# Towards Effective and Efficient Context-aware Nucleus Detection in Histopathology Whole Slide Images

**Zhongyi Shui[1,2*], Honglin Li[1,2*], Yunlong Zhang[1,2], Yuxuan Sun[1,2], Yiwen Ye[3], Pingyi Chen[1,2], Ruizhe Guo[1,2], Lei Cui[4], Chenglu Zhu[2†], Lin Yang[2,5,6†]**

[1] College of Computer Science and Technology, Zhejiang University
[2] School of Engineering, Westlake University
[3] College of Computer Science and Technology, Northwestern Polytechnical University
[4] School of Information Science and Technology, Northwest University
[5] The Institute of Advanced Technology, Westlake Institute for Advanced Study
[6] Center for Interdisciplinary Research and Innovation, MuyuanLaboratory

## Abstract

Nucleus detection in histopathology whole slide images (WSIs) is crucial for a broad spectrum of clinical applications. The gigapixel size of WSIs necessitates the use of sliding window methodology for nucleus detection. However, mainstream methods process each sliding window independently, which overlooks broader contextual information and easily leads to inaccurate predictions. To address this limitation, recent studies additionally crop a large Filed-of-View (LFoV) patch centered on each sliding window to extract contextual features. However, such methods substantially increase whole-slide inference latency. In this work, we propose an effective and efficient context-aware nucleus detection approach. Specifically, instead of using LFoV patches, we aggregate contextual clues from off-the-shelf features of historically visited sliding windows, which greatly enhances the inference efficiency. Moreover, compared to LFoV patches used in previous works, the sliding window patches have higher magnification and provide finer-grained tissue details, thereby enhancing the classification accuracy. To develop the proposed context-aware model, we utilize annotated patches along with their surrounding unlabeled patches for training. Beyond exploiting high-level tissue context from these surrounding regions, we design a post-training strategy that leverages abundant unlabeled nucleus samples within them to enhance the model's context adaptability. Extensive experimental results on three challenging benchmarks demonstrate the superiority of our method.

**Code** — https://github.com/windygoo/PathContext

## Introduction

Nucleus detection in histopathology whole slide images (WSIs) is a fundamental task in computational pathology. It allows quantitative analysis of WSIs, which can lead to better cancer diagnosis, grading, prognosis, and treatment planning while maintaining medical interpretability (Diao et al. 2021; Ryu et al. 2023; Yang et al. 2024; Ignatov, Yates, and Boeva 2024; Xu et al. 2025b). Therefore, the development of precise automatic nucleus detection algorithms has become a critical research focus in recent years. Current nucleus detection pipeline involves training a nucleus detector using expert-annotated histopathology patches and then deploys it to detect nuclei in gigapixel WSIs through a sliding window technique (Huang et al. 2020; Shui et al. 2022; Zhang et al. 2022; Huang et al. 2023b), as illustrated in Fig. 1 (a). However, for accurate nucleus localization, the annotated and sliding window patches are cropped at high magnification but small Field-of-View (FoV) (Ryu et al. 2023). As a result, the nucleus detector can only see a limited context without understanding broader tissue information, which can easily lead to inaccurate predictions. In clinical practice, pathologists first examine tissue context at large FoVs and then zoom in to observe detailed nuclear morphology for accu-

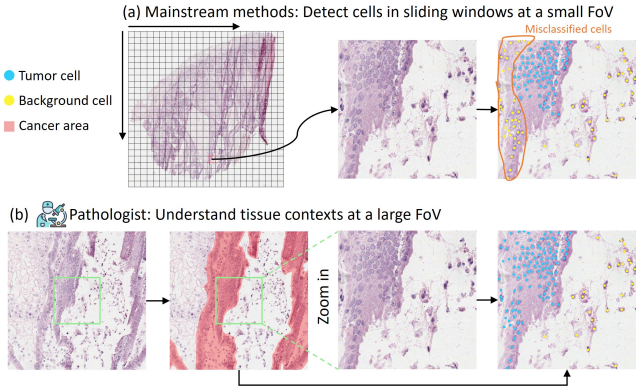


Figure 1: Nucleus detection on gigapixel WSIs necessitates a sliding window strategy. (a) Mainstream methods detect nuclei in each window patch without understanding broader tissue structure, which easily leads to inaccurate predictions. (b) Pathologists first zoom out to examine tissue context at large FoVs and then zoom in to observe detailed nuclear morphology for accurate nucleus classification (Ryu et al. 2023).

*These authors contributed equally.
†Corresponding author.

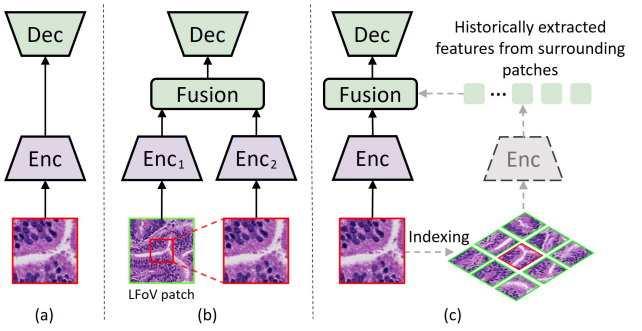


Figure 2: (a) Mainstream nucleus detection methods operate on patches at a single FoV without considering tissue context. (b) Previous context-aware nucleus detection approaches leverage large FoV (LFoV) patches to extract contextual information, which substantially increases the whole-slide inference time due to additional I/O-intensive data preparation. (c) The proposed context-aware method aggregates contextual information from off-the-shelf features of historically visited window patches, which greatly improves the whole-slide inference efficiency.

rate assessments, as depicted in Fig. 1 (b).

Inspired by this clinical workflow, recent works (Bai et al. 2020, 2022; Shui et al. 2024b; Ryu et al. 2023; Millward, He, and Nibali 2023; Torbati et al. 2025) utilize a low-magnification large FoV (LFoV) patch, which has the same size as sliding window patches but encompass broader WSI regions, as a supplementary input of nucleus detectors to enable context-aware nucleus detection, as shown in Fig. 2 (b). Despite the improved outcomes, we argue that this line of approaches exhibits two critical limitations: (1) The requirement for additional processing the LFoV patches significantly increases the whole-slide inference time. (2) The LFoV patches at low magnification inherently lack fine-grained tissue details, thereby diminishing the potential performance gains.

