MUSE: Multi-Scale Dense Self-Distillation for Nucleus Detection and Classification

Zijiang Yang[1,2,*†], Hanqing Chao[2,3,5,*], Bokai Zhao[2,*], Yelin Yang[4], Yunshuo Zhang[4], Dongmei Fu[1], Junping Zhang[5], Le Lu[2], Ke Yan[2,3], Dakai Jin[2], Minfeng Xu[2], Yun Bian[4], Hui Jiang[4]

[1]University of Science and Technology Beijing
[2]DAMO Academy, Alibaba Group
[3]Hupan Lab
[4]Shanghai Institution of Pancreatic Disease
[5]Fudan University
{hanqing.chq, eric.xmf}@alibaba-inc.com, bianyun2012@foxmail.com, jianghui5131@163.com

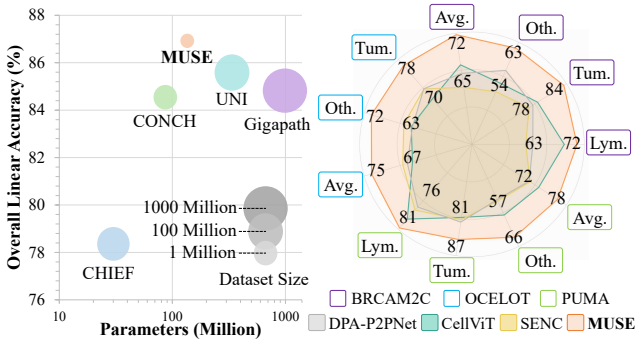
Abstract

Nucleus detection and classification (NDC) in histopathology analysis is a fundamental task that underpins a wide range of high-level pathology applications. However, existing methods heavily rely on labor-intensive nucleus-level annotations and struggle to fully exploit large-scale unlabeled data for learning discriminative nucleus representations. In this work, we propose **MUSE** (**MU**lti-scale den**SE** self-distillation), a novel self-supervised learning method tailored for NDC. At its core is **NuLo** (**Nu**cleus-based **Lo**cal self-distillation), a coordinate-guided mechanism that enables flexible local self-distillation based on predicted nucleus positions. By removing the need for strict spatial alignment between augmented views, NuLo allows critical cross-scale alignment, thus unlocking the capacity of models for fine-grained nucleus-level representation. To support MUSE, we design a simple yet effective encoder-decoder architecture and a large field-of-view semi-supervised fine-tuning strategy that together maximize the value of unlabeled pathology images. Extensive experiments on three widely used benchmarks demonstrate that MUSE effectively addresses the core challenges of histopathological NDC. The resulting models not only surpass state-of-the-art supervised baselines but also outperform generic pathology foundation models.

Code — https://github.com/alibaba-damo-academy/MUSE
Extended version — https://arxiv.org/abs/2511.05170

Introduction

Nucleus detection and classification (NDC) is a foundational task in histopathological diagnosis (Page et al. 2023; Zhang et al. 2025). Core pathological workflows, including disease diagnosis, biomarker evaluation, and prognosis prediction, critically depend on the precise localization and identification of specific types of nuclei (Wang et al. 2023; Corredor et al. 2019). In histopathology, accurate recognition of nucleus types relies on both nuclear morphology and the structural context of the surrounding tissue. However, manual annotation for NDC is extremely labor-intensive and time-consuming. To reduce annotation costs while ensuring representativeness, most existing datasets annotate only small high-magnification tiles (typically around 128 μm per side, containing dozens to hundreds of nuclei). Despite this compromise, these annotated samples remain insufficient to capture the full variability in tissue architecture, nuclear morphology, and staining conditions, limiting the generalizability of the supervised models (Graham et al. 2019; Hörst



(a) Classification (ACC %) (b) Detection and Classification (F1)

Figure 1: Comparison of MUSE and SOTA methods. (a) Compared with pathology pretraining methods, MUSE achieves better nucleus classification performance with smaller backbones and only 0.5 million samples. (b) After fine-tuning, MUSE outperforms SOTA methods on nucleus detection and classification. Lym., Tum., Oth., and Avg. denote the F1 score of lymphocytes, tumor nucleus, other nucleus, and average, respectively.

*These authors contributed equally.
†This work was done when Zijiang Yang conducted an internship at DAMO Academy, Alibaba Group.

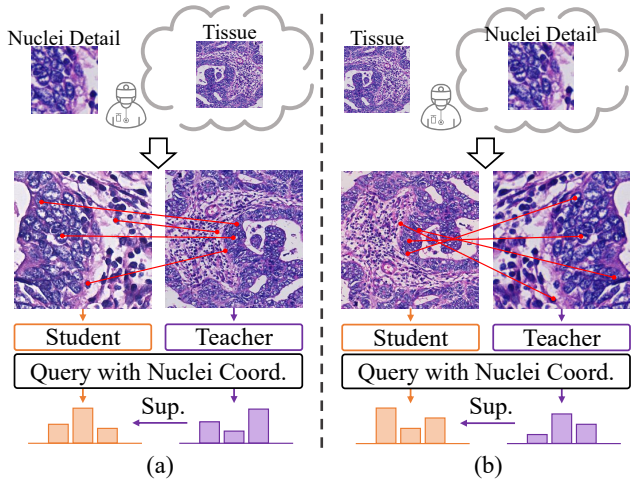


Figure 2: Motivation of NuLo. Pathologists can flexibly associate nuclei with tissue. This inspires us to introduce two cross-scale self-distillation processes on matched nuclei: (a) inferring tissue information with nuclei detail and (b) inferring nuclei detail with tissue information.

et al. 2024). Furthermore, the inherently limited field of view (FoV) in such small tiles restricts access to broader tissue-level context, further contributing to performance bottlenecks (Chai et al. 2025; Hörst et al. 2025).

