels or anchor boxes in medical images correspond to the background class, leading to an imbalanced distribution between foreground and background samples.

Mathematically, the focal loss is expressed as:

$$\text{Focal loss} = -\alpha\,(1 - p_t)\,\gamma \log\,(p_t), \ldots\ldots\ldots\ldots (5)$$

where ***pt*** is the predicted probability of the ground truth class, ***α*** and ***γ*** is the class balancing and modulating factor. In our practice we set ***α*** to 0.5 and ***γ*** to 2.0.

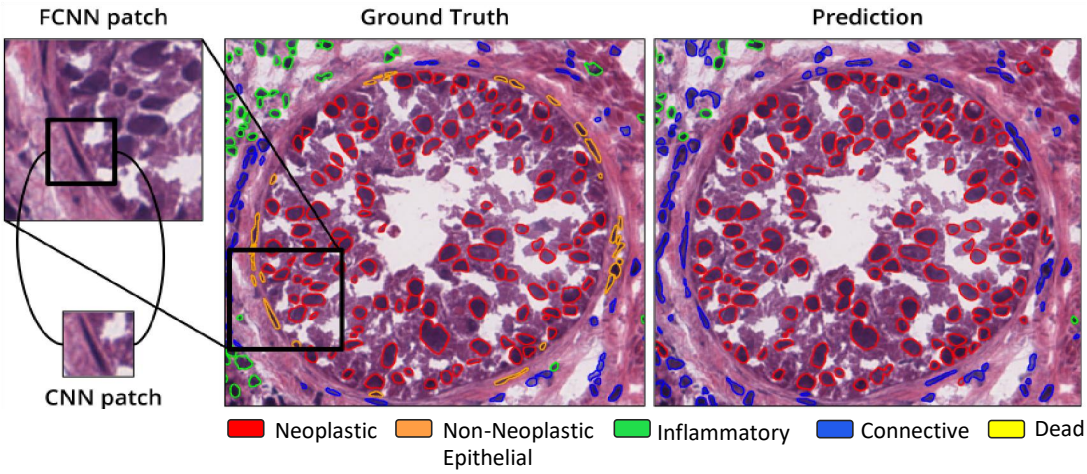

Fig: 2 Model prediction for bladder tissue visual field, pathologists have validated ground truth labels and segmentation masks. The 224×224 fcnn patch is often used for training fully convolution segmentation models. The patch below it is the one that [35] used to train a cnn to categorize each individual nucleus.

4. EXPERIMENTS

For our investigation, we utilized the PanNuke[9] dataset as our primary dataset for training and evaluating our model. This dataset encompasses 7,904 images, each measuring 256 × 256 pixels, and contains 189,744 meticulously annotated nuclei. These nuclei span across 19 diverse tissue types. They are categorized into 5 distinct cell categories: connective, inflammatory, dead/necrotic, non-neo epithelial, and neoplastic, as illustrated in Fig. 2, captured at a magnification of ×40. The cell images preserve a high level of resolution and intricate details of both nuclei and cells.

We also conducted a comprehensive evaluation utilizing a dataset obtained from HuBMAP[10], with a specific focus on glomeruli segmentation across 15 whole slide images. The dataset consists of 6694 patches extracted at a resolution of 2048x2048 pixels. Employing a three-fold cross-validation approach, 80% of the data was dedicated to training and validation, while the remaining 20% was exclusively reserved for testing.

