DPA-P2PNet: Deformable Proposal-Aware P2PNet for Accurate Point-Based Cell Detection

Zhongyi Shui[1,2*], Sunyi Zheng[2*], Chenglu Zhu[2], Shichuan Zhang[1,2], Xiaoxuan Yu[1,2], Honglin Li[1,2], Jingxiong Li[1,2], Pingyi Chen[1,2], Lin Yang[2†]

[1]College of Computer Science and Technology, Zhejiang University
[2]School of Engineering, Westlake University
{shuizhongyi, yanglin}@westlake.edu.cn

Abstract

Point-based cell detection (PCD), which pursues high-performance cell sensing under low-cost data annotation, has garnered increased attention in computational pathology community. Unlike mainstream PCD methods that rely on intermediate density map representations, the Point-to-Point network (P2PNet) has recently emerged as an end-to-end solution for PCD, demonstrating impressive cell detection accuracy and efficiency. Nevertheless, P2PNet is limited to decoding from a single-level feature map due to the scale-agnostic property of point proposals, which is insufficient to leverage multi-scale information. Moreover, the spatial distribution of pre-set point proposals is biased from that of cells, leading to inaccurate cell localization. To lift these limitations, we present DPA-P2PNet in this work. The proposed method directly extracts multi-scale features for decoding according to the coordinates of point proposals on hierarchical feature maps. On this basis, we further devise deformable point proposals to mitigate the positional bias between proposals and potential cells to promote cell localization. Inspired by practical pathological diagnosis that usually combines high-level tissue structure and low-level cell morphology for accurate cell classification, we propose a multi-field-of-view (mFoV) variant of DPA-P2PNet to accommodate additional large FoV images with tissue information as model input. Finally, we execute the first self-supervised pre-training on immunohistochemistry histopathology image data and evaluate the suitability of four representative self-supervised methods on the PCD task. Experimental results on three benchmarks and a large-scale and real-world interval dataset demonstrate the superiority of our proposed models over the state-of-the-art counterparts. Codes and pre-trained weights are available at https://github.com/windygoo/DPA-P2PNet.

Introduction

Identifying various types of cells such as tumor cells, lymphocytes, and fibroblasts in histopathology whole slide images (WSIs) is crucial for numerous downstream tasks including tumor microenvironment analysis (Jiao et al. 2021), cancer diagnosis (Cheng et al. 2022) and prognosis (Howard, Kanetsky, and Egan 2019). Predominant cell detection methodologies embrace a instance segmentation paradigm. Although these approaches showcase impressive capability in capturing intricate details of cell morphology, their training requires a formidable investment of valuable resources due to the laborious nature of cell mask annotation. As a matter of fact, the expensive annotation has long plagued the advancement and application of mask-based cell detection models.

To reduce the annotation cost while maintaining sufficient clinical support, point-based cell detection (PCD) has emerged as a promising and rapidly evolving frontier in computational pathology (Zhou et al. 2018; Huang et al. 2020; Abousamra et al. 2021; Cai et al. 2021; Zhang et al. 2022; Ryu et al. 2023). The goal of PCD is to predict a 2D point set that represents the coordinates and classes of cells present in an input image. To accomplish this, prevalent PCD methods connect to off-the-shelf segmentation models via carefully crafted pseudo mask labels derived from point annotations. Subsequently, a series of post-processing steps comprising thresholding, local maxima detection and non-maximum suppression are applied to the predicted density maps to locate cells. However, the heuristic post-processing not only demands tedious hyper-parameter tuning but also results in low throughput. To address these issues, a recent study (Shui et al. 2022) introduced Point-to-Point Network (P2PNet) (Song et al. 2021) to establish an end-to-end PCD system. Specifically, P2PNet adopts a detection paradigm, where the cell coordinates and categories can be directly obtained by refining and classifying pre-defined point proposals on an input image. Moreover, P2PNet employs a one-to-one matching scheme to suppress duplicate predictions, eliminating the need for error-prone and time-consuming post-processing. Because of these improvements, P2PNet can achieve superior accuracy and efficiency over the mainstream density map-based PCD methods.

Despite its successful application, we contend that the performance and flexibility of P2PNet can be limited from the following two aspects. (i) Unlike anchor boxes that naturally account for object scales, point proposals are scale-agnostic. As a result, P2PNet can only decode from a single-level feature map, which is inadequate to represent multi-scale information. Considering the substantial variability in cell size and morphology, as well as the heterogeneity in in-

*These authors contributed equally.
†Corresponding author.

tensity distribution (Ryu et al. 2023), decoding from hierarchical feature maps becomes imperative. (ii) The spatial distribution of artificially placed point proposals is inherently sub-optimal, with many of them situated far from cell centroids. This constitutes a significant challenge for achieving high-quality localization. On the other hand, P2PNet lacks the ability to perceive the positions of point proposals, which restricts it to use a fixed set of point proposals to ensure model convergence. To lift these limitations, this study proposes ***d***eformable ***p***roposal-***a***ware P2PNet, dubbed as DPA-P2PNet. Overall, we incorporate two improvements into the vanilla P2PNet. First, we straightforwardly extract multi-scale decoding features for each point proposal according to its coordinates on feature pyramid. This modification not only enhances the model's capability but also makes it to be proposal-aware. Based on this, we further design deformable point proposals to reduce the distribution bias between proposals and potential cells to improve the localization quality and categorical discriminability of extracted features.