In this work, we propose to aggregate contextual information from patches that surround with the region-of-interest (ROI) patch, as illustrated by Fig. 2 (c). Notably, in the training stage, the ROI patch comes from the annotated set while during inference on WSIs, the ROI patch and its surrounding patches are both part of sliding windows. Therefore, this design eliminates the I/O-intensive and time-consuming step of additionally preparing LFoV patches in previous studies. Moreover, since the ROI patch and its surrounding patches have the same magnification (*i.e.*, data distribution), we employ a shared image encoder to process both and utilize features extracted from surrounding patches as contextual clues. With this design, we can directly re-use the off-the-shelf features extracted from historically visited surrounding windows to perform context-aware nucleus detection, which further improves the whole-slide inference speed.

Additionally, we observe that current context-aware methods exclusively exploit high-level contextual features in LFoV patches while neglecting the massive unlabeled nuclei within these patches. To harness this untapped resource, we introduce a post-training stage in this work. Specifically, we employ the pre-trained detector to detect these unlabeled nuclei and generate pseudo labels for them using a novel cross-labeling strategy. Then, we use these pseudo-labeled samples to fine-tune the detector, empowering it to classify nuclei at different spatial locations (*i.e.*, various context conditions) and thus enhancing the model's context adaptability. Besides, we discover for the first time that incorporating high-level contextual features inherently diminishes the model's perception of low-level nuclear morphological details, which potentially comprises the model's accuracy. To address this limitation, we introduce an lightweight auxiliary branch to compensate for these morphological features. Our main contributions can be summarized as follows:

1. We propose a novel context aggregation approach that exploits features from surrounding sliding windows for effective and efficient nucleus detection in WSIs.
2. We propose a cross-labeling strategy to effectively utilize unlabeled nuclei in surrounding patches to improve the model's context adaptability.
3. Extensive experiments on three challenging benchmarks demonstrate the advantages of our method over the state-of-the-art counterparts on both nucleus detection and instance segmentation tasks.

## Related Works

### Nucleus Detection

Current methods for nucleus detection can be divided into two categories: density map-based and end-to-end.

Density map-based methods (Graham et al. 2019; Abousamra et al. 2021; Zhang et al. 2022; Pan et al. 2023; Lou et al. 2024a; Hörst et al. 2024) first regress nucleus probability maps, and then apply post-processing including thresholding, local maxima detection, and non-maximum suppression to identify nuclei centroids. In contrast, end-to-end methods can directly predict nuclear positions without hand-crafted post-processing procedures. These approaches can be further categorized into two distinct paradigms: anchor-based and anchor-free. Anchor-based methods, primarily built upon P2PNet (Song et al. 2021), localize nuclei by predicting relative offsets from pre-defined anchor points across input images (Song et al. 2021; Shui et al. 2022, 2024b), while anchor-free approaches (Huang et al. 2023b,a; Pina, Dorca, and Vilaplana 2024) directly predict absolute nuclear positions through learnable queries following the DETR architecture (Carion et al. 2020). Due to the substantial variability in nuclei size, shape and spatial distribution, it is impossible to design a universal post-processing approach that generalizes well across all scenarios. Consequently, density map-based methods generally exhibit inferior detection performance and efficiency compared to end-to-end approaches.

### Context-aware Pathology Image Analysis

Due to the gigapixel scale of WSIs, dense prediction tasks such as tissue segmentation and nucleus detection on them must be conducted in a sliding window manner. However,

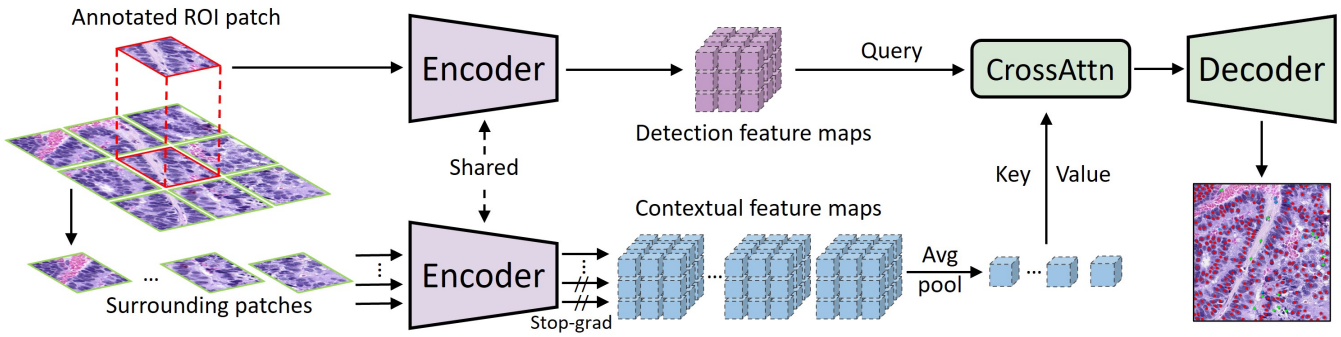


Figure 3: The training pipeline of our proposed context-aware nucleus detection method. We use a shared visual encoder to encode the annotated patch and its surrounding patches. To avoid GPU memory overflow from simultaneously processing numerous surrounding patches with gradient computation, we randomly select a small subset to participate in back-propagation while the rest undergo forward pass only. The contextual features are first downsampled through grid average pooling and then incorporated into the detection branch via cross-attention.

this strategy can lead to the loss of broader spatial context when window patches are processed independently. To address this issue, several recent works have explored context-aware approaches for tissue segmentation (Kamnitsas et al. 2017; Tokunaga et al. 2019; Schmitz et al. 2021; Van Rijthoven et al. 2021) and nucleus detection (Bai et al. 2020, 2022; Ryu et al. 2023; Shui et al. 2024b).

Current context-aware nucleus detection approaches can be divided into two categories: explicit and implicit. Explicit approaches (Ryu et al. 2023; Schoenpflug and Koelzer 2023; Torbati et al. 2025) rely on additional tissue region annotations. Specifically, they utilize an auxiliary model for tissue segmentation at LFoV, and the predicted tissue mask is fed into the detection model to enhance nucleus classification accuracy. In contrast, the latter approaches (Bai et al. 2020, 2022; Shui et al. 2024b) eliminate the need for tissue mask labels and learn context features from LFoV patches implicitly, as illustrated in Fig. 2 (b). Despite notable performance improvements, all these methods introduce LFoV patches as additional input, which significantly increases the inference overhead. Essentially, the method proposed in this work belongs to the implicit family. However, different from previous works that utilize LFoV patches, we aggregate contextual information from off-the-shelf features of historically visited sliding window patches, which greatly improves the whole-slide inference efficiency.