To alleviate these limitations, recent studies have explored the integration of unlabeled data. Some approaches incorporate additional low-magnification images with larger FoVs to provide richer contextual information (Shui et al. 2024), while others employ semi-supervised learning techniques such as pseudo-labeling and consistency regularization (Su et al. 2023; Bai et al. 2021). However, these methods still depend heavily on the limited set of labeled data. Large-FoV (LFoV) inputs enrich contextual information per tile but do not expand the diversity or quantity of training samples. Meanwhile, semi-supervised learning methods typically assume that labeled and unlabeled data are drawn from the same distribution (Yang et al. 2022), which constrains the inclusion of diverse unlabeled samples and often leads to unstable performance when this assumption is violated. Recently, self-supervised learning (SSL) has emerged as a powerful paradigm for learning generalizable visual representations from large-scale unlabeled data (Zhou et al. 2021; Oquab et al. 2023). In computational pathology, SSL-pretrained foundation models have demonstrated promising improvements across various downstream tasks (Chen et al. 2024; Wang et al. 2024), offering a potential path forward for addressing the dilemma faced by NDC. However, most existing pathology foundation models directly adopt SSL methods originally developed for natural images (e.g., DINOv2 (Oquab et al. 2023)), primarily focusing on image-level representation learning without adapting to the local nucleus-level demands of dense prediction tasks such as NDC. While some of these models are trained on millions (Chen et al. 2024) or even billions (Xu et al. 2024) of histology tiles, they often struggle to capture discriminative nucleus-level features, limiting their effectiveness for NDC (Figure 1a). This limitation arises from three key issues. First, existing SSL methods typically enhance local representations through patch-level masked image modeling (e.g., iBOT (Zhou et al. 2021)). However, these methods require strict spatial alignment between patch tokens, limiting essential spatial augmentations such as scale jittering, which is crucial for learning local-to-global alignment (Caron et al. 2021). Second, current SSL models typically lack supervision across different feature levels. As NDC is a dense prediction task, existing studies have shows that it benefits from multi-level feature fusion (Shui et al. 2024). Third, most foundation models operate on small pathology tiles with a limited FoV (e.g., 256×256 pixels, ≈16.4K μm^2), which restricts their capacity to capture broader tissue context.

In this work, we propose MUSE (**MU**lti-scale den**SE** self-distillation), a novel SSL method tailored for NDC. Built around this pretraining strategy, we develop a simple yet effective framework encompassing a unified encoder-decoder model architecture, pretraining strategy, and downstream fine-tuning pipeline that collectively enable efficient utilization of both annotated and unannotated data. We first introduce a lightweight and flexible encoder-decoder backbone that supports variable input sizes (from 96 to 1024 pixels), enabling large-FoV training and multi-level feature fusion. Based on this, MUSE incorporates a novel Nucleus-based Local Self-distillation (NuLo) mechanism, which enhances global SSL with flexible local self-distillation. NuLo employs a lightweight nucleus detector to estimate nuclear coordinates and performs local self-distillation based on feature interpolation around each nucleus. This coordinate-guided local self-distillation removes the need for strict spatial alignment between augmented views, enabling spatial transformations, including flipping and scale changes, which are essential for NDC. This design also allows the model to learn cross-scale alignment, akin to how pathologists reason about the relationship between individual nuclei and their tissue context (Figure 2). To further enhance contextual learning, MUSE progressively expands the pretraining FoV up to 65.5K μm^2 (512×512 pixels). During fine-tuning, we propose a simple yet effective large-FoV semi-supervised strategy: labeled patches are expanded to 262.1K μm^2 (1024×1024 pixels), where supervised learning is applied to annotated regions, and pseudo-labeling is applied to surrounding unlabeled areas. Overall, the proposed MUSE framework effectively addresses key challenges in histopathological NDC. Extensive experiments on three widely used benchmarks demonstrate that MUSE-pretrained models not only significantly surpass existing supervised NDC methods but also outperform generic pathology foundation models even with smaller models and fewer samples. Our contributions are summarized as follows:

- We propose MUSE, a novel SSL approach specifically tailored for nucleus detection and classification. By introducing the novel Nucleus-based Local Self-distillation (NuLo) mechanism, MUSE enables flexible multi-scale local self-distillation. Combined with large-FoV pretraining, it allows the model to capture discriminative nu-

cleus features.

- Built around MUSE, we develop a complete framework that includes an encoder-decoder architecture, a pretraining strategy, and a downstream fine-tuning pipeline, allowing effective utilization of unlabeled data and LFoV across all training stages.
- Extensive experiments demonstrate that our MUSE framework effectively addresses the key challenges of histopathological NDC. The resulting pretrained models not only surpass state-of-the-art supervised methods but also outperform generic pathology foundation models.

Related Work

Nucleus Detection and Classification

Current methods for NDC can be broadly categorized into map-based (Graham et al. 2019; Yao et al. 2023; Pan et al. 2023; Hörst et al. 2024, 2025) and point-based methods (Abousamra et al. 2021; Ryu et al. 2023; Shui et al. 2024). Despite extensive research on model designs, these methods still follow the supervised learning paradigm, which relies on large-scale, fully annotated nucleus-level datasets. On the other hand, semi-supervised learning has been explored to leverage unlabeled patches for better performance (Bai et al. 2021; Su et al. 2023). However, these methods typically rely on strong assumptions such as the cluster assumption (Yang et al. 2022), which limit the usage of large-scale heterogeneous pathology patches. In this paper, we propose a novel SSL framework to leverage abundant unlabeled patches for better performance on NDC.

Pathology Pretraining

Self-supervised learning is an efficient strategy for learning generalizable representations from large-scale unlabeled data (Misra and Maaten 2020; Chen, Xie, and He 2021; Zhou et al. 2021; He et al. 2022; Caron et al. 2021; Assran et al. 2023; Oquab et al. 2023). Recently, pathological pretraining methods have been extensively explored (Wang et al. 2022a; Azizi et al. 2023; Wang et al. 2022b; Chen et al. 2024; Xu et al. 2024; Wang et al. 2024). After being pretrained on large-scale unlabeled pathology patches, these methods substantially advance WSI-level tasks (Wang et al. 2024). In this work, we experimentally shows that these methods still exhibit limited performance for nucleus representation. To address this issue, we propose NuLo to achieve better local self-distillation based on matched nuclei.

