PointNu-Net: Simultaneous Multi-tissue Histology Nuclei Segmentation and Classification in the Clinical Wild

Kai Yao, Kaizhu Huang, Jie Sun, Amir Hussain and Curran Jude

Abstract—Automatic nuclei segmentation and classification plays a vital role in digital pathology. However, previous works are mostly built on data with limited diversity and small sizes, making the results questionable or misleading in actual downstream tasks. In this paper, we aim to build a reliable and robust method capable of dealing with data from the ‘the clinical wild’. Specifically, we study and design a novel method to simultaneously detect, segment, and classify nuclei from Haematoxylin and Eosin (H&E) stained histopathology data, and evaluate our approach using the recent largest dataset: PanNuke. We address the detection and classification of each nucleus as a novel semantic keypoint estimation problem to determine the center point of each nuclei. Next, the corresponding class-agnostic masks for nuclei center points are obtained using dynamic instance segmentation. By decoupling two simultaneous challenging tasks, our method can benefit from class-aware detection and class-agnostic segmentation, thus leading to a significant performance boost. We demonstrate the superior performance of our proposed approach for nuclei segmentation and classification across 19 different tissue types, delivering new benchmark results.

Index Terms—Pathology, Object segmentation, Artificial neural networks

I. INTRODUCTION

Analyzing digital pathology images can provide valuable diagnostic and prognostic cancer indicators to prompt precision medicine. These images usually contain tens of thousands of nuclei with various types. Their distribution and appearance are essential markers for cancer study. Automated analysis of these images can quantify biomarkers and histopathological features, reduce the workload of pathologists, and standardize clinical practices. This creates a huge demand for high-throughput nuclei detection, segmentation, and classification [1]. For instance, nuclear morphometric and appearance features can deliver massive tumor information for clinical prediction, e.g., diagnosis of disease grade and type [2], disease prognosis and survival prediction [3], and cancer metastases detection [4]. Moreover, classifying nuclei instead of patches in tissue may improve the prediction of recurrence and outcome in some cases, e.g., in stage III colon cancer [5].

Convolutional neural network (CNN) has exhibited superior performance in many applications of computer vision. It has been introduced as a novel tool to process the digital pathology images and identify morphological patterns in computational pathology [6]. In general, on histology nuclei segmentation and classification, the present CNN based approaches can be categorized into either bottom-up or top-down methods. The bottom-up structure is adopted by most existing methods [7]–[11] which first generate high-resolution semantic segmentation masks and then group the pixels into an arbitrary number of object instances, as shown in Figure 1 (a). Relying on complicated pixel grouping post-processing to extract object instances, the performance of bottom-up approaches is highly dependent on segmentation results and grouping methods. Meanwhile, top-down methods, e.g., Mask-RCNN [12], first locate class-agnostic objects in prior bounding boxes to generate region proposals and then segment and classify object instances within region-of-interests (ROIs) (see Figure 1 (b)). Though Mask-RCNN can well separate touching nucleus, it can only segment the instances within bounding boxes with a
very low resolution (i.e., $28 \times 28$). Furthermore, if bounding boxes are predicted less accurately and smaller than the actual instances, top-down methods may lead to poor segmentation.

In addition to the drawbacks as mentioned above, more seriously, most of these present works are typically built and evaluated on small datasets with limited diversity [7], [9], mainly because the annotation of digital pathology images requires a large amount of time and effort from domain experts. As a result, the validity of the above proposed methods may be questionable [13] and may cause misleading results in downstream analysis. In fact, there are a number of even more difficult challenges in ‘the clinical wild’. First, the appearance of nuclei is obviously affected by manual operations such as slightly out-of-focus, imbalanced H&E staining and blur boundaries; folding tissue and hollow structure may also cause inconsistency of digital pathology images. Second, significantly morphological changes can be observed across the real images. The nuclei of a single cell type can show varied shape and size due to different diseases. Third, the density of the malignant cells in neoplastic tissues is usually much higher than the cell density in normal tissues, requiring that the analysis methods should have good generalization capabilities to process mutually occluded and overlapping nuclei under such extremely high cellularity and nuclei pleomorphism scenario. All these challenges were not properly manifested in the current small data and are consequently not addressed effectively for nuclei images in ‘the clinical wild’.

Aiming to alleviate the problems in bottom-up and top-down methods as well as building a reliable and robust method capable of dealing with challenges from the ‘the clinical wild’, in this paper, we study and design a new method to simultaneously detect, segment, and classify the nuclei from Haematoxylin and Eosin (H&E) stained histopathology data\(^1\). Specifically, different from both the bottom-up and top-down methods, we address the nuclei segmentation and classification problem from a new perspective and propose a novel keypoint-aware network which directly outputs instance segmentation and classification in histology images, as shown in Figure 1 (c). In particular, we develop a novel framework to segment and classify touching or overlapping nuclei by considering the problem as two simultaneous prediction problems. The network exploits keypoint heatmap regression to predict the center point of each instance (which inspires the model name PointNu-Net) that can detect and classify nuclei effectively; for each detected center point of instance, a high-resolution binary segmentation mask is then predicted using dynamic convolution.