# Method

## Background

Mainstream nucleus detectors takes a single patch as input (Graham et al. 2019; Shui et al. 2024b; Hörst et al. 2024). Such models are trained on an annotated patch set $\mathcal{D} = \{(x_i, p_i, y_i)\}_{i=1}^{N}$. For each patch $x_i \in \mathbb{R}^{H\times W\times 3}$, the annotation $p_i$ and $y_i$ represent centroids and categories of all nuclei in it, respectively. To develop context-aware nucleus detector, each annotated patch $x_i$ is complemented by its surrounding unlabeled patches $\{x_{i,j,k} \in \mathbb{R}^{H\times W\times 3} \mid j,k \in \{-\delta,\ldots,\delta\}\}$, where $\delta$ denotes the size of context area considered by the detector. For instance, when $\delta = 1$, the model learns to detect nuclei in an annotated patch while considering tissue context from its corresponding $3 \times 3$ neighborhood. Following (Shui et al. 2024b), we adopt P2PNet, an end-to-end nucleus detector, as the base model. Fig. 3 depicts the training pipeline of our proposed method.

## Extraction of Context Features

Current context-aware nucleus detection works (Bai et al. 2020, 2022; Shui et al. 2024b; Ryu et al. 2023) all leverage LFoV patches to extract contextual information that improves nucleus detection accuracy in ROI patches at small FoV. As these two types of patches have different magnifications and thereby lie in distinct data distributions (Chen and Krishnan 2022), these approaches employ separate image encoders to process them. Differently, since we extract contextual features from patches that surround the ROI patch and they have the same magnification, a shared image encoder is employed in this work.

During training, we encode an annotated patch $x_i$ into $\mathcal{F}_i \in \mathbb{R}^{h\times w\times d}$ and its surrounding patches into context feature maps $\{\mathcal{F}_{i,j,k} \in \mathbb{R}^{h\times w\times d} \mid j,k \in \{-\delta,\ldots,\delta\}\}$. Unlike previous methods that involve only two patches in each iteration, our approach requires encoding $(2\delta+1)^2$ patches concurrently. When $\delta = 1$, this amounts to 9 patches, and enabling gradient computation for all of them leads to prohibitive memory requirements. To address this challenge, we propose a selective gradient computation strategy. Specifically, in each iteration, we randomly select $k$ surrounding patches for back-propagation while the rest $(2\delta+1)^2 - k$ patches undergo feature extraction in the gradient-free manner. This approach preserves the model's capability to capture informative contextual features while significantly reducing memory consumption.

To eliminate spatial redundancy (Li et al. 2025) in each context feature map, we downsample $\mathcal{F}_{i,j,k}$ by partitioning it into a uniform $s \times s$ grid and apply average pooling within each grid cell. This reduces the resolution of $\mathcal{F}_{i,j,k}$ from $hw$ to $s^2$, where $s \ll \min(h,w)$ is set empirically in this work. Finally, we concatenate all compressed context feature maps

as $\mathcal{F}_i^{ctx} \in \mathbb{R}^{(2\delta+1)^2 \times s \times s \times d}$.

## Injection of Context Features

To enable context-aware nucleus detection, we inject the context feature maps $\mathcal{F}_i^{ctx}$ into the hidden embedding $\mathcal{F}_i$ extracted from the annotated patch via cross-attention:

$$\mathcal{F}_i' = \mathrm{CrossAttn}\left(\mathrm{Q} = \mathcal{F}_\mathrm{i}, \mathrm{K} = \mathcal{F}_\mathrm{i}^\mathrm{ctx}, \mathrm{V} = \mathcal{F}_\mathrm{i}^\mathrm{ctx}\right) \quad (1)$$

where $\mathcal{F}_i$ serves as query and $\mathcal{F}_i^{ctx}$ serves as both key and value. Finally, the context-enriched $\mathcal{F}_i'$ is fed into the decoder to predict nucleus centroids and categories in the annotated patch. It is worth noting that we observe no performance gains when adding positional embeddings in Eq. 1. We hypothesize that this is because histopathology slides are inherently continuous, with adjacent patches displaying coherent visual content along their boundaries. This continuity implicitly encodes relative positional relationship.

## Enhancing Context Adaptability with Unlabeled Nuclei

We observe that the LFoV patches or surrounding patches provide not only high-level tissue context but also abundant unlabeled nucleus samples. To leverage this resource, we introduce a post-training stage that enables the model to classify nuclei at different spatial locations (*i.e.*, various context conditions) to enhance its context adaptability.

In general, the proposed context-aware nucleus detection method can be decomposed into two steps: (1) generating context-enriched embedding $e \in \mathbb{R}^d$ for each point proposal (Song et al. 2021) and (2) performing classification via a classifier head $\phi : e \rightarrow Y \in \mathbb{R}^{C+1}$, where $C$ represents the number of nucleus categories and the extra class is background. For each annotated patch $x_i$, we utilize the nucleus detector pre-trained in the above sections to identify nuclei in its surrounding patches $\{x_{i,j,k} \mid j,k \in \{-\delta,\ldots,\delta\}\}$ while concurrently generating pseudo class labels for them. This yields additional training data $\{(x_{i,j,k}, \hat{p}_{i,j,k}, \hat{y}_{i,j,k}) \mid j,k \in \{-\delta,\ldots,\delta\}\}$ along with context-enriched embeddings $e$ for each identified nucleus. Then, we train a MLP head $\phi' : e \rightarrow Y' \in \mathbb{R}^C$, where the embeddings $e$ corresponding to nuclei in surrounding patches are supervised by pseudo labels $\hat{y}_{i,j,k}$ while those from the annotated patch use ground-truth class labels $y_i$.

However, we observe that using pseudo labels predicted by the nucleus detector itself leads to marginal improvements, even when we apply a high confidence threshold (*e.g.*, 0.9) to filter out unreliable pseudo labels. This can be attributed to the confirmation bias inherent in self-training approaches (Arazo et al. 2020), which leads to error accumulation when the model's own predictions are used for self-supervision. To mitigate this problem, we propose a cross-labeling technique. Specifically, we convert point annotations to pseudo mask labels following (Lin et al. 2024) and train a lightweight auxiliary model for multi-class nucleus segmentation. Afterwards, we feed each surrounding patch $x_{i,j,k}$ into it and extract pseudo labels $\hat{y}_{i,j,k}$ for all predetected nuclei according to their coordinates $\hat{p}_{i,j,k}$ on the predicted class maps. We find although the auxiliary model shows only comparable classification accuracy to our detector, the substantial difference on model architecture (density map-based vs. end-to-end) and training regime leads to distinct classification patterns and greatly alleviates the error accumulation problem.