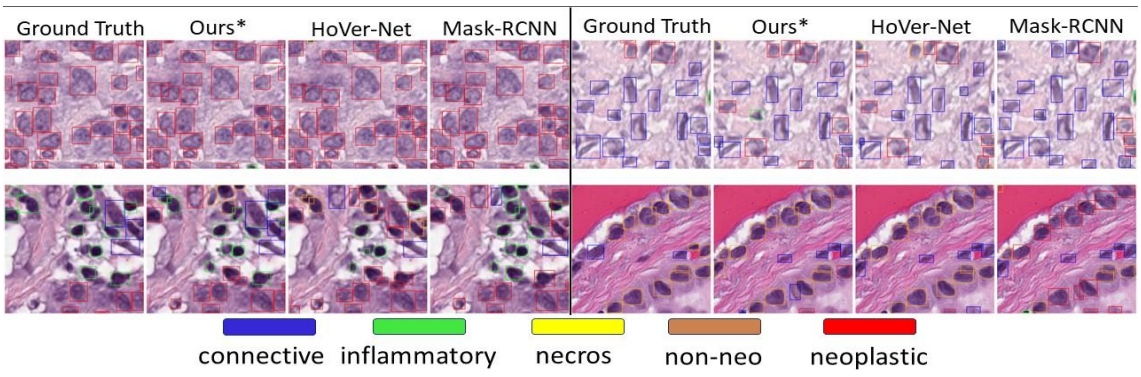

Fig. 3. Bounding box detection quality and nuclei boundary visualization.

A noteworthy aspect of our primary dataset is the PanNuke dataset's inherent imbalance, with certain classes, namely dead and non-neo epithelial classes being highly underrepresented, exhibiting considerable underrepresentation. The dataset's imbalance classes make it harder for models to learn and predict.

4.1. Experimental Setup

Our network undergoes 50 epochs with an initial learning rate of 3e-4, progressively reduced to 3e-7 through a reduction-on-plateau scheduler. Augmentations include rotation, flipping, noise injection, and color jittering. We implement a three-fold training approach with a holdout validation set.For our primary dataset, model evaluation is conducted on a 20% test set randomly sampled from the three folds. This sampling ensures each fold contains an equal representation of the smallest class.

TABLE 1. COMPARING BPQ AND MPQ SCORES OF OUR METHOD AGAINST BOUNDING BOX-BASED DETECTION FOR MASK QUALITY

Network	bPQ	mPQ
Mask-RCNN	0.5589	0.3688
(Ours) Auto-Prom	**0.6380**	**0.2979**

Table I

TABLE II. BOUNDING BOX QUALITY ANALYSIS RESULT ON THE SECONDARY DATASET. WE USE WIDELY ADOPTED AVERAGE PRECISION (AP) FOR THE QUALITY ANALYSIS BETWEEN THE TWO MODELS.

Method	Average Precision
Faster R-CNN (Res-101)	74.34 %
(ours) Auto-Prom	**76.73 %**

Table II

4.2. Baseline and Evaluation metrics

In this study, we conducted three experiments. One of them is for bounding box prediction accuracy, which is done by comparing f_1 accuracy. We also experiment with segmentation quality using panoptic quality analysis. Additionally, for our baseline comparison, we employ the baseline metrics provided by PanNuke[9] dataset organizers. All these metrics on the primary dataset for our network are acquired from the test split. In our comparative evaluation for mask quality analysis, we utilize Mask-RCNN as the baseline, as it stands as the sole bounding box-based detection method amongst the baselines.

In addition, our neural network undergoes training on the secondary dataset for 50 epochs for the third experiment, with an exclusive emphasis on utilizing bounding boxes. We chose Faster RCNN as our benchmark for comparing the performance of our network on the secondary dataset. We quantified the Intersection over Union (IoU) overlap at 0.5 against the ground truth bounding boxes. If an IoU overlap with the ground truth is detected, the corresponding bounding box is associated with the respective bounding box. Regarding the secondary data set, evaluation is performed on a randomly sampled test set, as detailed in Section 4.1.

5. RESULTS AND DISCUSSIONS

For binary Panoptic Quality (PQ) analysis, the multiclass mask is converted into a binary mask to facilitate PQ metric generation. Our approach achieves superior performance over Mask-RCNN in binary PQ detailed in Table 1, offering insights into the quality of masks.

The results on the F_1 accuracy are presented in Table III. Our approach achieves state-of-the-art performance in multi-class classification across four distinct categories: neoplastic, inflammatory, connective and non-neo epithelial. Furthermore, we construct a confusion matrix based on our model's predictions. Our approach demonstrates an Average Precision (AP) score of 79.89%, representing a notable improvement of 4.5% over our baseline. The generated mask for the separate test set is visually depicted in Fig. 3 for reference.