In clinical practice, pathologists generally perform accurate cell classification in two steps. They first zoom out to comprehend broad tissue structures and then zoom in to classify cells based on their morphology and the surrounding context. However, most computer-assisted PCD methods operate with a single image as input, deviating from the authentic diagnostic procedure. To mitigate this inconsistency, (Bai, Xu, and Xing 2020; Bai et al. 2022; Ryu et al. 2023) adapt cell segmentation models to incorporate input of multi-field-of-view (mFoV) images, resulting in heightened accuracy of density map-based PCD methods. Nonetheless, their improvements cannot be readily applied to end-to-end PCD models, owing to the intrinsic disparities in model architecture and operating mechanism. This study presents mFoV DPA-P2PNet as the first end-to-end PCD model capable of utilizing mFoV images for better cell classification.

Self-supervised learning (SSL) aims to learn a generic representation applicable to various downstream tasks. Recent researches (Wang et al. 2021; Li et al. 2023; Kang et al. 2023) have demonstrated that domain-aligned pre-training outperforms traditional transfer learning from ImageNet in several medical imaging tasks. Yet, the applicability of self-supervised domain-aligned pre-training to the dense prediction task of PCD has never been explored. To remedy this deficiency, we carry out an inaugural investigation into the effect of four representative SSL methods for the PCD task, including MoCo v2 (Chen et al. 2020b), SwAV (Caron et al. 2020), DINO (Caron et al. 2021), and MAE (He et al. 2022). Aside from the difference in the downstream task compared to prior studies, we are also the first to perform SSL on a large-scale and highly valuable immunohistochemistry (IHC) WSI dataset that involves three biomarkers, namely Ki-67, PD-L1, and HER-2.

Related Work

Crowd Counting and Localization

Crowd counting aims to estimate the number of people in an image. The mainstream idea is to regress a pseudo density map generated from point annotations (Liang et al. 2022). The final crowd count is calculated by 2D integration over the estimated density map.

To tackle the issues of local inconsistency (Lian et al. 2019) and weak interpretability of density map-based crowd counting methods, recent studies (Wan, Liu, and Chan 2021; Song et al. 2021; Liang, Xu, and Bai 2022; Liang et al. 2022; Lin and Chan 2023) have redirected their focus towards the localization-based crowd counting problem, where the crowd count is represented by the number of localized human heads. Existing crowd localization methods can be categorized into two types. The first line of methods conducts additional post-processing on the estimated density maps to localize individual human heads. Another type of methods comprising P2PNet (Song et al. 2021) and CLTR (Liang, Xu, and Bai 2022) achieve end-to-end crowd localization by directly regressing the point coordinates. Currently, P2PNet represents the state-of-the-art in the field of crowd localization (Lin and Chan 2023).

While crowd localization and PCD share similar high level spirit, there are two concretized differences between them. Firstly, in terms of localization, human heads have relatively regular and consistent shapes, whereas cells typically exhibit a wide range of shapes and sizes. This large variation poses a greater challenge for accurate cell localization. Secondly, when it comes to classification, crowd localization only concerns binary classes, distinguishing between foreground (heads) and background, while PCD generally involves multiple categories. Furthermore, classifying cells is significantly more demanding as it requires the integration of both coarse-grained tissue structure and fine-grained cell morphology. These two factors stress the necessity of leveraging multi-scale information for accurate cell detection.

Point-based Cell Detection

PCD aims to localize and classify cells in a pathology image, with each cell represented by a class-aware point. Mainstream PCD methods (Abousamra et al. 2021; Cai et al. 2021; Zhang et al. 2022; Ryu et al. 2023) operate similarly with density map-based crowd localization approaches but regress multiple density maps, each corresponding to a distinct cell type. Recently, (Shui et al. 2022) introduces the advanced P2PNet to perform PCD in an end-to-end manner. However, the original P2PNet model can only decode from a single-level feature map, which is insufficient to squeeze the most out of the multi-scale information. To alleviate this deficiency, (Shui et al. 2022) downsamples the shallow feature maps and executes feature fusion by element-wise summation, resulting in an enhanced feature map to decode. Nevertheless, we argue that this approach would lead to a loss of fine-grained information to some extent and the improved P2PNet still suffers from the biased distribution of pre-defined point proposals.

Leveraging Large Field of Views

Several studies (Tokunaga et al. 2019; Ho et al. 2021; Schmitz et al. 2021; Van Rijthoven et al. 2021) extract a large FoV region as an additional input to improve the segmentation performance on smaller FoV regions. In the area of PCD, (Bai, Xu, and Xing 2020) and (Bai et al. 2022) have

devoted pioneering efforts in this direction. They propose a feature aggregation module that combines visual representations extracted from two images with different FoVs to enhance cell detection, where the tissue structure information is learned implicitly. A recent study (Ryu et al. 2023) incorporates the contextual knowledge explicitly via multi-task objectives at different FoVs. Specifically, they build two models for tissue and cell segmentation at large and small FoVs, respectively. The predicted tissue probability map is blended into the cell detection branch to promote cell classification. In this study, we adopt the same experimental setup as in (Bai, Xu, and Xing 2020; Bai et al. 2022) since the costly tissue mask annotation is unavailable in general. However, these two pioneering methods are both based on density map regression. So their improvements cannot be seamlessly applied to the advanced end-to-end PCD models due to the inherent discrepancy in model architecture and operating mechanism.