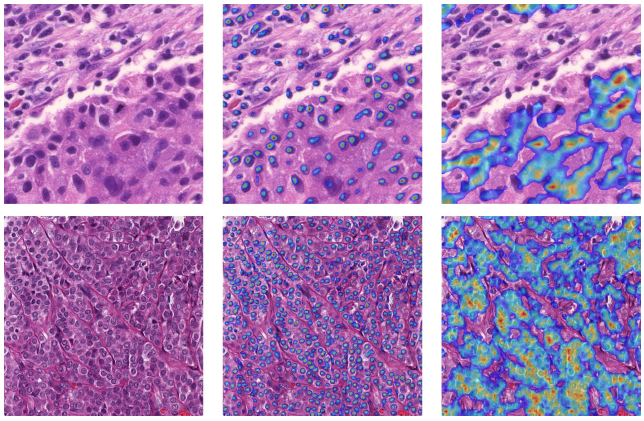

Figure 4: Input image (left) and corresponding Grad-CAM++ attention maps of context-free (middle) and context-aware (right) P2PNet models. It can be observed that incorporating high-level contextual features dilutes the model's attention of local nuclear morphological details.

## Revitalizing Nuclear Morphological Perception

Current context-aware nucleus detection methods primarily focus on incorporating richer tissue context to improve model performance (Shui et al. 2024b). However, we discover that integrating high-level contextual features inevitably dilutes the model's attention of fine-grained nuclear morphological details, as depicted in Fig. 4. This degraded perception could compromise the model's performance, as nuclear shape, size, chromatin pattern and texture are critical for accurate nucleus classification (Hörst et al. 2024).

Given that the segmentation task naturally models nuclear morphology, we utilize the auxiliary model developed in the previous section to alleviate this limitation. Specifically, we extract morphology-rich $m$ from the input feature maps of the model's last layer, which are specifically optimized to delineate nuclei regions and thereby contain abundant nuclear morphological information (Chen et al. 2023). Finally, we replace the input of $\phi'$ from $e$ to $[e; m]$, leveraging complementary merits of nuclear contextual and morphological features for accurate classification.

During inference, we first feed $e$ to $\phi$ to identify foreground nuclei, then feed $[e; m]$ of each detected nucleus to $\phi'$ for category prediction.

# Experiments

## Experimental Setup

**Datasets.** To the best of our knowledge, there are currently no publicly available datasets with exhaustive WSI-level nucleus annotations. Consequently, we conduct experiments on three patch-level benchmarks that provide context images for each annotated patch.

| Method | BRCA | | | | OCELOT | | | PUMA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $F^{Inf.}$ | $F^{Tum.}$ | $F^{Str.}$ | $F_{avg}$ | $F^{Tum.}$ | $F^{Back.}$ | $F_{avg}$ | $F^{Tum.}$ | $F^{Tils.}$ | $F^{Oth.}$ | $F_{avg}$ |
| Hover-net | 62.31±0.91 | 75.25±0.24 | 49.58±0.36 | 62.38±0.22 | 69.27±0.08 | 61.03±0.46 | 65.15±1.57 | 78.36±0.24 | 75.72±0.15 | 49.53±0.16 | 67.87±0.13 |
| MCSpatNet | 64.45±0.61 | 79.34±0.28 | 55.31±0.39 | 66.37±0.17 | 70.69±0.39 | 56.37±1.23 | 63.53±0.78 | 83.30±0.16 | 77.74±0.55 | 52.76±1.31 | 71.27±0.39 |
| P2PNet | 64.25±0.50 | 79.21±0.15 | 55.20±0.28 | 66.22±0.10 | 72.55±0.27 | 61.64±0.21 | 67.09±0.23 | 83.35±0.16 | 80.55±0.13 | 56.93±0.37 | 73.61±0.17 |
| Semi-P2PNet | 64.72±0.73 | 81.11±0.30 | 57.40±0.33 | 67.74±0.38 | 73.12±0.36 | 62.03±0.05 | 67.58±0.20 | 83.69±0.13 | 80.56±0.34 | 57.77±0.27 | 74.01±0.07 |
| AC-Former | 59.77±2.02 | 74.67±0.33 | 48.15±1.59 | 60.87±0.03 | 69.84±0.13 | 58.74±0.40 | 64.29±0.20 | 79.51±0.11 | 74.86±0.48 | 54.15±0.53 | 69.51±0.20 |
| SMILE | 71.82±0.41 | 80.27±0.12 | 52.11±0.07 | 68.06±0.13 | 70.15±0.16 | 61.08±0.49 | 65.62±0.31 | 79.89±0.46 | 80.34±0.18 | 53.10±0.32 | 71.11±0.27 |
| PointNu-Net | 70.95±0.51 | 80.04±0.27 | 54.16±0.16 | 68.38±0.29 | 70.10±0.29 | 59.01±0.30 | 64.55±0.28 | 81.03±0.12 | 76.26±0.16 | 53.53±0.30 | 70.27±0.18 |
| CellViT | 68.12±0.60 | 78.60±0.47 | 50.27±0.75 | 65.66±0.49 | 70.25±0.51 | 60.78±0.64 | 65.52±0.48 | 81.56±0.24 | 78.85±0.41 | 57.04±0.68 | 72.49±0.21 |
| SENC | 56.89±0.33 | 76.95±0.12 | 50.31±0.07 | 61.38±0.17 | 73.70±0.09 | 63.95±0.23 | 68.83±0.16 | 79.87±0.77 | 76.38±0.60 | 53.14±0.71 | 69.80±0.69 |
| CGT | 56.50±0.11 | 76.61±0.02 | 51.80±0.05 | 61.63±0.05 | 72.08±0.06 | 62.80±0.13 | 67.44±0.09 | 79.89±0.05 | 76.69±0.02 | 54.51±0.05 | 70.36±0.03 |
| TopoCellGen | 65.20±0.40 | 81.70±0.60 | 58.20±0.50 | 68.40±0.40 | - | - | - | - | - | - | - |
| MFoVCE-Net | 60.30±1.04 | 78.79±0.07 | 55.77±1.77 | 64.95±0.22 | 72.81±0.59 | 63.17±0.15 | 67.99±0.37 | 79.80±0.84 | 75.96±1.13 | 55.16±1.17 | 70.31±0.84 |
| MFoV-P2PNet | 63.52±0.59 | 80.71±0.24 | 55.84±0.57 | 66.69±0.37 | 74.70±0.09 | 63.48±0.08 | 69.09±0.06 | 84.17±0.29 | 79.96±0.54 | 59.70±0.60 | 74.61±0.28 |
| **Ours** | **72.68**±0.19 | **83.49**±0.10 | **59.87**±0.27 | **72.01**±0.13 | **75.24**±0.28 | **66.43**±0.10 | **70.83**±0.15 | **86.45**±0.15 | **82.17**±0.29 | **63.45**±0.62 | **77.36**±0.30 |