In essence, our results underscore the prowess of our approach in attaining superior multiclass classification performance and elucidate the effectiveness of our methodology in the context of binary panoptic quality analysis and multi-class detection. The confusion matrix provides a comprehensive snapshot of prediction outcomes, contributing to a holistic understanding of our model's behaviour.

6. CONCLUSION

The purpose of this research was to offer a dataset with the help of annotated and quality checked. The dataset included specific borders and class labels for five primary kinds of nuclei that were found in numerous distinct types of malignant tissue.

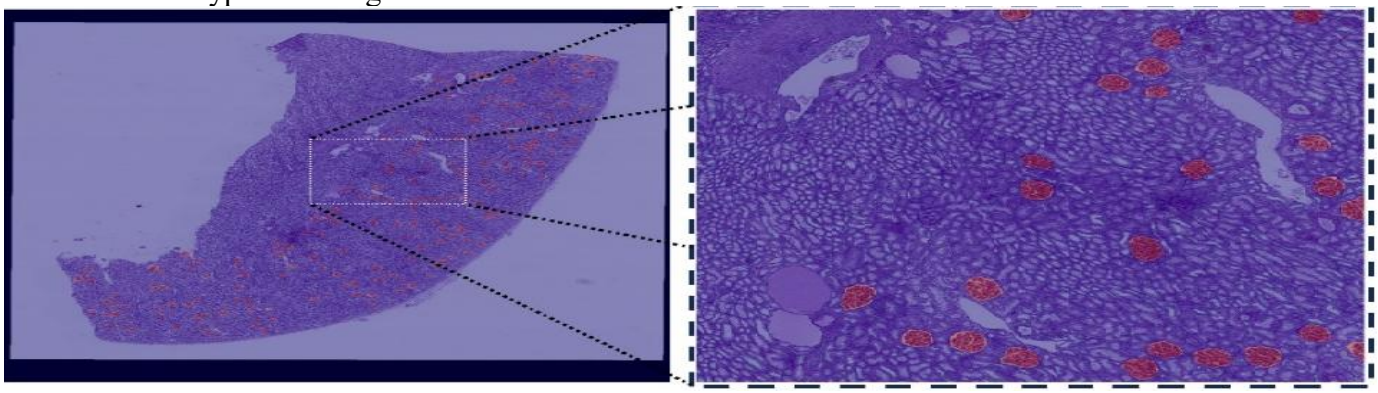

Fig. 3. Predicted mask visualization on WSI from a test set of HuBMAP kidney segmentation dataset. The red regions correspond to our model's prediction of glomeruli [11].

TABLE III. COMPARISON OF CLASSIFICATION METRICS WITH THE BASELINE MODELS. FOR THE PRECISION (P), RECALL (R) AND F1 METRIC (F1), THE CENTROID OF THE SEGMENTATION MASK IS USED. THE MODELS ARE LABELED AS FOLLOWS: M1: DET U-NET, M2: DIST, M3: MASK-RCNN, M4: MICRO-NET, AND M5: HOVER-NET.

Nuclei Labels							
		M1	M2	M3	M4	M5	**Ours**
Neoplastic	P	0.40	0.49	0.55	0.59	0.58	**0.71**
	R	0.47	0.55	0.63	0.66	0.67	**0.74**
	F_1	0.43	0.50	0.59	0.62	0.62	**0.71**
Non-Neo	P	0.27	0.38	0.52	0.63	0.54	**0.70**
	R	0.31	0.33	0.52	0.54	0.54	**0.55**
	F_1	0.29	0.35	0.52	0.58	0.56	**0.62**
Inflammatory	P	0.32	0.42	0.46	0.59	0.56	**0.63**
	R	0.45	0.45	0.54	0.46	0.51	**0.61**
	F_1	0.37	0.42	0.50	0.54	0.54	**0.61**
Connective	P	0.34	0.42	0.42	0.50	0.52	**0.55**

	R	0.38	0.37	0.43	0.45	0.47	**0.53**
	F_1	0.36	0.39	0.42	0.47	0.49	**0.53**
Dead	P	0.00	0.00	0.17	0.23	0.28	**0.41**
	R	0.00	0.00	0.30	0.17	0.35	**0.11**
	F_1	**0.00**	**0.00**	**0.22**	**0.19**	**0.31**	**0.14**