Self-supervised Learning in Medical Imaging

SSL has proven to be an effective method to learn a good representation from vast unlabeled images by solving a pretext task. Pre-trained models from SSL are widely used in fine-tuning downstream tasks faster or for better accuracy (Chen et al. 2020a). For applications in medical imaging, transfer from ImageNet has become the de-facto approach. Yet, (Matsoukas et al. 2022) observes that domain-specific SSL methods can further improve the performance of models fine-tuned on downstream medical image-related tasks. This has been confirmed on numerous medical image analysis tasks such as pathology image classification (Sowrirajan et al. 2021; Wang et al. 2021; Kang et al. 2023; Chen et al. 2022a; Li et al. 2023) and retrieval (Gildenblat and Klaiman 2019), survival outcome prediction (Chen et al. 2022a), nuclei instance segmentation (Kang et al. 2023) and MRI brain tumor segmentation (Zhou et al. 2022). However, the applicability of self-supervised domain-aligned pre-training for the PCD task remains unrevealed.

Approach

In this section, we first briefly review the vanilla P2PNet. Then we elaborate the proposed DPA-P2PNet, including the multi-scale decoding (MSD) strategy and the generation of deformable point proposals (DPP). After that, we present mFoV DPA-P2PNet that supports mFoV images as input for better cell classification. Lastly, we detail the IHC dataset collected for SSL pre-training.

Revisiting P2PNet

Similar with modern object detectors (Ren et al. 2015), P2PNet comprises three parts: backbone, neck and heads. Taken an image $I \in \mathbb{R}^{H\times W\times 3}$ as input, the backbone and neck produces hierarchical visual representations $\{P_i\}_{i=2}^{L}$. Let s_i denote the downsampling ratio of P_i, then s_i equals 2^i.

Since point proposals are scale-agnostic, P2PNet only selects one feature map (e.g., P_i) for decoding. Each location of P_i corresponds to a $s_i \times s_i$ patch of the input image. To densely detect the cells, $m \times n$ point proposals are placed at each patch in a grid distribution, where m and n separately represent the number of rows and columns of point proposals. Conditional on P_i, two task-specific convolutional heads are employed to generate regression offset map and classification logit map of channels $2mn$ and $(C+1)mn$ while retaining the spatial resolution. C is the number of cell types and the extra class is background. The 2D offsets and logits at each location are assigned to the point proposals within the corresponding patch.

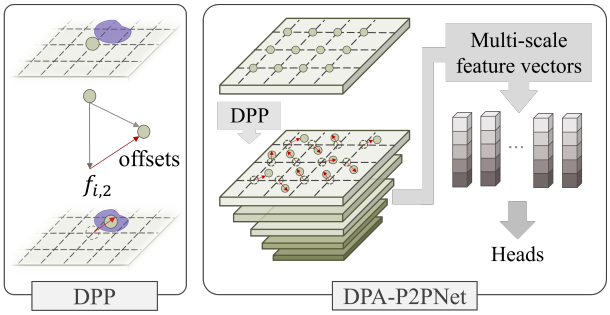

Figure 1: Framework of DPA-P2PNet.

DPA-P2PNet

Fig. 1 presents the overall framework of our proposed DPA-P2PNet. In the following sections, we use the set $\mathcal{S} = \{p_i\}_{i=1}^{M}$ to denote the pre-defined point proposals.

Decoding from feature pyramid To effectively exploit the information at different granularities, we construct ROI features from hierarchical feature maps based on the coordinates of each point proposal. Specifically, the ROI feature vectors $\{f_{i,j}\}_{j=2}^{L}$ for proposal p_i are extracted from the feature pyramid via the bilinear interpolation method:

$$f_{i,j} = \sum_{q} G(p_i, q) \cdot P_j(q), \tag{1}$$

where j denotes the feature level, q enumerates all integral spatial locations around p_i in the feature map P_j and $G(\cdot,\cdot)$ is the bilinear interpolation kernel. Then, we concatenate $\{f_{i,j}\}_{j=2}^{L}$ and fed it into two dedicated MLP heads for decoding offsets and logits with respect to p_i.

The decoding strategy described above not only facilitates the exploitation of multi-scale information but also endows our model with a proposal-aware ability. Intuitively, in the original P2PNet, a point proposal passively receives the decoded content, whereas in our model, proposals actively query the distances to potential cells. As a result, P2PNet is confined to a fixed set of point proposals to ensure model convergence. However, a considerable portion of these manually placed proposals deviate far from cell centroids, which could easily lead to inaccurate localization. This distribution bias can also cause the extracted features to be less discriminative for cell classification in our model. The proposal-aware nature opens up possibilities for dynamically refining the point proposals without concerns of model dispersion.

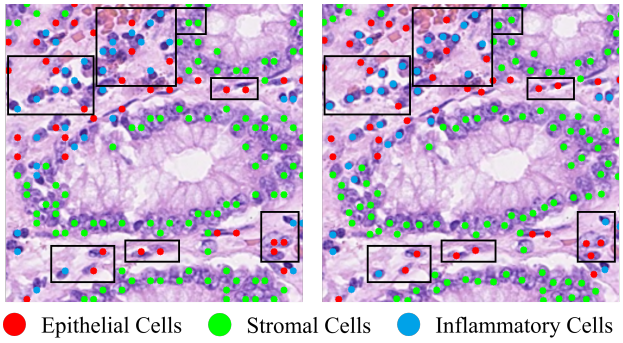

Figure 2: Illustration of our proposed deformable point proposals. The positions of foreground proposals before (left) and after (right) the deformation are depicted. We highlight some ROI regions with boxes for comparison. For a clearer view, we encourage readers to zoom in.

Deformable point proposals Improving the quality of object proposals has been a long-standing research topic in the field of object detection (Zhu et al. 2020). In this paper, we propose a structurally simple yet highly potent solution to mitigate the positional bias between pre-defined point proposals and potential cells.