Preliminaries

Nucleus Detection and Classification

Nucleus detection and classification is defined as follows:

$$\hat{C} = M(I), \hat{C} = \{[\hat{t^{(i)}}, \hat{x^{(i)}}, \hat{y^{(i)}}]^T\}_{i=1}^{N_p} \quad (1)$$

where I is the input pathology patch, M denotes the model, N_p is the total number of predicted nuclei, $\hat{C}$ is the predicted set, $\hat{t^{(i)}}$, $\hat{x^{(i)}}$, and $\hat{y^{(i)}}$ are the predicted type, the coordinate of the x-axis and the y-axis of the i-th predicted nucleus, respectively. In this work, we further define a simplified task, termed nucleus classification, to focus on evaluating nucleus representation performance:

$$\{\hat{t^{(i)}}\}_{i=1}^{N_g} = M(I; \{[x^{(i)}, y^{(i)}]^T\}_{i=1}^{N_g}), \quad (2)$$

where N_g is the total number of ground truth nuclei, $x^{(i)}$ and $y^{(i)}$ are the ground truth x-axis and the y-axis coordinates of the i-th nucleus, respectively.

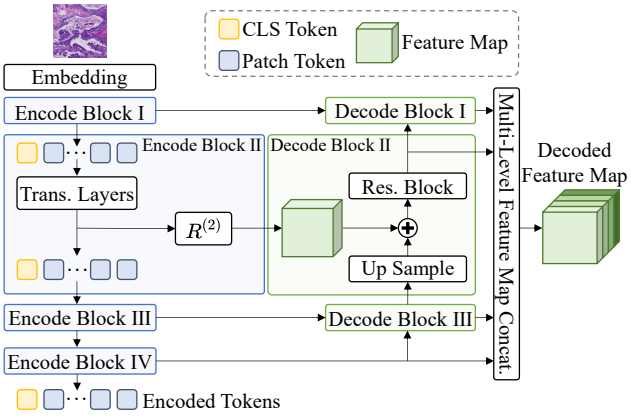


Figure 3: Illustration of the architecture. Given a patch, this architecture produces both image-level representation (CLS token) and high-resolution dense representations for multi-scale self-distillation. $R^{(2)}$ denotes the reassembly layer of the second encoder block.

Self-Distillation with No Labels

Self-Distillation with No Labels (DINO) (Caron et al. 2021) employs a self-supervised teacher-student architecture to learn visual representations without labels. The student network M_s and teacher network M_t share identical architectures but maintain separate parameters. Given two augmented views, denoted as $I^{(1)}$ and $I^{(2)}$, of an input image, the output of M_s is required to match the output of M_t:

$$\mathcal{L}_{image} = -P_t^{(cls)}(M_t(I^{(1)})) \log P_s^{(cls)}(M_s(I^{(2)})), \quad (3)$$

where $P_t^{(cls)}$ and $P_s^{(cls)}$ denote the operator to compute output probability of M_t and M_s, respectively, $\mathcal{L}_{image}$ is the cross-entropy loss at the image level. In practice, the multi-crop strategy is applied to generate a set of views $V = \{I_g^{(1)}, I_g^{(2)}, I_l^{(i)}\}_{i=1}^{N_l}$ from an image to build a group of paired samples and encourage local-to-global learning, where N_l is the number of local views, I_g denote global view with large resolution, and I_l denote local view with smaller resolution. M_s are updated with stochastic gradient descent to minimize $\mathcal{L}_{image}$. M_t are initialized as the same of M_s and updated with an Exponential Moving Average (EMA) of M_s. In this work, we mainly follow the framework of DINO and further introduce nucleus-level self-distillation to improve the representation of nuclei.

Method

As shown in Figure 3, we introduce a lightweight and flexible encoder-decoder backbone to extract features. The

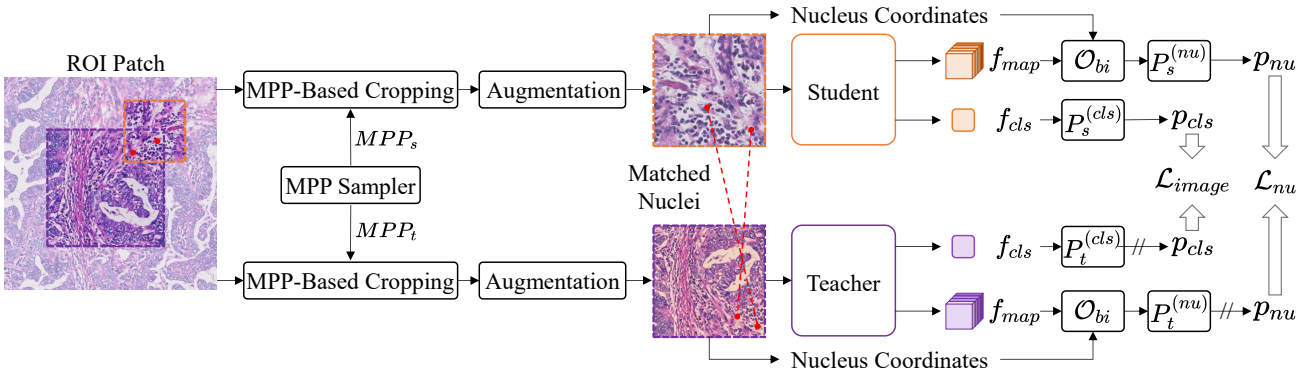


Figure 4: Illustration of MUSE. MPP-Based Cropping is first employed to generate paired views based on the ROI patch and random MPP. After data augmentation, we extract image-level representations (CLS tokens, f_{cls}) and dense representations (feature maps, f_{map}) of the paired views with teacher and student networks. MUSE minimizes two losses: 1) image-level self-distillation between CLS tokens and 2) nucleus-level self-distillation between features of matched nuclei. Specifically, nucleus features are interpolated from the feature maps based on their coordinates.

framework of MUSE is illustrated in Figure 4. First, MPP-based cropping is employed to obtain multi-scale paired views. Second, multi-level representations of views are extracted with teacher and student networks. Third, these representations are further utilized for image-level and nucleus-level self-distillation. For downstream NDC tasks, the pre-trained model is applied to the specific dataset with a large-FoV semi-supervised fine-tuning pipeline.