This study is prompted by the observation that the usage and validity of results in most challenge competitions is problematic owing to the restricted nature of challenge datasets. This observation is what drove this effort. For instance, even on ImageNet [12], which is a significant number of orders of magnitudes greater than [13], there is evidence of overfitting because of hypothesis testing in the process of construction. The selection bias in the availability of the models and the size of the dataset in comparison to any other study, is just a tiny step in the direction of a safe and robust use of CV in C Path. In the same vein as Esteva et al. [14], our approach delivers promising results and significantly enhances efficiency in pathological image analysis. By implementing our method, we achieve finer object boundaries without needing a ground truth segmentation dataset, relying solely on bounding boxes during training. This innovation could expedite the annotation process, allowing experts to outline bounding boxes and enabling automatic mask generation in SAM. Notably, our approach eliminates the need for prompts during inference, reducing human involvement.

8. REFERENCES

1. Mazurowski, M. A., Dong, H., Gu, H., Yang, J., Konz, N., & Zhang, Y. (2023). Segment anything model for medical image analysis: An experimental study. *Medical Image Analysis*, *89*, 102918. https://doi.org/10.1016/j.media.2023.102918
2. Zhai, X., Kolesnikov, A., Houlsby, N., & Beyer, L. (2022). Scaling vision transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr52688.2022.01179
3. Graham, S., Vu, Q. D., Raza, S. E., Azam, A., Tsang, Y. W., Kwak, J. T., & Rajpoot, N. (2019). Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, *58*, 101563. https://doi.org/10.1016/j.media.2019.101563.
4. He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2020). Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 386-397. https://doi.org/10.1109/tpami.2018.2844175.
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. *Lecture Notes in Computer Science*, 213-229. https://doi.org/10.1007/978-3-030-58452-8_13.
6. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W., Dollár, P., & Girshick, R. (2023). Segment anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/iccv51070.2023.00371.
7. Lin, T., Goyal, P., Girshick, R., He, K., & Dollar, P. (2020). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 318-327. https://doi.org/10.1109/tpami.2018.2858826.
8. Javed, S., Mahmood, A., Fraz, M. M., Koohbanani, N. A., Benes, K., Tsang, Y., Hewitt, K., Epstein, D., Snead, D., & Rajpoot, N. (2020). Cellular community detection for tissue phenotyping in colorectal cancer histology images. *Medical Image Analysis*, *63*, 101696. https://doi.org/10.1016/j.media.2020.101696.
9. Kang, Y., Liu, R., Wu, J., & Chen, L. (2019). Structural insights into the mechanism of human soluble guanylate cyclase. *Nature*, *574*(7777), 206-210. https://doi.org/10.1038/s41586-019-1584-6.
10. ang, Y., Hayat, M., Jin, Z., Zhu, H., & Lei, Y. (2023). Zero-shot point cloud segmentation by semantic-visual aware synthesis. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/iccv51070.2023.01064.
11. Ranjan, A., Singh, A. K., & Sahana, B. C. (2020). A Review on Deep Learning-Based Channel Estimation Scheme. *Advances in Intelligent Systems and Computing*, 1007–1016. https://doi.org/10.1007/978-981-15-4032-5_90.
12. Foucart, A., Debeir, O., & Decaestecker, C. (2022). Comments on "MoNuSAC2020: A multi-organ nuclei segmentation and classification challenge". *IEEE Transactions on Medical Imaging*, *41*(4), 997-999. https://doi.org/10.1109/tmi.2022.3156023.
13. Graham, S., Vu, Q. D., Jahanifar, M., Raza, S. E., Minhas, F., Snead, D., & Rajpoot, N. (2023). One model is all you need: Multi-task learning enables simultaneous histology image segmentation and classification. *Medical Image Analysis*, *83*, 102685. https://doi.org/10.1016/j.media.2022.102685
14. Hering, J., & Kybic, J. (2020). Multiple instance learning via deep hierarchical exploration for histology image classification. *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. https://doi.org/10.1109/isbi45749.2020.9098616 .