As illustrated in Fig. 1, we deform the pre-set point proposals to shift them towards neighboring cells. To achieve this, we use a MLP layer to generate the deformation offsets from $f_{i,2}$, informed by that the high-resolution P_2 contains the finest-grained features essential for small object localization (Lin et al. 2017; Liu et al. 2021). It is noteworthy that the offsets are learned implicitly without direct supervision. Based on the deformed point proposals $\mathcal{S}'$, we proceed with the decoding procedure, as described in the preceding section, to perform classification and finer localization.

Fig. 2 presents an intuitive demonstration of the deformation process. Clearly, the pre-defined point proposals are adaptively transported onto nearby cells by this transformation. On the basis of $\mathcal{S}'$, we can extract more discriminative features for classification. Moreover, through additional refinement during the subsequent decoding stage, the localization quality can be further improved.

mFoV DPA-P2PNet

Concentric images $\{I^k\}_{k=1}^{K}$ with the same resolution yet captured at different objective magnification can be interpreted as having multiple FoVs. In this paper, we use I^K to represent the image with the highest magnification yet smallest FoV, while $\{I^k\}_{k=1}^{K-1}$ to denote the set of images with larger FoVs. Following the setup of (Bai, Xu, and Xing 2020; Bai et al. 2022), the magnification of I^k is twice that of I^{k-1}.

Fig. 3 presents the framework of mFoV DPA-P2PNet. Specifically, we first use a separate backbone and neck to construct the feature pyramid $\{P_i^k\}_{i=2}^{L}$ for image I^k. After that, we aggregate the contextual information extracted from

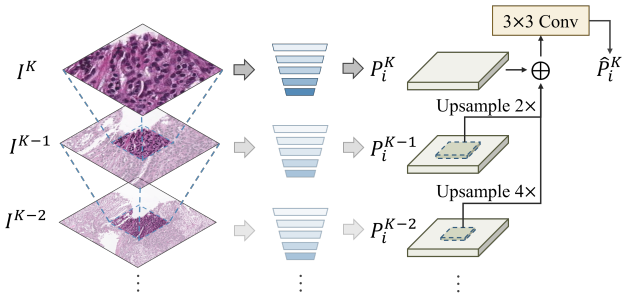

Figure 3: Schematic of mFoV DPA-P2PNet.

IHC biomarker	# WSIs	# Patches	Organs
Ki-67	403	233,308	breast, stomach
PD-L1	1,208	607,302	lung
HER2	1,101	597,563	breast

Table 1: IHC data for pre-training.

large FoV images into the feature maps $\{P_i^K\}_{i=2}^{L}$ obtained from I^K, which can be expressed as:

$$\hat{P}_i^K = \mathrm{Conv}\left(P_i^K + \sum_{k=1}^{K-1} \mathrm{Upsample}\left(\mathrm{Crop}\left(P_i^k\right)\right)\right) \quad (2)$$

where $\mathrm{Crop}(\cdot)$ is the center square crop operation. For P_i^k, the cropping limit normalized by dividing the height or width is $\left[\frac{2^{K-k}-1}{2^{K-k+1}}, \frac{2^{K-k}+1}{2^{K-k+1}}\right)$. Then we upsample the cropped region with a factor of 2^{K-k} to match the resolution of P_i^K and fuse the mFoV features by element-wise summation. Unless specified otherwise, the upsampling process is completed by a shared ConvTranspose2d layer in this work. Thereafter, in order to eliminate the aliasing effect of upsampling (Lin et al. 2017), we employ a 3×3 convolution on the merged map to generate the final feature map $\hat{P}_i^K$. Finally, we utilize the enhanced feature pyramid $\{\hat{P}_i^K\}_{i=2}^{L}$ for decoding.

Data Collection for SSL

Tab. 1 provides an overview of the immunohistochemistry (IHC) dataset used for pre-training. We first collect 2,712 WSIs that cover three types of IHC biomarkers: Ki-67, programmed death-ligand 1 (PD-L1), and human epidermal growth factor receptor 2 (HER2). To increase the diversity and informativeness of the pre-training dataset, we extract at most 500 patches of resolution 512 × 512 pixels and objective magnification 40× (0.25 μm/px) from each WSI. Moreover, a pre-trained cell detector is deployed to guarantee that each preserved patch contains at least 30 cells. As a result, we collect a total of 1.4M patches, slightly exceeding the scale of the ImageNet-1K training set (Deng et al. 2009).

Broader impact Beyond assessing the effect of various SSL algorithms on the PCD task, we are the first to con-

Datasets	Objective magnification	No. of patches			No. of cells			No. of categories	Image resolution	Organs
		train	val	test	train	val	test			
CoNSeP	20×	22	5	14	13,040	2,515	8,777	3	500×500	colon
BCData	40×	803	133	402	93,838	21,804	65,432	2	640×640	breast
PD-L1	40×	1215	405	405	358,832	116,148	116,418	10	1024×1024	lung
OCELOT	50×	252	78	70	41,645	13,574	10,618	2	1024×1024	kidney, etc.

Table 2: Profiles of four histopathology datasets. A 20× objective magnification corresponds to approximately 0.5 μm/px.

duct self-supervised pre-training with large-scale IHC image data. In contrast to HE data that is widely adopted in previous relevant works, IHC staining allows the detection and localization of specific proteins or antigens within tissue sections, which is particularly valuable for identifying specific cell types, biomarkers, or pathological changes associated with diseases. In clinical practice, pathologists usually combine the complementary merits of HE and IHC stained WSIs to obtain a more comprehensive understanding of tissue samples. By making the pre-trained weights publicly available, we hope to advance the development of the computational pathology community.