Table 1: Comparison of nucleus detection performance across three benchmarks. $F^{Inf.}$, $F^{Tum.}$, $F^{Str.}$, $F^{TILs.}$ and $F^{Oth.}$ denote the F1-socre for the inflammatory, tumor, stromal, TILs and other nuclei, respectively. $F_{avg}$ represents the average F1-score. The best and second-best performance are highlighted in **bold** and underlined, respectively.

| Method | BRCA | | | | OCELOT | | | PUMA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $PQ^{Inf.}$ | $PQ^{Tum.}$ | $PQ^{Str.}$ | $PQ_{avg}$ | $PQ^{Tum.}$ | $PQ^{Back.}$ | $PQ_{avg}$ | $PQ^{Tum.}$ | $PQ^{Tils.}$ | $PQ^{Oth.}$ | $PQ_{avg}$ |
| Hover-net | 46.97±0.69 | 64.58±0.16 | 45.52±0.27 | 52.36±0.19 | 61.15±0.40 | 43.44±0.63 | 52.29±0.29 | 64.17±0.18 | 56.36±0.06 | 35.17±0.19 | 51.90±0.09 |
| MCSpatNet | 51.24±0.93 | 69.66±0.18 | 44.65±0.37 | 55.18±0.38 | 65.89±0.18 | 38.88±1.30 | 52.39±0.65 | 66.74±0.21 | 55.30±0.45 | 31.96±1.40 | 51.33±0.54 |
| P2PNet | 49.92±1.46 | 69.15±0.18 | 46.74±0.14 | 55.27±0.48 | 66.21±0.28 | 44.44±0.36 | 55.33±0.28 | 67.16±0.17 | 58.11±0.25 | **39.95**±0.29 | 54.74±0.17 |
| Semi-P2PNet | 48.93±0.44 | 71.17±0.24 | 47.69±0.16 | 55.93±0.19 | 66.70±0.38 | 44.76±0.95 | 55.73±0.67 | 67.13±0.29 | 58.40±0.12 | 39.10±0.36 | 54.88±0.18 |
| AC-Former | 42.54±0.17 | 63.44±1.52 | 40.87±0.32 | 48.95±0.34 | 64.97±0.41 | 42.87±0.53 | 53.92±0.16 | 64.78±0.14 | 55.11±0.44 | 34.58±0.26 | 51.49±0.21 |
| SMILE | 45.71±0.94 | 70.95±0.24 | 44.66±0.18 | 53.77±0.33 | 63.07±0.22 | 45.56±0.64 | 54.32±0.29 | 65.44±0.43 | 57.27±0.17 | 37.44±0.46 | 53.38±0.14 |
| PointNu-Net | 50.34±0.25 | 71.16±0.33 | 46.78±0.13 | 56.09±0.14 | 64.08±0.28 | 43.95±0.27 | 54.01±0.23 | 66.76±0.23 | 54.76±0.12 | 36.64±0.54 | 52.72±0.28 |
| CellViT | 45.52±1.07 | 70.59±0.19 | 43.51±0.36 | 53.21±0.43 | 64.46±0.69 | 46.13±0.91 | 55.29±0.57 | 64.68±0.33 | 54.24±0.89 | 36.24±0.54 | 51.72±0.46 |
| SENC | 36.98±0.17 | 67.86±0.14 | 41.44±0.11 | 48.76±0.11 | 66.77±0.04 | 44.67±0.24 | 55.72±0.13 | 64.88±0.66 | 51.75±0.19 | 33.05±0.22 | 49.89±0.34 |
| CGT | 38.18±0.05 | 67.98±0.02 | 42.26±0.04 | 49.47±0.02 | 63.08±0.11 | 41.87±0.09 | 52.47±0.02 | 65.08±0.04 | 53.83±0.03 | 34.46±0.10 | 51.12±0.03 |
| MFoVCE-Net | 51.46±1.49 | 69.09±1.13 | 45.86±1.26 | 55.47±0.54 | 66.74±0.07 | 44.42±0.14 | 55.58±0.09 | 65.26±0.38 | 55.10±1.10 | 34.94±0.69 | 51.76±0.26 |
| MFoV-P2PNet | 50.01±1.79 | 70.85±0.23 | 46.63±0.56 | 55.83±0.68 | 67.18±0.29 | 45.52±0.16 | 56.35±0.15 | 68.42±0.24 | 57.83±0.35 | 38.23±0.54 | 54.83±0.28 |
| **Ours** | **54.82**±0.54 | **73.30**±0.25 | **49.24**±0.24 | **59.12**±0.18 | **67.62**±0.29 | **48.85**±0.36 | **58.24**±0.29 | **69.53**±0.13 | **58.53**±0.28 | 39.92±0.22 | **55.99**±0.13 |

Table 2: Comparison of nucleus instance segmentation performance across three benchmarks. To enable nucleus detection models to produce instance masks, we train the segmentor component of PromptNucSeg (Shui et al. 2024a) on the training sets of these datasets and employ different detection models as the prompter component within the PromptNucSeg framework.

- **BRCA** (Abousamra et al. 2021) is a breast cancer dataset and consists of 120 patches at 20× magnification belonging to 113 patients, collected from TCGA (Weinstein et al. 2013). The training, validation, and testing sets contain 80, 10, and 30 patches, respectively. The nuclei in this dataset are categorized into three types: tumor, inflammatory, and stromal.
- **OCELOT** (Ryu et al. 2023) comprises 664 patches at 40× magnification extracted from 303 WSIs. The dataset contains a total of 113,026 nuclei, annotated to differentiate between tumor and non-tumor nuclei. The training, validation, and test sets contain 400, 137, and 126 patches, respectively.
- **PUMA** (Shahamiri et al. 2025) comprises 206 patches extracted from melanoma tissue scanned at 40× magnification. It provides annotations for three nuclei types: tumor, tumor infiltrating lymphocytes (TILs), and other nuclei. We randomly divide this dataset into training, validation, and test subsets at a ratio of 6:2:2.

Additionally, we manually annotate instance masks for nuclei in the BRCA and OCELOT datasets. These annotations serve two purposes: enabling the training of baseline methods like CellViT (Hörst et al. 2024) that require mask supervision and facilitating the evaluation of different methods on context-aware nucleus instance segmentation.