Architecture

The encoder-decoder framework has been extensively validated for NDC (Graham et al. 2019; Shui et al. 2024). In this work, an encoder-decoder backbone based on Vision Transformer (ViT) (Dosovitskiy et al. 2020) is constructed.

The encoder is composed of a ViT and reassembly layers. Specifically, we first extract multi-level encoded features $F_e = \{f_e^{(i)}\}_i^{N_e}$ from N_e layers, where $f_e^{(i)}$ is the encoded feature from the i-th layer. Subsequently, the reassembly layers $R(\cdot)$ are used to convert F_e into the 2D feature maps required by the decoder: $F'_e = \{R^{(i)}(f_e^{(i)})\}_{i=1}^{N_e}$. For ViT, $f_e^{(i)} \in \mathbb{R}^{(HW/p^2+1)\times c^{(i)}}$, where p denotes the patch size, c denotes the dimension of each token, H denotes the height of the input image, and W denotes the width of the input image. The reassembly layer $R^{(i)}(\cdot)$ discards the CLS token, reassembles the sequence of tokens into a 2D feature map $\in \mathbb{R}^{c^{(i)}\times(H/p)\times(W/p)}$, adjusts the feature map to the target channel dimension via a 1×1 convolution, and finally resamples it to the target spatial size.

The decoder comprises a series of residual-block-based modules that progressively fuse feature maps from the encoder. Let $f_d^{(i)}$ denote the i-th decoded feature map, and let the full set of decoder outputs be $F_d = \{f_d^{(i)}\}_{i=1}^{N_e}$. Decoded feature maps are further mapped to the target channel dimension and spatial size, and then concatenated to form a unified feature map $f_{map} = Cat(F_d)$, which serves as the dense representation of the input image. In addition, the CLS token output of the encoder, denoted as f_{cls}, is used as the image-level representation.

Following common practice in the encoder-decoder framework, N_e is set to 4. F'_e is constructed from equally spaced transformer layers (Ranftl, Bochkovskiy, and Koltun 2021). This design serves as the default backbone configuration unless otherwise specified.

MUSE

ROI Patches. WSIs typically exhibit gigapixel-scale dimensions, presenting challenges for constructing effective image pairs with partial overlap. To address this and generate suitable training samples, we introduce a sequential dual-cropping strategy: 1) Region of Interest (ROI) cropping and 2) multi-scale view cropping. An initial crop operation is applied to the source WSI to isolate a relevant ROI. A subsequent crop operation is performed on the extracted ROI to yield training samples. In this work, a dataset comprising 483,627 ROI patches based on the Cancer Genome Atlas Program (TCGA) (Liu et al. 2018) is constructed, termed TCGA_{nu}. Nuclei coordinates are auto-detected. Please refer to the extended version in arXiv for details.

MPP-Based Cropping. The physical scale of pathology images is critical for effective self-distillation in MUSE. To generate views of specified physical resolution, we propose a novel cropping method based on Microns-Per-Pixel (MPP). The cropping operator $\mathcal{O}_{crop}$ is formally defined as:

$$\{I_o, C_o\} = \mathcal{O}_{crop}(I_e, C_e; MPP_e, MPP_o, R_o), \quad (4)$$

where I_e is the input image, C_e is the input nucleus coordinates, I_o is the output image, C_o is the output nucleus coordinates, MPP_e is the input MPP, MPP_o is the output MPP, and R_o is the the pixels resolution of I_o. MPP_o is random generated by MPP Sampler. $\mathcal{O}_{crop}$ crops a region of size $MPP_o/MPP_e \times R_o$ from I_e, then resizes it to R_o.

After cropping, C_o is aligned to the local coordinate system of I_o, which makes it difficult to match cells in multiple output images from the same I_s. We further introduce

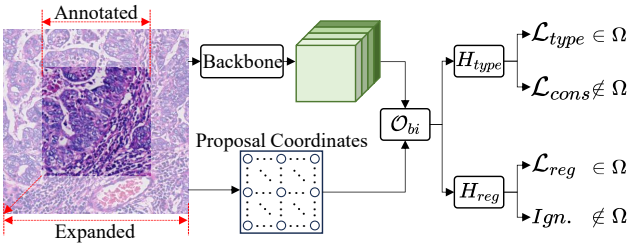


Figure 5: Illustration of the fine-tuning. Annotated samples are first extended to new samples with a larger field of view. The pretrained backbone is then used to extract feature maps, which are further utilized to obtain features at proposal coordinates. $Ign.$ denotes that the coordinate regression is not applied to unlabeled regions.

a coordinate-agnostic indexing scheme to uniquely identify each nucleus in I_e for cross-view nucleus matching:

$$\{I_o, C_o, K_o\} = \mathcal{O}_{crop}(I_e, C_e, \mathcal{O}_{in}(C_e)); \cdots), \quad (5)$$

where $\mathcal{O}_{in}$ is the nucleus index construction operator, K_o is the nucleus index list of C_o, and $\cdots$ denotes $[MPP_e, MPP_o, R_o]^T$. For any two samples $\{I_o^{(1)}, C_o^{(1)}, K_o^{(1)}\}$ and $\{I_o^{(2)}, C_o^{(2)}, K_o^{(2)}\}$ derived from I_e, the nucleus-level match is efficiently established via the intersection of $K_o^{(1)}$ and $K_o^{(2)}$.