To our best knowledge, this is a brand new idea which was first applied on nuclei segmentation and classification. With the novel design, our method enjoys many benefits: compared with bottom-up methods, better detection results can be obtained by keypoint estimation without complex pixel grouping post-processing; compared with top-down methods, we utilize center points to represent nucleus instead of bounding boxes which can separate touching objects more precisely, and output high-resolution masks directly via dynamic instance segmentation. Therefore, overlapping and clustered nucleus can be separated as keypoints, leading to much better nuclei detection performance. In addition, our method can directly output instance segmentation and classification without any post-processing, enabling a very fast inference.

To validate our method, we first present comparative results on various datasets including two famous multi-tissue histology image datasets. Our novel methods demonstrates state-of-the-art performance compared to other recently proposed methods on nuclei segmentation and classification tasks. Importantly, we also conduct evaluations on the PanNUke data, the recently released largest pan-cancer histology dataset containing over 20,000 WSIs for nuclei segmentation and classification [14], which is statistically similar to ‘the clinical wild’ and seldom investigated by previous methods. In such ‘clinical wild’, our proposed method leads to much robust and promising performance, which is substantially and consistently higher than all the comparison models for all the 19 different tissue types.

The rest of this paper is organized as follows. We first conduct an overview of related works. Then, we detail the proposed PointNu-Net, including network architecture, training scheme, and inference strategy. Finally, we evaluate the proposed method on three public available datasets qualitatively and quantitatively.

## II. Related Work

In this section we give an overview of the related work. We first review bottom-up and top-down nuclei segmentation and classification methods. Then we review some recently-developed keypoint-based object detection approaches.

### A. Bottom-up Nuclei Segmentation And Classification

Most histology nuclei segmentation and classification approaches are designed bottom-up. Kumar et al. [7] predicted three-class segmentation masks containing inside, outside, and edges of nuclei. Then pixels were grouped into instances using inside pixel seeded region growth. DIST [8] formulated the inside nuclei segmentation problem as a regression of distance maps. CIA-Net [10] utilized additional contour supervision to obtain segmentation with more accurate edges of nuclei. Triple U-Net [11] designed the parallel feature aggregation network to fuse features from Hematoxylin and RGB images progressively. Thus it learned a more precise nuclei boundaries. Hover-Net [9] was the first to achieve simultaneous nuclei segmentation and classification. It outputs the instance center by regressing vertical and horizontal distances of nuclear pixels to their centres of mass, and then predicted an additional nuclei type map to reach nuclei classification.

Although these approaches achieved encouraging results, the detection performance of the bottom-up methods highly depends on the segmentation results and pixel grouping methods, thus limiting their generalization in complex scenarios, e.g., in ‘the clinical wild’. In contrast, we propose to regress a keypoint heatmap independently from nuclei segmentation. Thus there is no need of pixel grouping to separate nucleus, which enables better generalization in the complex scenarios.

\(^1\) Codes available at: https://github.com/Kaiseem/PointNu-Net.
B. Top-down Nuclei Segmentation And Classification

Top-down methods, e.g., Mask-RCNN [12], have made rapid progress in natural image segmentation. Despite their potentials in handling touching and clustered nucleus, they were seldomly used in nuclei segmentation and classification. One major limitation for top-down methods is the difficulty in segmenting precise nuclei boundaries. This may be caused by several reasons. 1) Since the prior anchors were usually designed densely, one candidate region proposal may correspond to many overlapped bounding boxes. Thus, non-maximum suppression (NMS) is often performed to filter the bounding boxes with low scores. This may result in poor segmentation when the bounding boxes cannot cover the whole nuclei. 2). Region proposal methods often have a fixed resolution of output segmentation masks (e.g., 28×28). The predicted masks are then resampled to the corresponding bounding boxes’ sizes, which may also introduce quantization errors.

On the contrary, since our method utilizes the peak of keypoint heatmap to represent object instances, one candidate proposal tends to be one point only. Thus, NMS can be optional but not necessary. In addition, our method predicts high-resolution masks for all instances via dynamic convolution, where the predicted masks have the same size as the input images. This can boost the performance not only in detection but also in segmentation compared with top-down methods.

C. Keypoint-based Object Detection

Unlike anchor-based detectors which regress objects by fitting the anchor boxes, keypoint-based object detectors directly regress object locations by utilizing features at certain pre-defined object keypoint. There are mainly two types of keypoint-based object detectors, i.e., group-based and group free, in the literature. Group-based detectors predict multiple keypoints for each object and group them to get bounding boxes. CornerNet [15] detected top-left and bottom-right corners of the object and then matched corners of the same object by computing the distance of points in the feature space. Duan et al. [16] added a center detection branch to improve the performance by center point validation. RepPoints [17] took advantage of deformable convolutional networks to get sets of points to represent objects. CentripetalNet [18] improved CornerNet by predicting the centripetal shift in order to pair corner keypoints from the same object. On the other hand, group-free detectors directly predicted the center keypoints of objects so as to avoid the complex grouping process. CenterNet [19] located objects by the center points and regressed the corresponding size. SaccadeNet [20] improved center point detection by extracting more informative corner features during training.