**Evaluation metrics.** For nucleus detection, following previous studies (Abousamra et al. 2021; Ryu et al. 2023), we employ F1-score as the evaluation metric. If a detected nucleus is within a valid distance ($\sigma$) from an annotated nucleus and the nuclear class matches, it is counted as a true positive (TP), otherwise a false positive (FP). Then, the F1-score for the $c$-th class is calculated as: $\mathrm{F_c} = \frac{2\mathrm{TP_c}}{2\mathrm{TP_c}+\mathrm{FP_c}+\mathrm{FN_c}}$, and the average F1-score is $\bar{F} = \frac{1}{C}\sum_{c=1}^{C}\mathrm{F_c}$. In accordance with the official settings, $\sigma$ is set to 6, 15 and 15 pixels for the BRCA, OCELOT and PUMA benchmarks, respectively. For nucleus instance segmentation, we adopt Panoptic Quality (PQ) (Kirillov et al. 2019) as the evaluation metric. Following (Ryu et al. 2023), we repeat all experiments for 5 times with different random seeds and report the mean and 95% confidence interval of the performance metrics.

**Implementation details.** We use ResNet-50 (He et al. 2016) pre-trained on ImageNet-1K as the image encoder. The detector is trained using the AdamW (Loshchilov and Hutter 2017) optimizer with learning rate of 1e-4 for 200 epochs.

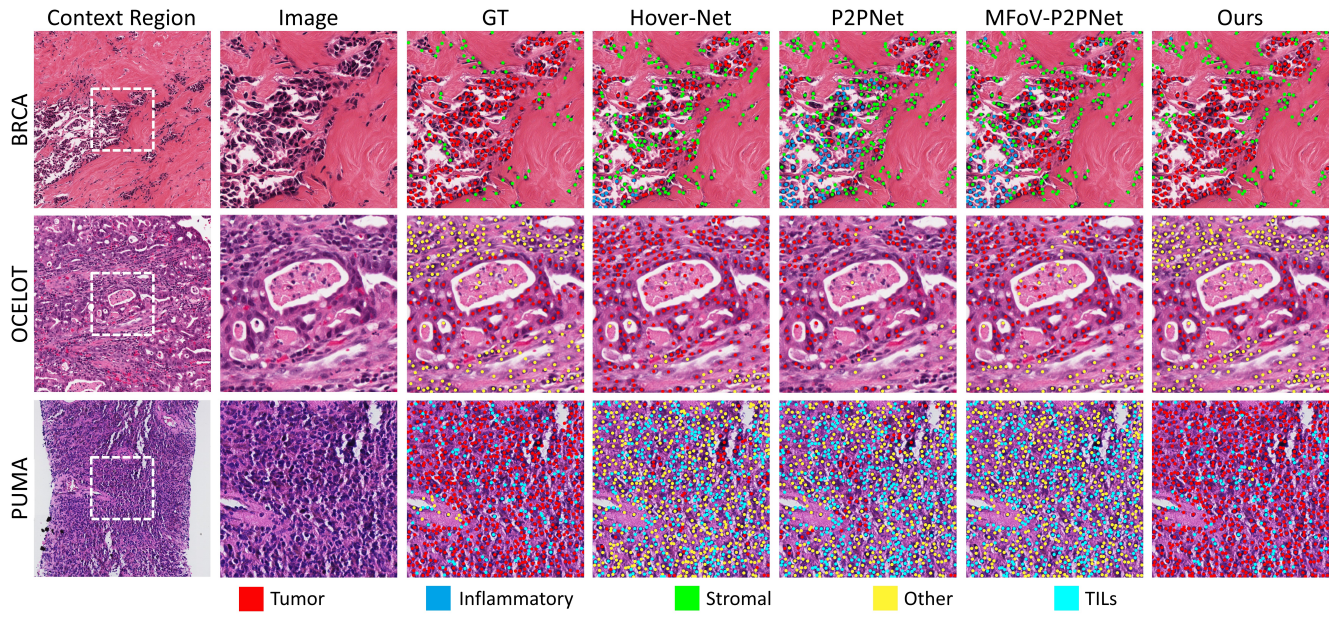


Figure 5: Qualitative comparison on three benchmarks. The white dashed boxes indicate the positions of ROI detection patches within the context images.

| Method | Params (M) | Time (s) | FLOPs (G) | $F_{avg}$ |
|---|---|---|---|---|
| CellViT | 142.85 | 3027.04 | 3566.71 | 65.66 |
| P2PNet | **27.26** | **148.86** | **100.08** | 66.22 |
| MFoV-P2PNet | 53.22 | 486.20 | 212.75 | 66.69 |
| Ours w/ ME | 48.08 | 205.81 | 186.74 | **72.01** |
| Ours w/o ME | 44.08 | 156.07 | 115.28 | 71.23 |

Table 3: Comparison on model size, computational cost, inference efficiency and performance of different methods. The inference time is measured using ten TCGA-BRCA WSIs with an average of 5k sliding window patches of size 1024×1024. ME represents retaining the auxiliary model to supplement nuclear morphology-rich embedding.

We use a batch size of 2, with all data distributed across 4 V100 GPUs. The auxiliary cross-labeling model comprises 12 convolutional blocks (Liu et al. 2022), FPN (Lin et al. 2017) and two PixelShuffle layers. It accounts for 9% of the total network parameters and is trained for 20 epochs. The head $\phi'$ consists of two linear layers and is trained for 100 epochs with cross-entropy loss. We set $k = 3$ and pooling size $o = 6$ in our experiments. Unless otherwise specified, the size of context area $\delta$ is set to 1.

## Comparison with the State-of-the-art Methods

We compare the proposed method with state-of-the-art (SOTA) nucleus instance segmentation methods including Hover-Net (Graham et al. 2019), SMILE (Pan et al. 2023), PointNu-Net (Yao et al. 2024) and CellViT (Hörst et al. 2024), as well as nucleus detection methods, including MC-SpatNet (Abousamra et al. 2021), P2PNet (Song et al. 2021), Semi-P2PNet (Shui et al. 2023), CGT (Lou et al. 2024a), SENC (Lou et al. 2024b), TopoCellGen (Xu et al. 2025a), MFoVCE-Net (Bai et al. 2020), and MFoV-P2PNet (Shui et al. 2024b). Among these methods, MFoVCE-Net and MFoV-P2PNet are context-aware approaches, while the others are context-free. Semi-P2PNet is a semi-supervised learning approach that also utilizes unlabeled nucleus samples in surrounding patches.

Tab. 1 and Tab. 2 exhibit the performance comparison results on the detection and segmentation tasks, respectively. For nucleus detection, the proposed method outperforms the SOTA counterparts by 3.61, 1.74 and 2.75 points in $F_{avg}$ on the BRCA, OCELOT and PUMA benchmarks. Compared to the baseline P2PNet model, our method shows substantial improvements of 5.79, 3.74 and 3.75 points on average F1 score. In task of nucleus instance segmentation, our method achieves 3.03, 1.89 and 1.11 points improvement over the SOTA methods in $PQ_{avg}$ across three datasets.