Nucleus-Based Local Self-Distillation (NuLo). We introduce self-distillation at the nucleus level to encourage the model to distinguish nuclei in pathological images and to learn stable representations of nuclei across scales. For any sample $\{I_o, C_o, K_o\}$, f_{map} is extracted with the encoder-decoder architecture. Each nucleus feature is then obtained via bilinear interpolation from f_{map} based on coordinates: $f_c = \mathcal{O}_{bi}(f_{uni}; [x_c, y_c]^T)$, where f_c denotes the feature vector of the nucleus, $\mathcal{O}_{bi}$ is the bilinear interpolation operator, and $[x_c, y_c]^T$ is the nucleus coordinates. The set of nucleus features is denoted as $F_c = \{f_c\}^{C_o}$. Given two views $\{I_o^{(1)}, C_o^{(1)}, K_o^{(1)}\}$ and $\{I_o^{(2)}, C_o^{(2)}, K_o^{(2)}\}$ derived from the same ROI patch, the corresponding nucleus features $F_c^{(1)}$ and $F_c^{(2)}$ are extracted through teacher and student networks, respectively. Nucleus-level self-distillation loss ($\mathcal{L}_{nu}$) of each paired views is further defined as follows:

$$\mathcal{L}_{nu} = \sum_{K_{cap}} -P_t^{(nu)}(F_c^{(1)}[K_{cap}]) \log P_s^{(nu)}(F_c^{(2)}[K_{cap}]),$$
$$s.t.\ K_{cap} = K_o^{(1)} \cap K_o^{(2)}, \quad (6)$$

where $P_t^{(nu)}$ and $P_s^{(nu)}$ denote the operator to compute output probability of $F_c^{(1)}$ and $F_c^{(2)}$, $[\cdot]$ is the query operator based on index, K_{cap} is the intersection set of $K_o^{(1)}$ and $K_o^{(2)}$. Equation (6) enables the student network to distinguish nuclei with morphological differences and learn cross-scale consistent representations of nuclei.

Optimization. MUSE adopts the teacher-student network update method of DINO as described in Preliminaries. Student network is optimized with both image-level ($\mathcal{L}_{image}$) and nucleus-level ($\mathcal{L}_{nu}$) self-distillation losses. $\mathcal{L}_{image}$ is applied to the CLS token outputs of encoders to preserve global image representation learning. Meanwhile, $\mathcal{L}_{nu}$ operates on the decoder outputs to enhance representation of nuclei. The composite loss $\mathcal{L}_{MUSE}$ is further defined as:

$$\mathcal{L}_{MUSE} = \lambda_{image}\mathcal{L}_{image} + \lambda_{nu}\mathcal{L}_{nu}, \quad (7)$$

where λ_{image} and λ_{nu} are the loss weights of $\mathcal{L}_{image}$ and $\mathcal{L}_{nu}$, respectively. λ_{image} and λ_{nu} are set to 1 in experiments, unless otherwise mentioned.

LFoV Pretraining. Following the common practice of DINO, MUSE sets the pixel resolutions of global views and local views to 224 and 96, respectively. In addition, we further extend the pixel resolutions of global and local views to 512 and 208 to explore the impact of incorporating more tissue context on nucleus representation.

Downstream Fine-Tuning

To transfer the MUSE-pretrained model to NDC tasks, we propose a novel fine-tuning pipeline (Figure 5). We first naturally expand the small FOV samples to include more tissue context, and then employ the point-based method to regress nucleus coordinates and predict nucleus types based on independent heads H_{reg} and H_{type}, respectively.

LFoV Samples. While the framework can be generalized to arbitrary region shapes, we primarily discuss the square-shaped annotated regions, which align with the predominant structure of most existing datasets (Schuiveling et al. 2025; Ryu et al. 2023; Abousamra et al. 2021). Training samples, comprising cropped WSI regions with corresponding nucleus annotations, are located by their top-left corner coordinates $[x_a, y_a]^T$ and side length of the annotation region r_a. Furthermore, the LFoV samples are generated by extending these annotated regions to incorporate surrounding unlabeled tissue areas:

$$x'_a = x_a - \mathcal{O}_{sample}(0, r'_a - r_a),$$
$$y'_a = y_a - \mathcal{O}_{sample}(0, r'_a - r_a),$$
$$s.t.\ r'_a \geq r_a, \quad (8)$$

where $\mathcal{O}_{sample}$ is a random sampling operator, and $[x'_a, y'_a]^T$ and r'_a are the top-left corner coordinates and side length of the LFoV sample, respectively. In Equation (8), region offsets are randomly sampled from the range $[0, r'_a - r_a]$. These samples permit the annotated region to appear at any position within the sample and provide expanded contextual tissue information beyond the annotation region.

Semi-Supervised Fine-Tuning. For LFoV samples containing both annotated and unannotated regions, we naturally introduce a semi-supervised fine-tuning: 1) the coordinate regression loss $\mathcal{L}_{reg}$ and classification loss $\mathcal{L}_{type}$ of the annotated region, and 2) the consistency prediction regularization term $\mathcal{L}_{cons}$ of the unannotated region:

$$\mathcal{L}_{ft} = \lambda_{reg}\mathcal{L}_{reg} + \lambda_{type}\mathcal{L}_{type} + \lambda_{cons}\mathcal{L}_{cons}, \quad (9)$$

where λ_{reg}, λ_{cls}, and λ_{cons} are loss weights, and $\mathcal{L}_{ft}$ is the total loss for fine-tuning. Specifically, we perform nucleus detection and classification on the entire LFoV sample, then split all predictions into two groups based on the