Since cell nucleus tend to cluster with high density, we utilize the group-free keypoint-based object detector to detect the center point of nucleus efficiently and effectively. Different from the above mentioned methods, our proposed method decouples instance segmentation and classification into class-aware keypoint regression and class-agnostic mask prediction. This allows us to directly predict semantic instance masks in one shot without the need of pixel grouping post-processing, thus enabling both more precise and faster detection. To the best of our knowledge, our proposed method is the first approach that utilizes keypoint-based detection into nuclei instance segmentation.

III. PROPOSED METHOD

The main idea of PointNu-Net is to rephrase the nuclei instance segmentation as two simultaneous prediction problems. Concretely, our system represents one nuclei by a single point at their instance segmentation mask center. The corresponding instance segmentation is then regressed directly using dynamic convolution at the center location. As shown in Figure 2, our method first outputs three predictions, i.e., keypoint heatmap, kernel prediction, and feature prediction. The local peaks are then filtered from the keypoint heatmap to locate center points of nucleus. Next, the kernel vectors are selected from the kernel prediction according to the center point position. At last, the instance masks of all the identified center points
A network architecture consists of image feature extractor, feature fusion module, and three task-specific prediction branches, as shown in Figure 3. In the following, we will first describe the prediction branches including the keypoint heatmap regression for nuclei detection and classification as well as the dynamic convolution for nuclei segmentation. Afterwards, details of the employed CNN for feature extraction and fusion will be discussed. Finally, we introduce the inference process of our proposed method.

### A. Detection And Classification By Keypoints

Given an input image of width \( W \) and height \( H \), our aim is to regress a low-resolution keypoint heatmap \( \hat{Y} \in [0, 1]^{\frac{W}{R} \times \frac{H}{R} \times C} \), where \( R \) is the output downsampling factor and \( C \) is the number of keypoint types. The default output downsampling factor of \( R = 4 \) is applied following the literature [21], [22]. Keypoint types include \( C = 1 \) for nuclei detection, or \( C > 1 \) for nuclei detection and classification. A prediction \( \hat{Y}_{x,y,c} = 1 \) corresponds to a detected keypoint, whilst \( \hat{Y}_{x,y,c} = 0 \) is background.

We train the keypoint heatmap prediction branch by following [19], as shown in Figure 3. We render all ground truth keypoints in a heatmap \( \hat{Y} \in [0, 1]^{\frac{W}{R} \times \frac{H}{R} \times C} \) using an unnormalized Gaussian kernel, \( \exp(-\frac{(x-p_x)^2+(y-p_y)^2}{2\sigma^2}) \), where \( \sigma \) is an object size-adaptive standard deviation [15]. If two Gaussian of the same class overlap, we take the element-wise maximum [21]. The training objective is a penalty-reduced pixel-wise logistic regression with the focal loss [23]:

\[
L_{	ext{keypoint}} = -\frac{1}{N_{\text{pos}}} \sum_{xyc} \left\{ \begin{align*}
(1 - \hat{Y}_{xyc})^\alpha \log (\hat{Y}_{xyc}) & \text{ if } Y_{xyc} = 1 \\
(1 - \hat{Y}_{xyc})^\beta (\hat{Y}_{xyc})^\alpha \log (1 - \hat{Y}_{xyc}) & \text{ otherwise }
\end{align*} \right.
\]

where \( \alpha \) and \( \beta \) are the hyper-parameters of the focal loss, \( N_{\text{pos}} \) is the number of keypoints in image \( I \), \( \hat{Y} \) is the ground truth heatmap, and \( Y \) is the prediction. Normalization by \( N_{\text{pos}} \) is chosen to normalize all positive focal loss instances to 1. We use \( \alpha = 2 \) and \( \beta = 4 \) in our experiments, following [15] and [19].

### B. Instance Segmentation By Dynamic Convolution

We regress the full-size binary instance mask \( \hat{M}_{x,y} \in \mathbb{R}^{W \times H} \) for each point at \((x,y)\) on the keypoint heatmap and a 3D tensor with \( \frac{W \times H}{R^2} \) channels is generated. Ideally, a one-to-one correspondence can be established between the semantic keypoint heatmap and the class-agnostic mask. However, it may result in unnecessary computing redundancy and out of memory problem to directly predict such a tensor.
To overcome this problem, we take advantages of dynamic convolution [24], [25] and divide the mask branch into a feature branch and a kernel branch. For each point in the heatmap, we engage dynamic convolution to produce the final instance mask prediction, which can be written as: \(M_{x,y} = K_{x,y} \ast F\), where \(F \in \mathbb{R}^{W \times H \times E}\) is the predicted feature, \(K_{x,y} \in \mathbb{R}^{1 \times 1 \times E}\) is the convolution kernel generated by kernel branch, \(E\) is the length of kernel size, and \(\ast\) is the convolution operation.