Experiment

Experimental Setup

Dataset In this study, we evaluate the advantages of DPA-P2PNet over the state-of-the-art counterparts on three histopathology datasets with varied staining types, including the HE stained CoNSeP (Graham et al. 2019), IHC Ki-67 stained BCData (Huang et al. 2020) datasets, and an internal IHC PD-L1 dataset. To validate the efficacy of our proposed mFoV DPA-P2PNet, we conduct comprehensive experiments on the OCELOT (Ryu et al. 2023) dataset, which offers the mapping from annotated patches to their source WSIs in TCGA (Hutter and Zenklusen 2018) so that we can crop patches at arbitrary FoVs. We divide the PD-L1 and OCELOT dataset into *training*, *validation*, and *test* subsets at a ratio of 6:2:2. To avoid information leaking among the subsets, we randomly split the dataset per WSI, ensuring that different patches from the same WSI are not included in multiple subsets. Tab. 2 provides the statistics of these datasets, and detailed cell categories are available in the supplementary material.

Implementation Details The interval of pre-defined point proposals is set to 8 pixels on the CoNSeP dataset while 16 pixels on the other three datasets. By default, we use ResNet-50 (He et al. 2016) and FPN (Lin et al. 2017) as the trunk and neck networks, respectively. All MLPs are structured as FC-ReLu-Dropout-FC. With the same label assignment scheme and loss functions as P2PNet (Song et al. 2021), we adopt the AdamW optimizer with weight decay 1e-4 to optimize our proposed models. During the training stage, data augmentations including random scaling, shifting and flipping are applied on the fly. For the pre-training, we utilize the configurations recommended in (Kang et al. 2023) to execute various SSL algorithms on our collected IHC dataset. All models are trained on NVIDIA A100 GPUs.

Evaluation Metrics We adopt macro-average F1-score and average precision (AP) as metrics to measure the cell detection performance of all models. If a detected cell is within a valid distance from an annotated cell and the cell class matches, it is counted as a true positive (TP), otherwise a false positive (FP). Following the official evaluation protocol, we set the matching distance as 6 and 10 pixels for the CoNSeP and BCData datasets. As for the OCELOT and PD-L1 datasets, we apply thresholds of 9 and 12 pixels, respectively.

Experimental Results

In the following parts, we first compare the capability of DPA-P2PNet with the state-of-the-art PCD and crowd localization competitors, which encompass density map (DM)-based approaches containing U-Net (Ronneberger, Fischer, and Brox 2015), DeepLabV3+ (Chen et al. 2018; Ryu et al. 2023), U-CSRNet (Huang et al. 2020), MCSpatNet (Abousamra et al. 2021), FIDT (Liang et al. 2022) and OT-M (Lin and Chan 2023), as well as end-to-end methods including P2PNet (Song et al. 2021), CLTR (Liang et al. 2022) and E2E (Shui et al. 2022) on the CoNSeP, BCData and PD-L1 datasets. Subsequently, we demonstrate the superiority of mFoV DPA-P2PNet over the precursor studies (Bai, Xu, and Xing 2020; Bai et al. 2022) on the OCELOT dataset. Thereafter, we conduct several ablation experiments to show the effectiveness of MSD and DPP. Finally, we test the validity of diverse SSL methods on the PCD task.

Comparison with SOTA Methods Table 3 shows the quantitative comparison results of our approach with the counterparts. Briefly, our proposed DPA-P2PNet achieves the highest F1 and AP scores on all datasets. The DM-based methods generally exhibit inferior performance because they rely on post-processing to localize cell centroids on the predicted density maps. However, finding a set of post-processing parameters that work well in all scenes is impossible. For example, setting the confidence threshold too large inevitably filters out cells with weak intensity, while using a small confidence threshold makes it difficult to separate overlapping cells. The inferior performance of P2PNet compared to our method can be attributed to its insufficient utilization of multi-scale information and the sub-optimal distribution of pre-defined point proposals. E2E unifies the resolutions of multi-level feature maps and performs feature fusion by element-wise summation, which mitigates the former drawback of P2PNet but somewhat leads to a loss of information. In comparison, DPA-P2PNet can fully utilize multi-scale features via instant decoding from the uncom-

Datasets	Metrics	*DM-based Methods*						*End-to-end Methods*			
		U-Net	DeepLabV3+	U-CSRNet	MCSpatNet	FIDT	OT-M	P2PNet	CLTR	E2E	Ours
CoNSeP	F1	61.8	65.7	42.5	68.2*	<u>70.3</u>	36.7	70.0	40.8	70.2	**71.1**
	AP	-	53.7	27.7	52.2*	57.7	-	<u>60.5</u>	21.8	59.9	**62.9**
BCData	F1	85.7	85.8	85.7*	85.3	85.9	66.5	<u>86.3</u>	84.8	<u>86.3</u>	**86.7**
	AP	80.1	82.1	-	79.2	81.3	-	<u>83.8</u>	81.0	83.3	**84.4**
PD-L1	F1	46.0	53.7	25.7	43.6	51.9	27.1	54.4	54.3	<u>54.8</u>	**55.9**
	AP	33.4	41.3	21.1	30.8	36.4	-	42.1	<u>43.0</u>	42.5	**43.7**
PD-L1	Params(M)	31.0	40.3	10.1	26.1	21.5	66.6	27.3	41.0	28.9	32.3
	MACs(G)	876	184	1371	277	579	432	100	105	107	151
	FPS	14	28	3	12	10	2	41	22	42	39

Table 3: Quantitative comparison on three datasets. * indicates the previously publicly reported best results on the dataset. The best and second-best performance are highlighted in bold and underlined, respectively.