Method	Arch.	Params.	Dataset	N_s	Evaluation on 20x			Evaluation on 40x			Overall
					BRCA.	OCEL.	PUMA	BRCA.	OCEL.	PUMA	
Pretrained on Large-Scale General Datasets											
DINO (Caron et al. 2021)	ResNet-50	23M	IN-1k	1M	73.94	75.47	72.76	72.22	75.75	76.84	74.50
DINO (Caron et al. 2021)	ViT-S/16	21M	IN-1k	1M	77.85	81.76	78.36	82.33	80.37	78.80	79.91
DINO (Caron et al. 2021)	ViT-B/16	85M	IN-1k	1M	77.20	80.61	78.76	80.12	81.37	80.36	79.74
MAE (He et al. 2022)	ViT-B/16	85M	IN-1k	1M	76.78	78.46	77.83	77.56	81.38	77.27	78.21
iBOT (Zhou et al. 2021)	ViT-S/16	21M	IN-22k	14M	80.33	80.40	77.82	81.34	81.45	80.73	80.34
iBOT (Zhou et al. 2021)	ViT-B/16	85M	IN-22k	14M	79.57	82.48	77.93	82.48	82.17	79.07	80.62
DINOV2 (Oquab et al. 2023)	ViT-S/14	21M	LVD-142M	142M	82.39	80.13	79.66	81.72	79.13	81.55	80.76
DINOV2 (Oquab et al. 2023)	ViT-B/14	86M	LVD-142M	142M	84.39	81.66	80.66	85.02	79.37	82.93	82.34
Pretrained on Pathology Patches											
MoCoV2 (Kang et al. 2023)	ResNet-50	24M	TCGA	19M	80.71	82.17	79.60	79.27	82.20	79.90	80.64
DINO (Kang et al. 2023)	ViT-S/16	22M	TCGA	19M	83.63	85.30	81.92	84.30	84.44	81.13	83.45
DINOV2 (Oquab et al. 2023)	ViT-S/16	21M	TCGA_{nu}	484K	81.71	83.65	79.08	83.34	83.04	79.13	81.66
DINOV2 (Oquab et al. 2023)	ViT-B/16	86M	TCGA_{nu}	484K	83.60	82.87	79.00	84.25	82.71	79.10	81.92
CHIEF (Wang et al. 2024)	Swin-T	28M	custom*	15M	79.52	82.77	77.69	79.77	81.42	76.11	79.55
CTransPath (Wang et al. 2022b)	Swin-T	28M	custom*	15M	80.12	82.92	77.83	79.87	81.07	78.15	79.99
CONCH (Lu et al. 2024)	ViT-B/16	86M	custom*	$1\text{M}_{(\text{VL})}$	86.09	87.12	82.52	87.93	86.32	84.80	85.80
UNI (Chen et al. 2024)	ViT-L/16	300M	Mass-100k	100M	87.12	87.72	83.04	88.95	87.26	<u>84.84</u>	86.49
Prov-GigaPath (Xu et al. 2024)	ViT-G/14	1.1B	custom*	1.3B	86.99	**88.13**	83.46	88.00	86.03	83.04	85.94
MUSE (ours)	ResNet-50	86M	TCGA_{nu}	484K	86.29	86.30	81.69	88.26	84.85	80.42	84.64
	ViT-S/16	31M	TCGA_{nu}	484K	86.40	86.21	83.09	88.56	86.40	80.79	85.24
	ViT-B/16	123M	TCGA_{nu}	484K	88.43	86.03	84.18	89.60	86.87	82.46	86.26
LFoV-MUSE (ours)	ResNet-50	86M	TCGA_{nu}	484K	<u>88.70</u>	<u>87.87</u>	<u>84.49</u>	<u>89.74</u>	85.17	82.62	<u>86.43</u>
	ViT-S/16	31M	TCGA_{nu}	484K	86.29	87.54	83.81	86.59	**88.01**	84.56	86.13
	ViT-B/16	123M	TCGA_{nu}	484K	**89.29**	87.05	**84.84**	**90.26**	<u>87.87</u>	**85.74**	**87.51**

Table 1: Comparison of nucleus classification in ACC % (↑) for fine-tuning (FT). MUSE outperforms other pretraining methods in overall evaluations. The best results are highlighted in bold, and the second-best results are in underlined. N_s is the number of samples in each dataset. custom* denotes mixed dataset. BRCA. and OCEL. denote BRCAM2C and OCELOT, respectively.

annotated region Ω. Following common practice (Shui et al. 2024), Mean Squared Error (MSE) and cross-entropy losses are employed for regression and classification, respectively. Both $\mathcal{L}_{reg}$ and $\mathcal{L}_{type}$ are computed by ground truth annotations and corresponding predictions strictly within Ω. For unlabeled area, we first generate pseudo-labels with predicted probabilities, and then filter proposal points with prediction confidence above a specified threshold for $\mathcal{L}_{cons}$.

Experiments

Experiment Settings

Dataset & Metrics. BRCAM2C (Abousamra et al. 2021), OCELOT (Ryu et al. 2023), and PUMA (Schuiveling et al. 2025) are used to evaluate the performance of models on various tissues. Following common practice (Caron et al. 2021; Oquab et al. 2023) for evaluating pretrained models, we report Accuracy (ACC) of K-Nearest Neighbors (KNN), linear probing (LIN), and end-to-end fine-tuning (FT) for the nucleus classification task. For NDC, we follow other SOTA methods (Shui et al. 2024) to evaluate models with F1 score.

Baselines. We compare MUSE against strong SSL baselines pre-trained on large-scale general datasets, including iBOT (Zhou et al. 2021), MAE (He et al. 2022), DINO (Oquab et al. 2023), and DINOv2 (Oquab et al. 2023). Furthermore, we compare MUSE with the SOTA pathology foundation models, including PathBench (Kang et al. 2023), CHIEF (Wang et al. 2024), CTransPath (Wang et al. 2022b), CONCH (Lu et al. 2024), Prov-GigaPath (Xu et al. 2024), and UNI (Chen et al. 2024). Besides, we also compare fine-tuned MUSE-pretrained models with SOTA NDC methods, including MCSpatNet (Abousamra et al. 2021), PointNu-Net (Yao et al. 2023), CellViT (Hörst et al. 2024), SMILE (Pan et al. 2023), SENC (Lou et al. 2024b), CGT (Lou et al. 2024a), and DPA-P2PNet (Shui et al. 2024).

Implementation. Please refer to the extended version in arXiv for detailed hyper-parameters.