So far, the whole instance mask \(M \in \mathbb{R}^{W \times H \times W \times H}\) is no longer required to compute. Moreover, only the kernels corresponding to the positives are selected to compute the final prediction and calculate the training loss. Inspired by Wang et al. [26], we add locations in the neighborhood of the center as positives where the corresponding heatmap is larger than a threshold \(\tau\) (e.g., heatmap \(> \tau\)). The training objective of feature branch and kernel branch is a Soft Dice loss [27]

\[
\mathcal{L}_{\text{mask}} = \frac{1}{N_{\text{pos}}} \sum_{y_k} \mathbb{1}_{\left(\max_{c} (Y)^{c}_{x,y_k}) > \tau\right)} (1 - \text{DICE}(\hat{M}^{c}, M^{k}))
\]

where \(N_{\text{pos}}\) is the number of instance masks, DICE is Dice coefficient, \(M^{c}\) is the predicted soft mask, \(k\) is the \(k^{th}\) instance in image \(I\), and \(M^{k}\) is the ground truth mask corresponding to \(k^{th}\) instance. \(I\) is the indicator function, being 1 if \(\max_{c}(Y)^{c}_{x,y_k}) > \tau\) and 0 otherwise.

The overall objective is:

\[
\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{keypoint}} + \lambda_{\text{mask}} \mathcal{L}_{\text{mask}}.
\]

We set \(\lambda_{\text{mask}} = 1\) and \(\tau = 0.5\) empirically in all our experiments unless specified otherwise. We apply a single network to predict semantic keypoint \(\hat{Y}\), dynamic convolution kernel \(K\) and feature \(F\) simultaneously with three separated branches (shown in Figure 3), and generate class-agnostic instance masks \(\hat{M}\) dynamically.

**C. Feature Extraction And Fusion**

Standard feature extraction, like ResNet [28] and VGGNet [29], cannot extract a strong and representative set of features for small objects, since they encode the input image as a high-level low-resolution representation in a series manner. However, the size of nucleus varies among the tissues, and many nucleus are very small (less than \(8 \times 8\) pixels). This may result in unsatisfactory detection and segmentation for small nucleus. To overcome this problem, we take High Resolution Net (HRNet) [30] as our backbone which can maintain high-resolution representations as well as extracting low-resolution high-level features through the whole process. As shown in Figure 3, we omit HRNet structure for clarity, and Conv2-5 denotes multi-scale features. Different scales of features are upsampled to the same size as ‘Conv2’ using bilinear interpolation and concatenated together.

In order to better integrate the multi-scale features from the backbone, we design a simple module called Joint Pyramid Fusion Module (JPFM) to make sure each branch can make full use of the maintained high-resolution representations from the backbone, as shown in Figure 4. Since the multi-scale features have been fused repeatedly in HRNet, there is limited gain using the fusion module with a large receptive field. In contrast, we engage a dense design of dilated convolution \((d = 1, 2, 4, 8)\) to extract task-specific representation for each branch. We do not use a shared fusion block for all the branches since the importance of different scale for different branch is not the same. Instead, we use the unshared JPFM for each branch, and we will discuss it in the later section with detailed experiments.

Finally, Conv Blocks are stacked in each branch to generate final prediction, while Upsample Blocks are used in feature branch to output high-resolution features. Followed by [25], a normalized coordinates is added at the first Conv Blocks of the kernel branch to embed position information.

**D. From Points To Instance Segmentation**

At inference time, as shown in Figure 2, we first extract the peaks in the heatmap for each category independently. We detect all responses whose value is greater or equal to its 8-connected neighbors and a confidence threshold of 0.4 is used to filter out predictions with low confidence. Then the top 100 scoring local peaks and their corresponding kernels are selected, and dynamic convolution is performed between the predicted feature and selected kernels. Finally, a threshold of 0.5 is used to convert predicted soft masks to binary masks. Matrix NMS operation [25] is alternatively used to get rid of overlapped predictions.

**IV. Experiments**

**A. Datasets**

PanNuke is so far the largest publicly available nuclei segmentation and classification dataset [14], [31]. Images with 19 different tissues were obtained from The Cancer Genome Atlas (TCGA), where 256 \(\times\) 256 pre-extract patches were collected from more than 20,000 Whole Slide Images. PanNuke is split into 3 randomized training, validation, and testing folds, and we follow [14] to reproduce the results. This dataset is semi-automatically annotated and quality controlled by clinical pathologists, which is statistically similar to ‘the clinical wild’ and with minimal selection bias.

Kumar is a common nuclei segmentation dataset [7] consisting of 30 images from seven different tissues. They are divided into a training set of 16 images (4 breast, 4 liver, 4 kidney and 4 prostate) and a testing set of 14 images (2 breast, 2 liver, 2 kidney and 2 prostate, 2 bladder, 2 colon, 2 stomach) with the same protocol used in [7]–[9].

ConNSEp [9] containing 41 images with different cell types. Four classes are considered: epithelial, inflammatory, spindle-shaped, and miscellaneous. Training and test set partition follows previous works, where the training set contains 27 images, and the test set contains 14 images.

**B. Implementation And Training Details**

We implemented our framework with the open source software library PyTorch 1.8.0 on a workstation equipped with one NVIDIA GeForce RTX 3090 GPU. For Kumar dataset, we used stain normalization [32] to reduce the color differences.
between the stained images, and no stain normalization was performed for CoNSeP and PanNuke since the color difference between training and testing data is not large. The training objective function consists of the keypoint regression loss $L_{keypoint}$ and instance segmentation loss $L_{mask}$, and the weights for each loss are both set to 1. AdamW [33] has been used as an optimizer to minimize objective function with the mini-batch of 8 and weight decay of 0.0001. All models are trained for 100 epochs with an initial learning rate of 0.0001, which is then divided by 10 at the 80th and again at the 90th epoch. Various data augmentation techniques were employed including random cropping, flipping, color jittering, blurring and elastic transformation by following HoVer-Net [9].