Fig. 5 presents qualitative comparison results on three benchmarks. Taking the image from the BRCA dataset as an example, the leftmost image shows the context region, where a cluster of densely packed, hyperchromatic nuclei is clearly visible in the lower left corner and displays a tumor invasion area. Under sliding window inference, without access to this broader contextual information, context-free nucleus detection models incorrectly classify many of the tumor nuclei as inflammatory or stromal types. Among context-aware approaches, our method exhibits better nucleus detection accuracy than MFoV-P2PNet. This can be attributed that compared to MFoV-P2PNet that uses LFoV patches to extract contextual information, our approach leverages surrounding window patches with higher magnification that provide more detailed contextual features.

Tab. 3 presents a comprehensive comparison of model

| CA | CL | ME | $F_{avg}$ |
|---|---|---|---|
| | | | 66.22 |
| ✓ | | | 70.79 |
| ✓ | | ✓ | 70.95 |
| ✓ | ✓ | | 71.23 |
| ✓ | ✓ | ✓ | 72.01 |

Table 4: Effect of our proposed modules.

| $\delta$ | MFoV-P2PNet | Ours |
|---|---|---|
| 0 | 66.22 | 66.22 |
| 1 | 66.69 | 72.01 |
| 2 | 66.83 | 72.07 |
| 3 | 66.95 | 72.28 |
| 4 | 66.98 | 72.51 |

Table 5: Effect of context size $\delta$.

| Method | $F_{avg}$ | $\Delta$ |
|---|---|---|
| baseline | 70.95 | - |
| SL | 71.10 | +0.15 |
| CL | 72.01 | +1.06 |

Table 6: Effect of pseudo-labeling strategies.

| Method | $F_{avg}$ |
|---|---|
| Add | 67.50 |
| Concat | 68.85 |
| CrossAttn | 72.01 |

Table 7: Effect of context integration strategies.

size, computational cost, inference efficiency and performance across different methods. All metrics are measured on a system with a single RTX 3090 GPU and dual AMD EPYC 7542 processors (2.90GHz, 64 cores, 128 threads). Compared to the baseline context-free nucleus detector (*i.e.*, P2PNet), our method introduces minimal additional inference time. Notably, the proposed model runs 2.36× faster than previous context-aware detector (*i.e.*, MFoV-P2PNet) as we eliminate the time-consuming step of additionally preparing LFoV patches.

## Ablation study

We evaluate the effect of our proposed modules and hyper-parameters on the BRCA dataset. Due to page limitation, we report only mean results across five runs.

**Effect of our proposed modules**. Tab. 4 presents the ablation results of context-aware learning (CA), cross-labeling (CL) and the incorporation of morphology-rich embedding (ME). It can be observed that all proposed modules contribute to improving the model performance.

**Effect of $\delta$.** Tab. 5 shows the impact of context size $\delta$ on model performance. $\delta = 0$ denotes the baseline P2PNet model. The results show dramatic performance gains when $\delta$ increases from 0 to 1, whereas further increments yield marginal improvements. This is also observed with previous context-aware method, MFoV-P2PNet. We hypothesize that the nearest neighboring patches provide the most relevant contextual information, while incorporating broader context regions inevitably introduces background noise that impedes the model from identifying informative contextual features.

**Comparison of pseudo-labeling strategies**. Tab. 6 shows the model performance with different pseudo-labeling strategies. The baseline represents the model with CA and ME modules. It can be observed that self-labeling (SL), which uses the detector itself to generate pseudo class labels for nuclei in surrounding patches, yields marginal performance gains. The proposed cross-labeling (CL) strategy that utilizes another model to produce pseudo labels shows notable improvement. Although our evaluations show that the auxiliary model does not exhibit better classification accuracy than the base detector (70.74 vs 70.79), it effectively mitigates the error accumulation problem inherent in the SL strategy.

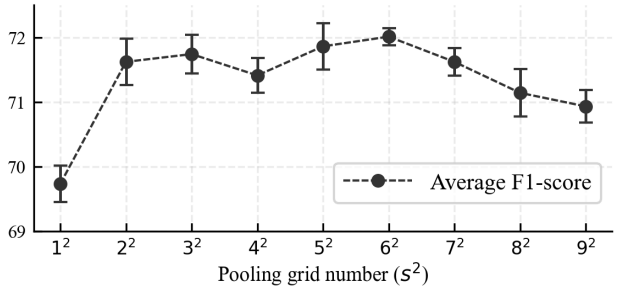


Figure 6: Effect of $s$. The error bars denote 95% confidence intervals across 5 independent runs.

**Comparison of context integration strategies**. Tab. 7 shows the effect of different context integration methods. The cross-attention (CrossAttn) operation achieves significantly better results than both addition (Add) and concatenation approaches (Concat).

**Effect of $s$**. Fig. 6 exhibits the impact of pooling grid number on model performance, with optimal results achieved at $s = 6$. Intuitively, increasing the grid number preserves more comprehensive contextual details and leads to better results. However, we observe performance degradation when $s$ exceeds 6. This can be attributed to the inherent spatial redundancy in high-level contextual feature maps (Li, Wen, and He 2023; Yang et al. 2025), where the feature points have already undergone sufficient information exchange in the earlier stages. This redundancy may impede the model in capturing informative contextual features and thus resulting in sub-optimal performance.

## Conclusion

In this paper, we propose an effective and efficient context-aware nucleus detection approach. Instead of using LFoV patches in previous methods, we aggregate contextual information from historically visited sliding windows during whole-slide inference, which significantly improves the detection efficiency and accuracy. Additionally, we propose a cross-labeling strategy to effectively leverage unlabeled nuclei in surrounding patches to enhance the model's context adaptability, and incorporate an auxiliary model to compensate for the loss of nuclear morphological details during contextual features integration. Extensive experiments on three benchmarks demonstrate the superiority of our method.

**Limitations and future work**. Compared to previous context-aware methods, our approach suffers from increased training time, as each iteration requires encoding a substantially larger number of patches. To address this limitation, we plan to explore feature caching strategies to accelerate model training in future work.

## Acknowledgements

This study was partially supported by "Pioneer" and "Leading Goose" R&D Program of Zhejiang (Grant 2025SDXHDX0003), the National Natural Science Foundation of China (Grant No.62506306), and foundation of Muyuan Laboratory (Program ID: 14106022401,14106022402). Furthermore, essential technical support was provided by the D-PathAI platform, including its hardware and software, which was developed by Hangzhou Dipath Technology Co., Ltd.

## 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.

Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N. E.; and McGuinness, K. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, 1–8. IEEE.