Main Results

Nucleus Classification. Table 1 and Table 2 report the results of nucleus classification on multi-tissue and multi-magnification datasets. Pathology pretraining significantly improves nucleus representation performance compared to models pretrained on general datasets. Furthermore, MUSE achieves better data efficiency and overall nucleus classification performance compared to existing methods. Specifically, ViT-B pre-trained with MUSE outperforms CONCH, which has the same encoder, in ACC % by 1.4 in LIN and 0.5 in FT. As CONCH uses over 1 million image-text pairs (versus about 0.5 million patches for MUSE), CONCH shows a

Method	Arch.	Dataset	BRCA. (20x)		OCEL. (20x)		PUMA (20x)		BRCA. (40x)		OCEL. (40x)		PUMA (40x)		Overall	
			KNN	LIN	KNN	LIN	KNN	LIN	KNN	LIN	KNN	LIN	KNN	LIN	KNN	LIN
Pretrained on Large-Scale General Datasets																
DINO	ResNet-50	IN-1k	76.29	75.91	73.00	78.39	74.04	75.14	75.37	72.06	70.72	73.23	74.06	76.71	73.91	75.24
DINO	ViT-S/16	IN-1k	77.25	69.86	71.94	74.52	70.22	73.42	77.78	73.89	70.72	73.88	71.29	75.19	73.20	73.46
DINO	ViT-B/16	IN-1k	78.03	77.09	73.11	76.83	71.40	76.80	76.28	74.59	72.30	76.92	72.86	78.43	74.00	76.78
MAE	ViT-B/16	IN-1k	65.45	73.54	67.04	77.34	61.79	76.16	67.45	71.92	66.18	76.17	64.14	76.70	65.34	75.31
iBOT	ViT-S/16	IN-22k	77.67	78.61	74.70	78.01	71.32	76.59	78.00	78.44	72.42	76.58	71.18	76.57	74.21	77.47
iBOT	ViT-B/16	IN-22k	79.06	76.20	76.66	77.54	71.09	76.35	80.63	79.97	74.86	78.85	73.34	78.21	75.94	77.85
DINOV2	ViT-S/14	LVD-142M	80.27	78.49	76.77	76.79	69.55	75.93	78.80	74.95	77.35	78.37	73.26	78.06	76.00	77.10
DINOV2	ViT-B/14	LVD-142M	79.11	78.96	75.25	78.22	68.81	76.09	80.12	74.68	76.06	79.28	73.50	79.45	75.48	77.78
Pretrained on Pathology Patches																
MoCoV2	ResNet-50	TCGA	79.37	81.90	76.83	79.31	74.52	78.92	78.23	80.65	78.03	81.46	75.48	79.55	77.08	80.30
DINO	ViT-S/16	TCGA	84.31	80.85	82.06	83.61	76.69	79.89	80.55	82.15	78.42	82.58	76.20	80.61	79.70	81.61
DINOV2	ViT-S/16	$TCGA_{nu}$	78.56	80.61	75.44	82.17	71.94	77.53	77.08	81.85	73.98	80.87	71.74	77.85	74.79	80.15
DINOV2	ViT-B/16	$TCGA_{nu}$	77.90	82.21	75.22	83.25	72.06	78.26	77.73	82.35	74.13	81.98	71.17	78.88	74.70	81.16
CHIEF	Swin-T	custom*	78.39	78.89	80.21	81.88	74.17	76.90	75.71	78.05	73.76	78.13	73.73	75.49	75.99	78.22
CTransPath	Swin-T	custom*	80.39	78.80	80.07	81.47	74.89	76.56	77.73	77.43	73.77	77.52	73.62	76.13	76.75	77.98
CONCH	ViT-B/16	custom*	86.68	85.13	**88.08**	86.41	81.95	82.02	86.71	86.20	**86.00**	83.80	83.05	83.66	85.41	84.54
UNI	ViT-L/16	Mass-100k	87.65	86.99	87.35	86.17	**82.21**	82.37	88.82	88.64	85.90	85.72	81.95	83.49	85.65	85.56
Prov-GigaPath	ViT-G/14	custom*	86.44	85.99	86.99	**87.50**	80.24	81.79	86.49	87.66	83.59	83.80	79.90	81.83	83.94	84.76
MUSE	ResNet-50	$TCGA_{nu}$	88.37	88.14	85.51	85.57	81.21	81.53	85.78	87.39	83.49	83.65	78.60	80.64	83.82	84.49
	ViT-S/16	$TCGA_{nu}$	86.88	87.79	86.13	85.42	80.00	81.34	87.67	89.66	85.45	85.20	79.71	80.17	84.31	84.93
	ViT-B/16	$TCGA_{nu}$	87.56	89.60	85.90	85.82	81.26	83.29	88.11	88.86	85.55	85.57	81.19	82.48	84.93	85.94
LFoV-MUSE	ResNet-50	$TCGA_{nu}$	**89.53**	**90.18**	86.21	86.19	**82.21**	83.85	87.44	88.86	85.18	85.78	79.88	82.76	85.07	86.27
	ViT-S/16	$TCGA_{nu}$	85.47	87.06	84.17	87.21	79.15	84.22	86.00	86.63	84.95	**86.57**	79.82	83.53	83.26	85.87
	ViT-B/16	$TCGA_{nu}$	89.03	89.20	87.38	86.10	81.11	**84.36**	**88.93**	**90.18**	85.52	86.43	**83.16**	**85.12**	**85.86**	**86.90**

Table 2: Comparison of nucleus classification in ACC % (↑) for K-Nearest Neighbor (KNN) and Linear Probing (LIN). MUSE outperforms other pretraining methods in overall evaluations. The best results are highlighted in bold, and the second-best results are in underlined. custom* denotes mixed datasets. BRCA. and OCEL. denote BRCAM2C and OCELOT, respectively.

marginal advantage over MUSE in KNN.

Furthermore, we conducted pretraining and inference of MUSE with a larger field of view (LFoV-MUSE). Compared to MUSE, ViT-S and ViT-B pretrained with LFoV-MUSE exhibit improvements in FT ACC % of 0.9 and 1.3, respectively. ViT-B pretrained with LFoV-MUSE also outperforms all other SOTA foundation models in KNN, LIN, and FT evaluations. Specifically, it outperforms CONCH by 0.5, 2.4, and 1.7 in KNN, LIN, and FT ACC %, respectively. Remarkably, it also exceeds the performance of much larger models: outperforming UNI (2.4× parameters) by 0.2, 1.3, and 1.0 in KNN, LIN, and FT ACC %, respectively, and surpassing Prov-GigaPath (8.9× parameters) by 1.9, 2.1, and 1.6 in KNN, LIN, and FT ACC %, respectively.