C. Results And Comparative Analysis

To quantify the instance segmentation performance of each method, we used panoptic quality (PQ), followed by Graham et al. [9]. Panoptic quality was further tear apart into Detection Quality (DQ) and Segmentation Quality (SQ) components for interpretability. We used multi-class PQ ($mPQ$) to evaluate the performance of segmentation and classification, while binary PQ ($bPQ$) was used to evaluate the performance of segmentation. Specifically, $mPQ$ was calculated independently for each positive class, and $bPQ$ assumes that all nucleus belong to one class. For Kumar and CoNSeP, we also report DICE and Aggregated Jaccard Index (AJI) [7] in direct comparison with previous works on small datasets.

On large dataset PanNuke, the proposed method was compared against several deep learning based methods, e.g., DIST [8], Mask-RCNN [12], Micro-Net [34] and HoVer-Net [9]. As shown in Table I, the proposed method achieves state-of-the-art performance not only on classification but also on segmentation. In particular, PointNu-Net outperformed the best method HoVer-Net, by 0.0328 and 0.0213 on the average $mPQ$ and $bPQ$ across tissues respectively. To take insight into how PointNu-Net performed for different types of nucleus, we reported $mPQ$ and $bPQ$ for all 19 tissue types separately. We observed that for most tissue types, PointNu-Net achieved the best performance against all the existing methods.

In some cases, distinguishing between neoplastic and non-neoplastic nuclei proved to be challenging, yet an improvement of 0.027 and 0.086 on the average PQ on the both nuclei can be observed in our method. In addition, when facing the complicated scenarios, e.g., out-of-focus, blur boundaries, and imbalanced staining, previous methods, e.g., HoVer-Net, 2Evaluation code: https://github.com/TIA-Lab/PanNuke-metrics
might perform poorly since there is no confidence mechanism to filter the poor instance results. In contrast, our method outputted instances with high confidence, resulting in more robust performance in 'the clinical wild'.

On small datasets Kumar and CoNSeP, we evaluated the performance of instance segmentation with the above mentioned methods. We also made comparisons with some additional methods. As observed from the results in Table III, our method generated the highest detection quality (DQ) on Kumar and CoNSeP, as PointNu-Net had better detection performance. On the other hand, our method resulted in marginally lower segmentation quality (SQ) and did not exhibit significant advantages on segmentation. This can be explained by the fact that our method only outputs the instances with high confidence, whilst the instance with low confidence may be ignored. This can be also reflected on DICE which measures the semantic segmentation results only. In addition, since AJI may over-penalize the overlapping region according to Graham et al. [9], panoptic quality (PQ) is more fair to evaluate each method. As a consequence, PointNu-Net had a relatively low AJI (since our method allows overlapped outputs), while it has comparable results on Kumar dataset, w.r.t. bPQ. The superior performance of the proposed method can be observed in the last column of Table III, i.e., multi-classes panoptic quality (mPQ). The keypoint-aware framework demonstrated superior performance of detection and classification of nuclei, making an improvement of 0.02 mPQ on CoNSeP over the best result given by HoVer-Net. It is noted that no mPQ could be available on Kumar since no classification annotation is conducted in the dataset. Visualization of nuclei segmentation results can be found in Figure 7.

D. Speed-accuracy Trade-off

Although it is rarely mentioned in the previous works, the speed of processing images also matters due to the extra large size image in WSIs. In order to investigate the trade-off between the efficiency and the accuracy quantitatively, we performed segmentation and classification on datasets