	Methods	F1	AP
DeepLabV3+	Baseline	58.8	42.7
	MFoVCE-Net	59.3	43.5
	MFoVCE-Net+	59.8	44.2
DPA-P2PNet	Baseline	59.3	44.4
	Ours (BI)	<u>61.9</u>	<u>48.9</u>
	Ours (TC)	**62.6**	**49.1**

Table 4: Performance comparison of various PCD methods with mFoV inputs on the OCELOT dataset. K is set as 2. The baselines represent using only the small FoV patch. BI and TC indicate bilinear interpolation and transposed convolution, respectively.

No. of FoVs	F1	AP	FPS
1	59.3	44.4	25
2	62.6	49.1	18
3	64.2	51.4	11
4	64.6	51.6	8

Table 5: Performance of mFoV DPA-P2PNet on the OCELOT dataset as K increases from 1 to 4.

MSD	DPP	F1	AP
		70.0	60.5
✓		70.7	61.7
✓	✓	71.1	62.9

Table 6: Effect of MSD and DPP on the CoNSeP dataset. P2PNet serves as the baseline.

pressed feature pyramid. Although the transformer-based CLTR model employs a more elegant query-based paradigm compared to our method, it requires a larger amount of labeled data to unleash its potential because vision transformer (ViT) based models have weaker inductive bias than CNNs in modeling visual structures and thus require much more labels to learn such bias implicitly (Xu et al. 2021). However, obtaining cell annotations on a large scale demands specialized expertise, and incurs a huge cost. From this perspective, our model is more label-efficient. Additionally, the capacity of CLTR, quantified by the number of queries, cannot scale with resolutions of input images, which greatly curtails its practical applications.

We also analyze the model size, computational cost and inference efficiency of different methods on the PD-L1 dataset in Tab. 3. The DM-based methods demonstrate high computational complexity because they need to regress the high-resolution cell density maps. In contrast, the end-to-end models operate with hidden features of lower resolution and directly output cell coordinates and categories without the need for time-consuming post-processing. Consequently, they require fewer computational resources and generally exhibit faster inference speeds compared to the DM-based methods. The CLTR model, however, stands as an exception due to the considerable time complexity of the multi-head self-attention layers. Despite a slight sacrifice in speed, our proposed DPA-P2PNet outperforms the original P2PNet notably by 1.5% on F1 and 1.6% on AP.

Effect of mFoV DPA-P2PNet We compare the performance of our proposed mFoV DPA-P2PNet with MFoVCE-Net (Bai, Xu, and Xing 2020) and its upgraded version MFoVCE-Net+ (Bai et al. 2022) on the OCELOT dataset. The comparison results are shown in Tab. 4. It can be seen that mFoV DPA-P2PNet that uses transposed convolution for upsampling achieves the highest performance, surpassing the currently SOTA method MFoVCE-Net+ by a remarkable margin of 2.8% on F1 and 4.9% on AP.

In Tab. 5, we further investigate the scalability of mFoV DPA-P2PNet by utilizing more patches with larger FoVs as model input, which has not been previously investigated. Overall, the model performance keeps growing as the number of FoVs increases. As a trade-off, the inference speed gradually slows down. Specifically, when K equals to 4, our approach achieves a substantial performance gain of 5.3% on F1 and 7.2% on AP compared to the baseline model that only sees the patch with the smallest FoV. However, the inference efficiency reduces by 68% as more images need to

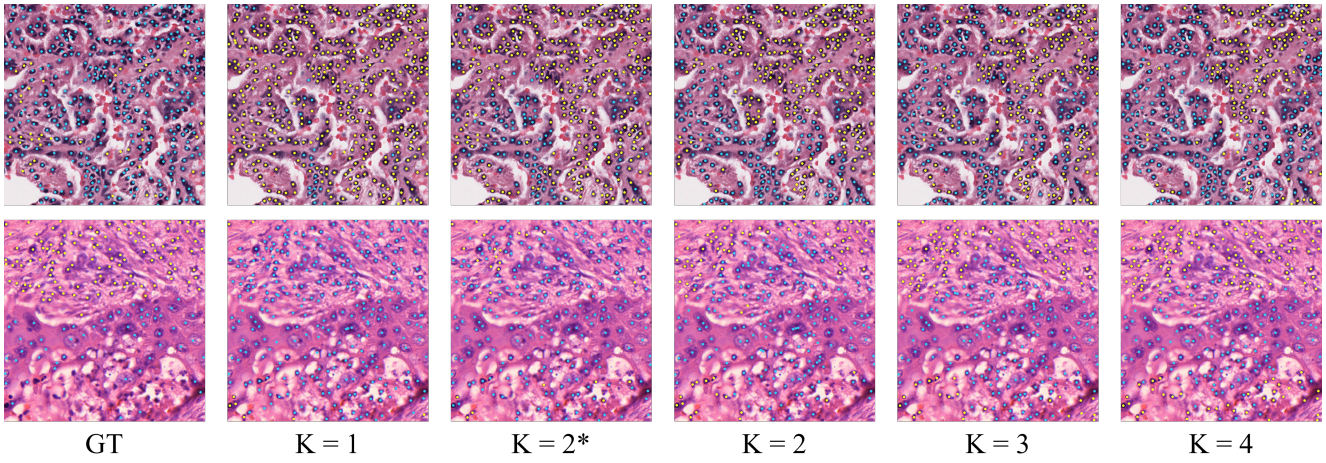

Figure 4: Qualitative comparison results on the OCELOT dataset. The symbol $*$ indicates that the predictions come from MFoVCE-Net+, while the rest are from our proposed mFoV DPA-P2PNet. Cyan: tumor cells. Yellow: background cells.