Bai, T.; Vu, Q. D.; 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.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, P.; Zhu, C.; Shui, Z.; Cai, J.; Zheng, S.; Zhang, S.; and Yang, L. 2023. Exploring unsupervised cell recognition with prior self-activation maps. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 559–568. Springer.

Chen, R. J.; and Krishnan, R. G. 2022. Self-supervised vision transformers learn visual concepts in histopathology. *arXiv preprint arXiv:2203.00585*.

Diao, J. A.; Wang, J. K.; Chui, W. F.; Mountain, V.; Gullapally, S. C.; Srinivasan, R.; Mitchell, R. N.; Glass, B.; Hoffman, S.; Rao, S. K.; et al. 2021. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nature communications*, 12(1): 1613.

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.; 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.

Hörst, F.; Rempe, M.; Heine, L.; Seibold, C.; Keyl, J.; Baldini, G.; Ugurel, S.; Siveke, J.; Grünwald, B.; Egger, J.; et al. 2024. Cellvit: Vision transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94: 103143.

Huang, J.; Li, H.; Sun, W.; Wan, X.; and Li, G. 2023a. Prompt-based grouping transformer for nucleus detection and classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 569–579. Springer.

Huang, J.; Li, H.; Wan, X.; and Li, G. 2023b. Affine-Consistent Transformer for Multi-Class Cell Nuclei Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21384–21393.

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.

Ignatov, A.; Yates, J.; and Boeva, V. 2024. Histopathological image classification with cell morphology aware deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6913–6925.

Kamnitsas, K.; Ledig, C.; Newcombe, V. F.; Simpson, J. P.; Kane, A. D.; Menon, D. K.; Rueckert, D.; and Glocker, B. 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis*, 36: 61–78.

Kirillov, A.; He, K.; Girshick, R.; Rother, C.; and Dollár, P. 2019. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9404–9413.

Li, H.; Shui, Z.; Zhang, Y.; Zhu, C.; and Yang, L. 2025. PathVQ: Reforming Computational Pathology Foundation Model for Whole Slide Image Analysis via Vector Quantization. *arXiv preprint arXiv:2503.06482*.

Li, J.; Wen, Y.; and He, L. 2023. Scconv: Spatial and channel reconstruction convolution for feature redundancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6153–6162.

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, Y.; Wang, Z.; Zhang, D.; Cheng, K.-T.; and Chen, H. 2024. BoNuS: boundary mining for nuclei segmentation with partial point labels. *IEEE Transactions on Medical Imaging*, 43(6): 2137–2147.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

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

Lou, W.; Li, G.; Wan, X.; and Li, H. 2024a. Cell graph transformer for nuclei classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3873–3881.

Lou, W.; Wan, X.; Li, G.; Lou, X.; Li, C.; Gao, F.; and Li, H. 2024b. Structure embedded nucleus classification for histopathology images. *IEEE Transactions on Medical Imaging*.

Millward, J.; He, Z.; and Nibali, A. 2023. Dense Prediction of Cell Centroids Using Tissue Context and Cell Refinement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 138–149. Springer.

Pan, X.; Cheng, J.; Hou, F.; Lan, R.; Lu, C.; Li, L.; Feng, Z.; Wang, H.; Liang, C.; Liu, Z.; et al. 2023. SMILE: Cost-sensitive multi-task learning for nuclear segmentation and classification with imbalanced annotations. *Medical Image Analysis*, 88: 102867.

Pina, O.; Dorca, E.; and Vilaplana, V. 2024. Cell-DETR: Efficient cell detection and classification in WSIs with transformers. In *Medical Imaging with Deep Learning*.

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.

Schoenpflug, L. A.; and Koelzer, V. H. 2023. SoftCTM: cell detection by soft instance segmentation and consideration of cell-tissue interaction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 109–122. Springer.

Shahamiri, N.; Rempe, M.; Heine, L.; Kleesiek, J.; and Hörst, F. 2025. Cracking the PUMA Challenge in 24 Hours with CellViT++ and nnU-Net. *arXiv preprint arXiv:2503.12269*.

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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 109–118. Springer.

Shui, Z.; Zhang, Y.; Yao, K.; Zhu, C.; Zheng, S.; Li, J.; Li, H.; Sun, Y.; Guo, R.; and Yang, L. 2024a. Unleashing the power of prompt-driven nucleus instance segmentation. In *European Conference on Computer Vision*, 288–304. Springer.

Shui, Z.; Zhao, Y.; Zheng, S.; Zhang, Y.; Li, H.; Zhang, S.; Yu, X.; Zhu, C.; and Yang, L. 2023. Semi-supervised Cell Recognition under Point Supervision. *arXiv preprint arXiv:2306.08240*.

Shui, Z.; Zheng, S.; Zhu, C.; Zhang, S.; Yu, X.; Li, H.; Li, J.; Chen, P.; and Yang, L. 2024b. DPA-P2PNet: deformable proposal-aware P2PNet for accurate point-based cell detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4864–4872.

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.

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.

Torbati, N.; Meshcheryakova, A.; Mechtcheriakova, D.; and Mahbod, A. 2025. A multi-stage auto-context deep learning framework for tissue and nuclei segmentation and classification in h&e-stained histological images of advanced melanoma. *arXiv preprint arXiv:2503.23958*.

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.

Weinstein, J. N.; Collisson, E. A.; Mills, G. B.; Shaw, K. R.; Ozenberger, B. A.; Ellrott, K.; Shmulevich, I.; Sander, C.; and Stuart, J. M. 2013. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10): 1113–1120.

Xu, M.; Gupta, S.; Hu, X.; Li, C.; Abousamra, S.; Samaras, D.; Prasanna, P.; and Chen, C. 2025a. TopoCellGen: Generating Histopathology Cell Topology with a Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Xu, Q.; Luo, Y.; Duan, W.; and Chen, Z. 2025b. Co-Seg++: Mutual Prompt-Guided Collaborative Learning for Versatile Medical Segmentation. *arXiv preprint arXiv:2506.17159*.

Yang, S.; Chen, Y.; Tian, Z.; Wang, C.; Li, J.; Yu, B.; and Jia, J. 2025. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19792–19802.

Yang, Z.; Qiu, Z.; Lin, T.; Chao, H.; Chang, W.; Yang, Y.; Zhang, Y.; Jiao, W.; Shen, Y.; Liu, W.; et al. 2024. From Histopathology Images to Cell Clouds: Learning Slide Representations with Hierarchical Cell Transformer. *arXiv preprint arXiv:2412.16715*.

Yao, K.; Huang, K.; Sun, J.; and Hussain, A. 2024. PointNu-Net: Keypoint-Assisted Convolutional Neural Network for Simultaneous Multi-Tissue Histology Nuclei Segmentation and Classification. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(1): 802–813.

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.