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|c|}
\hline
 & \textbf{DIST} & \textbf{Mask-RCNN} & \textbf{Micro-Net} & \textbf{HoVer-Net} & \textbf{PointNu-Net} \\
\hline
\textbf{Adrenal Gland} & 0.3442 & 0.5063 & 0.347 & 0.5546 & 0.4153 & 0.644 & 0.4882 & 0.6962 & 0.5115 & 0.7134 \\
\textbf{Bile Duct} & 0.3614 & 0.5384 & 0.3536 & 0.5567 & 0.4124 & 0.6232 & 0.4714 & 0.6696 & 0.4868 & 0.6814 \\
\textbf{Bladder} & 0.4463 & 0.5625 & 0.5065 & 0.6049 & 0.5357 & 0.6488 & 0.5792 & 0.7031 & 0.6065 & 0.7226 \\
\textbf{Breast} & 0.3790 & 0.5466 & 0.3882 & 0.5574 & 0.4407 & 0.6029 & 0.4902 & 0.6470 & 0.5147 & 0.6709 \\
\textbf{Cervix} & 0.3371 & 0.5309 & 0.3402 & 0.5483 & 0.3795 & 0.6101 & 0.4438 & 0.6652 & 0.5014 & 0.6899 \\
\textbf{Colon} & 0.2989 & 0.4508 & 0.3122 & 0.4605 & 0.3414 & 0.4972 & 0.4095 & 0.5575 & 0.4509 & 0.5945 \\
\textbf{Esophagus} & 0.3942 & 0.5295 & 0.4311 & 0.5691 & 0.4686 & 0.6011 & 0.5085 & 0.6427 & 0.5504 & 0.6766 \\
\textbf{Head & Neck} & 0.3177 & 0.4764 & 0.3946 & 0.5457 & 0.3668 & 0.5242 & 0.4530 & 0.6331 & 0.4838 & 0.6546 \\
\textbf{Kidney} & 0.3339 & 0.5727 & 0.3553 & 0.5092 & 0.4165 & 0.6321 & 0.4424 & 0.6836 & 0.5066 & 0.6912 \\
\textbf{Liver} & 0.3441 & 0.5818 & 0.4103 & 0.6085 & 0.4365 & 0.6666 & 0.4974 & 0.7248 & 0.5174 & 0.7314 \\
\textbf{Lung} & 0.2809 & 0.4798 & 0.3182 & 0.5134 & 0.337 & 0.5588 & 0.4004 & 0.6302 & 0.4048 & 0.6352 \\
\textbf{Ovarian} & 0.3789 & 0.5289 & 0.4337 & 0.5784 & 0.4387 & 0.6013 & 0.4863 & 0.6309 & 0.5484 & 0.6863 \\
\textbf{Pancreatic} & 0.3395 & 0.5343 & 0.3624 & 0.5460 & 0.4041 & 0.6074 & 0.4600 & 0.6491 & 0.4804 & 0.6791 \\
\textbf{Prostate} & 0.3810 & 0.5442 & 0.3959 & 0.5789 & 0.4341 & 0.6049 & 0.5101 & 0.6615 & 0.5127 & 0.6654 \\
\textbf{Skin} & 0.2627 & 0.5080 & 0.2665 & 0.5021 & 0.3223 & 0.5817 & 0.3429 & 0.6234 & 0.4011 & 0.6494 \\
\textbf{Stomach} & 0.3369 & 0.5553 & 0.3684 & 0.5976 & 0.3872 & 0.6293 & 0.4726 & 0.6886 & 0.4517 & 0.7010 \\
\textbf{Testis} & 0.3278 & 0.5548 & 0.3512 & 0.5420 & 0.4088 & 0.6300 & 0.4754 & 0.6890 & 0.5334 & 0.7058 \\
\textbf{Thyroid} & 0.2574 & 0.5596 & 0.3037 & 0.5712 & 0.3712 & 0.6555 & 0.4315 & 0.6983 & 0.4508 & 0.7076 \\
\textbf{Uterus} & 0.3487 & 0.5246 & 0.3683 & 0.5589 & 0.3965 & 0.5821 & 0.4393 & 0.6393 & 0.4846 & 0.6634 \\
\hline
\textbf{Average across tissues} & 0.3406 & 0.5346 & 0.3688 & 0.5528 & 0.4059 & 0.6053 & 0.4629 & 0.6596 & 0.4957 & 0.6808 \\
\textbf{STD across splits} & 0.0156 & 0.0097 & 0.0046 & 0.0076 & 0.0082 & 0.0050 & 0.0076 & 0.0036 & 0.0084 & 0.0030 \\
\hline
\end{tabular}
\caption{Comparative experiments of nuclei segmentation and classification on the PanNuke. We also provide the standard deviation (STD) across these splits in the final row.}
\end{table}

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
 & \textbf{Neo} & \textbf{Non-Neo} & \textbf{Epi} & \textbf{Inflam} & \textbf{Conn} & \textbf{Dead} \\
\hline
\textbf{DIST} & 0.439 & 0.290 & 0.343 & 0.275 & 0.000 & \\
\textbf{Mask-RCNN} & 0.472 & 0.403 & 0.290 & 0.300 & 0.069 & \\
\textbf{Micro-Net} & 0.504 & 0.442 & 0.333 & 0.334 & 0.051 & \\
\textbf{HoVer-Net} & 0.551 & 0.491 & 0.417 & 0.388 & 0.139 & \\
\textbf{PointNu-Net} & \textbf{0.578} & \textbf{0.577} & \textbf{0.433} & \textbf{0.409} & \textbf{0.154} & \\
\hline
\end{tabular}
\caption{Average PQ across three dataset splits on PanNuke for each nuclear category.}
\end{table}
Fig. 7. Examples of visual nuclei segmentation results on Kumar and CoNSeP. For each dataset, we displayed the 6 models from left to right. The different colours of the nuclear boundaries denote separate instances.