Metrics	*ResNet-50*					*ViT-B/16*			
	Random	IN	MoCo v2	SwAV	DINO	Random	IN	DINO	MAE
F1	52.3	55.9	**57.4**	56.4	56.6	56.7	58.0	60.8	**61.7**
AP	39.3	43.7	**45.4**	44.7	45.0	44.0	47.0	49.9	**51.6**

Table 7: Downstream evaluation of various SSL algorithms on the large-scale and real-world PD-L1 dataset. IN stands for using ImageNet-supervised pre-trained weights as model initialization.

be processed. To facilitate a straightforward comparison, we visualize the cell detection results under different input conditions in Fig. 4.

Ablation Studies We conduct ablation experiments on the CoNSeP dataset to validate the effectiveness of our proposed modules: multi-scale decoding (MSD) and deformable point proposals (DPP). The results are summarized in Tab. 6. When using only the MSD, our model achieves a performance gain of 0.7% on F1 and 1.2% on AP compared to the original P2PNet. When combined with the DPP, further improvements can be obtained with 1.1% and 2.4% on F1 and AP, respectively.

Effect of SSL To evaluate the transferability of learned weights by SSL, we apply the full fine-tuning protocol to train DPA-P2PNet with a pre-trained backbone on the PD-L1 dataset. Two architectures of backbones including ResNet-50 and ViT-B/16 (Dosovitskiy et al. 2020) are tested in our experiments. Moreover, to make the plain ViT model suitable for the dense prediction tasks (e.g., PCD), we introduce ViT-Adapter to inject the image-related inductive biases into the model and construct hierarchical features.

The downstream performance is shown in Table 7. We observe that supervised ImageNet pre-training is better than training from scratch but lags behind domain-specific SSL pre-training for both ResNet-50 and ViT-B/16 models, which aligns with the conclusions drawn in (Kang et al. 2023). Of the ResNet-50 based SSL methods, MoCo v2 achieves the most favorable results, outshining the ImageNet-supervised pre-training by 1.5% on F1 and 1.7% on AP. Regarding the ViT-B/16 based SSL methods, MAE demonstrates the best performance and it surpasses the ImageNet-supervised pre-training remarkably by 3.7% on F1 and 4.6% on AP. We attribute the superiority of MAE over DINO to the fact that contrastive learning primarily captures global relationships, while masked image modeling captures local relationships that is specially beneficial for dense prediction tasks (Park et al. 2023).

Conclusion

In this study, we present DPA-P2PNet for point-based cell detection. The key improvements of DPA-P2PNet over the prototype model are multi-scale decoding and deformable point proposals, which are designed to promote the utilization of multi-scale information within histopathology images and mitigate the distribution bias between pre-defined point proposals and potential cells, respectively. The ablation studies validate their efficacy and extensive comparison experiments on three histopathology datasets with various staining styles demonstrate the effectiveness and generalization of our proposed DPA-P2PNet model. Based on this, we also design mFoV DPA-P2PNet, the power and scalability of which are validated on the OCELOT dataset. Moreover, we execute the first self-supervised pre-training on large scale IHC image data and evaluate the efficacy of various SSL methods on the PCD task specially.

References

Abousamra, S.; Belinsky, D.; Van Arnam, J.; Allard, F.; Yee, E.; Gupta, R.; Kurc, T.; Samaras, D.; Saltz, J.; and Chen, C. 2021. Multi-class cell detection using spatial context representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4005–4014.

Bai, T.; Xu, J.; and Xing, F. 2020. Multi-field of view aggregation and context encoding for single-stage nucleus recognition. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, 382–392. Springer.

Bai, T.; Xu, J.; Zhang, Z.; Guo, S.; and Luo, X. 2022. Context-aware learning for cancer cell nucleus recognition in pathology images. *Bioinformatics*, 38(10): 2892–2898.

Cai, J.; Zhu, C.; Cui, C.; Li, H.; Wu, T.; Zhang, S.; and Yang, L. 2021. Generalizing nucleus recognition model in multi-source ki67 immunohistochemistry stained images via domain-specific pruning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, 277–287. Springer.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33: 9912–9924.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chen, R. J.; Chen, C.; Li, Y.; Chen, T. Y.; Trister, A. D.; Krishnan, R. G.; and Mahmood, F. 2022a. Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chen, T.; Liu, S.; Chang, S.; Cheng, Y.; Amini, L.; and Wang, Z. 2020a. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 699–708.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Chen, Z.; Duan, Y.; Wang, W.; He, J.; Lu, T.; Dai, J.; and Qiao, Y. 2022b. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*.

Cheng, G.; Zhang, F.; Xing, Y.; Hu, X.; Zhang, H.; Chen, S.; Li, M.; Peng, C.; Ding, G.; Zhang, D.; et al. 2022. Artificial intelligence-assisted score analysis for predicting the expression of the immunotherapy biomarker PD-L1 in lung cancer. *Frontiers in Immunology*, 13.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gildenblat, J.; and Klaiman, E. 2019. Self-supervised similarity learning for digital pathology. *arXiv preprint arXiv:1905.08139*.