|                  | Kumar                        | CoNSeP                       |
|------------------|------------------------------|------------------------------|
|                  | DICE | AJI | DQ | SQ | bPQ | DICE | AJI | DQ | SQ | bPQ | mPQ |
| U-Net [35]       | 0.758 | 0.556 | 0.691 | 0.690 | 0.478 | 0.724 | 0.482 | 0.488 | 0.671 | 0.328 | -   |
| DIST [8]         | 0.789 | 0.559 | 0.690 | 0.732 | 0.443 | 0.804 | 0.502 | 0.544 | 0.728 | 0.398 | 0.372 |
| Mask-RCNN [12]   | 0.760 | 0.546 | 0.704 | 0.720 | 0.509 | 0.740 | 0.474 | 0.619 | 0.740 | 0.460 | 0.450 |
| Micro-Net [34]   | 0.797 | 0.560 | 0.692 | 0.747 | 0.519 | 0.794 | 0.527 | 0.600 | 0.745 | 0.449 | 0.430 |
| CIA-Net [10]     | 0.818 | 0.620 | 0.754 | 0.762 | 0.577 | -     | -     | -     | -     | -     | -     |
| HoVer-Net [9]    | 0.826 | 0.618 | 0.770 | 0.773 | 0.597 | 0.853 | 0.571 | 0.702 | 0.778 | 0.547 | 0.516 |
| Triple U-Net [11]| 0.837 | 0.621 | -     | -     | 0.601 | 0.843 | 0.579 | -     | -     | 0.562 | -     |
| PointNu-Net      | 0.814 | 0.606 | 0.784 | 0.768 | 0.603 | 0.822 | 0.560 | 0.714 | 0.762 | 0.545 | 0.536 |

TABLE III

Comparative experiments on small datasets Kumar and CoNSeP. BPO reflects the performance on nuclei segmentation, while mPQ highlights the performance on nuclei segmentation and classification.

CoNSeP over 10 runs and calculated the average inference time on a $1,000 \times 1,000$ image, as shown in Table IV. As a region proposal based method that will generate redundant and overlapped region-of-interests, the top-down method Mask-RCNN required larger space to store intermediate results and longer time to post-process, resulting in almost 52 seconds processing time per image. The bottom-up method HoVer-Net took about 5.3 seconds per image thanks to its parallel separation of instances. The post-processing of HoVer-Net contained complex operations using CPU, which is a bottleneck for faster inference time. PointNu-Net detected each instance using heatmap peaks and segmented instances dynamically without any post-processing, which takes advantages of parallel calculation on GPU. As a consequence, our default version of PointNu-Net spent 3.29 seconds inference time to achieve the highest mPQ on CoNSeP, approximately 1.6 times faster than HoVer-Net. On the other hand, the inference time declined about 10% without non-maximum suppression, just leading to a slight decrease of performance. Meanwhile, because almost 95% inference time were costed on GPU, using a smaller model can significantly speed up the model. Here we propose PointNu-Net-M and PointNu-Net-S, which used HRNet-w32 and HRNet-w18 as the backbone, respectively. Both lighter versions of PointNu-Net share the same training protocol and architectures as the default version, except using 4 instead of 7 stacked convolutional layers on the keypoint heatmap branch and kernel branch. Eventually, we can halve the inference time compared with the default PointNu-Net, while achieving acceptable performance.
E. Ablation Study

We conducted ablation studies to gain full understanding of the main components in PointNu-Net. In order to get convincing results, we evaluate all the experiments on the large dataset PanNuke.

1) Keypoint Heatmap Branch: We investigated the importance of the keypoint heatmap strategy and focal loss. First, we replaced the prediction of the keypoint heatmap with that of center points to detect and classify each instance, and used binary cross entropy (BCE) loss as the loss function. We refereed to this combination as the vanilla strategy because cross entropy loss is the most commonly used loss function for classification. Then, on the basis of vanilla, we introduced Gaussian kernel strategy on each instance center to produce the keypoint heatmap, where the Gaussian kernel size was directly proportional to the object size. Finally, we introduced focal loss as the loss function, which is our default setting of PointNu-Net.

As shown in Table V, both the keypoint heatmap and the focal loss improved the performance of the vanilla setting. Specifically, the keypoint heatmap strategy boosted mPQ by 0.044 and bPQ by 0.062. The intuition behind the keypoint heatmap is that it made the ground truth smooth nearby each object center point and therefore reduced the difficulty of center point prediction. The focal loss further contributed an improvement of 0.049 and 0.050 in terms of mPQ and bPQ. This is benefited from that focal loss enforces the model to mine the hard examples, as it can reduce the weight of simple negative samples in training. These results indicated the effectiveness of the keypoint-aware prediction.

2) Backbone And Neck Selection: ResNet backbone and Feature Pyramid Networks (FPN) [36] are commonly considered as the default multi scale feature extractor in many computer vision tasks due to their simple but efficient architecture. However, as discussed before, a network in series may not be a good selection for local feature extraction. In order to validate this, we performed several experiments using different backbones and neck combination. As shown in Table VI, the combination of ResNet101 and FPN achieved 0.486 and 0.673 in terms of mPQ and bPQ, which was already higher than the previous work HoVer-Net. However, when we used ResNeXt-101-DCN which is considered more powerful in natural image processing, the performance became even worse. This might be the reason that better ability in catching global features may not benefit our task. Instead, HRNet-w64 can better utilize local features for classifying and detecting high density instances, as it keeps the high-resolution features along feature extraction. However, since HRNet-w64 has merged multi-scale features in parallel, the effect of FPN was very limited. Hence, we introduced Joint Pyramid Fusion Module (JPFM) to make full use of the features from HRNet-w64. For feature fusion, we have two options: to construct a shared JPFM to merge the features for all branches or to learn the feature in each branch separately, which are termed as shared and unshared JPFM respectively. Experimental results demonstrated that a shared JPFM outperformed the default neck FPN, while the unshared JPFM further boosted the network’s ability since different scale information may not have the same importance for the three separated branches.

To better understanding the unshared JPFM, we visualized the feature activation map in the three branches, as shown in Figure 8. As observed, keypoint heatmap and feature branches had more active feature map in \( d = 1 \), which means that the low-level high-resolution features are important to detection and classification. In contrast, the feature map in kernel branch activated strongly in \( d = 8 \), reflecting the crucial role of high-level low-resolution features in kernel generation.

| Table IV | Speed-Accuracy Trade-off on the dataset CoNSep. \(^1\) denotes model inference without matrix NMS operation. |
|----------|----------------------------------------------------------|
| Backbone | Neck | mPQ | bPQ |
| ResNet-101 | FPN | 0.486 | 0.673 |
| ResNet-101-DCN | FPN | 0.480 | 0.667 |
| HRNet-w64 | FPN | 0.490 | 0.675 |
| HRNet-w64 | Shared JPFM | 0.492 | 0.677 |
| HRNet-w64 | Unshared JPFM | 0.496 | 0.681 |

| Table V | Effectiveness of keypoint-aware model. |
|----------|--------------------------------------|
| mPQ | bPQ |
| Center point + BCE loss (vanilla) | 0.403 | 0.569 |
| Keypoint heatmap + BCE loss | 0.447 | 0.631 |
| Keypoint heatmap + Focal loss (ours) | 0.496 | 0.681 |

Fig. 8. Visualization of feature activation maps of all convolution layers in the JPFM in keypoint heatmap branch, kernel branch, and feature branch. \( d \) denotes the dilation of convolution in the JPFM.
V. CONCLUSION

In this paper, we have presented a keypoint-aware PointNu-Net for nuclei segmentation and classification within multi-tissue histology images, which detects and classifies nuclei by keypoint heatmap regression and segments nuclei simultaneously via dynamic convolution. We have proposed a more reasonable feature aggregation module for HRNet, called Joint by keypoint heatmap regression and segments nuclei simultaneously within multi-photonic images as paired keypoints, in "European Conference on Computer Vision, 2018.

K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, “Centernet: Keypoint triplets for object detection,” in "International Conference on Computer Vision, 2019, pp. 6569–6578.

Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, “Reppoints: Point set representation for object detection,” in "International Conference on Computer Vision, 2019, pp. 9657–9666.

Z. Dong, G. Li, Y. Liao, F. Wang, P. Ren, and C. Qian, “Centripetalnet: Pursuing high-quality keypoint pairs for object detection,” in "IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10519–10528.

X. Zhou, D. Wang, and P. Krähenbühl, “Objects as points,” arXiv preprint arXiv:1904.07850, 2019.

S. Lan, Z. Ren, Y. Wu, L. S. Davis, and G. Hua, “Saccadnet: A fast and accurate object detector,” in "IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10.397–10.406.

C. Zhe, T. Simon, S. E. Wei, and Y. Sheikh, “Real-time multi-person 2D pose estimation using part affinity fields,” in "IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

A. Newell, K. Yang, and D. Jia, “Stacked hourglass networks for human pose estimation,” in "European Conference on Computer Vision, 2016.

T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal Loss for Dense Object Detection,” in "2017 IEEE International Conference on Computer Vision (ICCV), oct 2017, pp. 2999–3007.

Z. Tian, C. Shen, and H. Chen, “Conditional convolutions for instance segmentation,” arXiv preprint arXiv:2003.05664, 2020.

X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, “Soltv2: Dynamic and fast instance segmentation,” in "Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020.

X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, “SOLO: Segmenting objects by locations,” in "European Conference on Computer Vision, 2020.

F. Milletari, N. Navab, and S. Ahmadi, “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation,” in "2016 Fourth International Conference on 3D Vision (3DV), 2016, pp. 565–571.

K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in "European conference on computer vision, Springer, 2016, pp. 630–645.

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2015.

J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang et al., “Deep high-resolution representation learning for visual recognition,” IEEE transactions on pattern analysis and machine intelligence, 2020.

J. Gamper, N. A. Koobhbanani, K. Benet, A. Khurram, and N. Rajpoot, “PanNuke: an open pan-cancer histology database for nuclei instance segmentation and classification,” in "European Congress on Digital Pathology, Springer, 2019, pp. 11–19.

M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, “A method for normalizing histology slides for quantitative analysis,” in "2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2009, pp. 1107–1110.

I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” arXiv preprint arXiv:1711.05101, 2017.

S. E. A. Raza, L. Cheung, M. Shaban, S. Graham, D. Epstein, S. Penglaris, M. Khan, and N. M. Rajpoot, “Micro-net: A unified model for segmentation of various objects in microscopy images,” Medical image analysis, vol. 52, pp. 160–173, 2019.

O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in "International Conference on Medical image computing and computer-assisted intervention, 2015, pp. 234–241.

T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in "IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2117–2125.