Graham, S.; Vu, Q. D.; Raza, S. E. A.; Azam, A.; Tsang, Y. W.; Kwak, J. T.; and Rajpoot, N. 2019. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58: 101563.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ho, D. J.; Yarlagadda, D. V.; D'Alfonso, T. M.; Hanna, M. G.; Grabenstetter, A.; Ntiamoah, P.; Brogi, E.; Tan, L. K.; and Fuchs, T. J. 2021. Deep multi-magnification networks for multi-class breast cancer image segmentation. *Computerized Medical Imaging and Graphics*, 88: 101866.

Howard, R.; Kanetsky, P. A.; and Egan, K. M. 2019. Exploring the prognostic value of the neutrophil-to-lymphocyte ratio in cancer. *Scientific reports*, 9(1): 1–10.

Huang, Z.; Ding, Y.; Song, G.; Wang, L.; Geng, R.; He, H.; Du, S.; Liu, X.; Tian, Y.; Liang, Y.; et al. 2020. Bcdata: A large-scale dataset and benchmark for cell detection and counting. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, 289–298. Springer.

Hutter, C.; and Zenklusen, J. C. 2018. The cancer genome atlas: creating lasting value beyond its data. *Cell*, 173(2): 283–285.

Jiao, Y.; Li, J.; Qian, C.; and Fei, S. 2021. Deep learning-based tumor microenvironment analysis in colon adenocarcinoma histopathological whole-slide images. *Computer Methods and Programs in Biomedicine*, 204: 106047.

Kang, M.; Song, H.; Park, S.; Yoo, D.; and Pereira, S. 2023. Benchmarking Self-Supervised Learning on Diverse Pathology Datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3344–3354.

Li, H.; Zhu, C.; Zhang, Y.; Sun, Y.; Shui, Z.; Kuang, W.; Zheng, S.; and Yang, L. 2023. Task-specific fine-tuning via variational information bottleneck for weakly-supervised

pathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7454–7463.

Lian, D.; Li, J.; Zheng, J.; Luo, W.; and Gao, S. 2019. Density map regression guided detection network for rgb-d crowd counting and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1821–1830.

Liang, D.; Xu, W.; and Bai, X. 2022. An end-to-end transformer model for crowd localization. *arXiv preprint arXiv:2202.13065*.

Liang, D.; Xu, W.; Zhu, Y.; and Zhou, Y. 2022. Focal inverse distance transform maps for crowd localization. *IEEE Transactions on Multimedia*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, W.; and Chan, A. B. 2023. Optimal Transport Minimization: Crowd Localization on Density Maps for Semi-Supervised Counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21663–21673.

Liu, Z.; Chen, H.; Feng, R.; Wu, S.; Ji, S.; Yang, B.; and Wang, X. 2021. Deep dual consecutive network for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 525–534.

Matsoukas, C.; Haslum, J. F.; Sorkhei, M.; Söderberg, M.; and Smith, K. 2022. What makes transfer learning work for medical images: feature reuse & other factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9225–9234.

Park, N.; Kim, W.; Heo, B.; Kim, T.; and Yun, S. 2023. What Do Self-Supervised Vision Transformers Learn? *arXiv preprint arXiv:2305.00729*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. 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*, 234–241. Springer.

Ryu, J.; Puche, A. V.; Shin, J.; Park, S.; Brattoli, B.; Lee, J.; Jung, W.; Cho, S. I.; Paeng, K.; Ock, C.-Y.; et al. 2023. OCELOT: Overlapped Cell on Tissue Dataset for Histopathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23902–23912.

Schmitz, R.; Madesta, F.; Nielsen, M.; Krause, J.; Steurer, S.; Werner, R.; and Rösch, T. 2021. Multi-scale fully convolutional neural networks for histopathology image segmentation: from nuclear aberrations to the global tissue architecture. *Medical image analysis*, 70: 101996.

Shui, Z.; Zhang, S.; Zhu, C.; Wang, B.; Chen, P.; Zheng, S.; and Yang, L. 2022. End-to-end cell recognition by point annotation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV*, 109–118. Springer.

Song, Q.; Wang, C.; Jiang, Z.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; and Wu, Y. 2021. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3365–3374.

Sowrirajan, H.; Yang, J.; Ng, A. Y.; and Rajpurkar, P. 2021. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, 728–744. PMLR.

Tokunaga, H.; Teramoto, Y.; Yoshizawa, A.; and Bise, R. 2019. Adaptive weighting multi-field-of-view CNN for semantic segmentation in pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12597–12606.

Van Rijthoven, M.; Balkenhol, M.; Siliņa, K.; Van Der Laak, J.; and Ciompi, F. 2021. HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. *Medical image analysis*, 68: 101890.

Wan, J.; Liu, Z.; and Chan, A. B. 2021. A generalized loss function for crowd counting and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1974–1983.

Wang, X.; Yang, S.; Zhang, J.; Wang, M.; Zhang, J.; Huang, J.; Yang, W.; and Han, X. 2021. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, 186–195. Springer.

Xu, Y.; Zhang, Q.; Zhang, J.; and Tao, D. 2021. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 34: 28522–28535.

Zhang, S.; Zhu, C.; Li, H.; Cai, J.; and Yang, L. 2022. Weakly supervised learning for cell recognition in immunohistochemical cytoplasm staining images. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 1–5. IEEE.

Zhou, L.; Liu, H.; Bae, J.; He, J.; Samaras, D.; and Prasanna, P. 2022. Self pre-training with masked autoencoders for medical image analysis. *arXiv preprint arXiv:2203.05573*.

Zhou, Y.; Dou, Q.; Chen, H.; Qin, J.; and Heng, P.-A. 2018. Sfcn-opi: Detection and fine-grained classification of nuclei using sibling fcn with objectness prior interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv preprint arXiv:2010.04159*.