For a fairer comparison, we pretrained models with DINOV2 on $TCGA_{nu}$. While ViT-S benefits from pretraining on $TCGA_{nu}$ compared to LVD-142M, ViT-B exhibits decreased performance, suggesting that the data size of $TCGA_{nu}$ is not enough for large-parameter model pretraining with DINOV2. In contrast, MUSE shows better data efficiency with the same dataset.

MUSE can be adapted to models with various encoder architectures. Table 1 and Table 2 shows that MUSE also significantly outperforms other methods when using ResNet-50 (He et al. 2016) as the encoder. The experiments based on different encoders verify the flexibility of MUSE.

Nucleus Detection and Classification. Table 3 reports the results of nucleus detection and classification. After pretraining with MUSE, simple downstream fine-tuning yields substantially better nucleus classification performance compared to both map-based and point-based SOTA methods. Specifically, ViT-B pretrained with LFoV-MUSE outperforms the best SOTA methods by an average F1-score margin of 3.95, 7.24, and 3.57 on the BRCAM2C, OCELOT, and PUMA datasets, respectively. Importantly, unlike previous work that requires spatial nucleus density statistics (Abousamra et al. 2021) or nucleus graph construction (Lou et al. 2024a), our method achieves superior results with a much simpler task-adaptive pipeline.

Ablation Studies

All experiments are conducted with ViT-B/16 and the small field-of-view, unless otherwise mentioned.

Decoder Pretraining. For models without a decoder, NuLo is applied to the last layer output of the encoder. As shown in Table 4, although the baseline without a decoder still outperforms DINOV2 pretrained with $TCGA_{Nu}$, there is a noticeable drop in performance. Specifically, the model employing both the decoder and multi-level hybrid nucleus representations outperforms the baseline by 3.01, 1.98, and 1.79 in

Method	BRCAM2C				OCELOT			PUMA			
	$F^{Lym.}$	$F^{Tum.}$	$F^{Oth.}$	$F^{Avg.}$	$F^{Tum.}$	$F^{Oth.}$	$F^{Avg.}$	$F^{Lym.}$	$F^{Tum.}$	$F^{Oth.}$	$F^{Avg.}$
MCSpatNet (Abousamra et al. 2021)	63.15	78.56	54.66	65.46	68.60	59.99	64.29	78.25	82.05	51.54	70.61
PointNu-Net (Yao et al. 2023)	<u>71.51</u>	76.02	51.95	66.50	66.72	56.96	61.84	76.31	79.57	52.68	69.52
SMILE (Pan et al. 2023)	**72.59**	<u>79.61</u>	51.06	<u>67.75</u>	66.99	60.10	63.55	<u>80.35</u>	77.54	52.52	70.14
SENC (Lou et al. 2024b)	57.94	76.50	49.42	61.29	70.02	<u>62.08</u>	<u>66.05</u>	77.38	81.51	54.38	71.09
CGT (Lou et al. 2024a)	56.42	75.98	50.44	60.95	68.77	61.30	65.03	76.55	79.66	54.20	70.14
CellViT (Hörst et al. 2024)	67.20	78.20	51.81	65.73	67.36	60.22	63.79	79.07	81.16	<u>57.96</u>	<u>72.73</u>
DPA-P2PNet (Shui et al. 2024)	59.65	77.26	<u>55.26</u>	64.06	<u>70.07</u>	59.92	64.99	76.80	<u>81.87</u>	54.04	70.90
LFoV-MUSE [ViT-B/16] (ours)	70.27	**83.48**	**61.36**	**71.70**	**76.37**	**70.20**	**73.29**	**80.75**	**84.53**	**63.62**	**76.30**

Table 3: Comparison of nucleus detection and classification in F1-score (↑). Pretrained ViT-B/16 with LFoV-MUSE significantly outperforms other methods. The best results are highlighted in bold, and the second-best results are in underlined.

Decoder	Multi-Level Context	KNN	LIN	FT
✗	✗	81.92	83.96	84.47
✓	✗	83.99	85.01	86.19
✓	✓	84.93	85.94	86.26

Table 4: Ablation study of encoder-decoder framework in ACC % (↑). Introducing a decoder and multi-level context to the pretraining effectively enhances model performance.

Global View [Min, Max]	Local View [Min, Max]	20x Evaluation KNN	LIN	FT	40x Evaluation KNN	LIN	FT
[20×, 20×]	[20×, 20×]	82.11	83.77	84.74	77.55	81.01	82.79
[40×, 40×]	[40×, 40×]	79.11	81.08	82.89	84.21	85.65	85.47
[20×, 40×]	[20×, 40×]	84.91	86.24	86.21	84.95	85.64	86.31

Table 5: Ablation study of multi-scale patching in ACC % (↑). Pretraining with Multi-scale patching enables the model to adapt to multiple magnifications.

KNN, LIN, and FT ACC %, respectively. Furthermore, removing multi-level hybrid nucleus representations also leads to a decline in performance.

Multi-Scale Patching. MPP-based Cropping support precise multi-scale patching based on physical resolution. Table 5 reports the ablation study of multi-scale patching. Pretraining at a fixed magnification leads to a significant performance drop at other magnifications. In contrast, pretraining with multi-scale patching enables the model to adapt to different magnifications and improves performance across all magnifications. These results show that multi-scale patching enables MUSE to learn more robust nucleus representations.

Pretraining Losses. Table 6 reports the ablation study of pretraining losses. Building up on the baseline with $\mathcal{L}_{image}$, further introducing $\mathcal{L}_{nucleus}$ leads to better performance of 3.56, 3.42, and 3.92 in KNN, LIN, and FT ACC %, respectively. These results verify the critical role of nucleus-level contrastive learning for better nucleus representation.

Fine-Tuning. Table 7 reports the ablation study of fine-tuning on OCELOT. Notably, even without introducing LFoV or $\mathcal{L}_{type}$, pretrained ViT-B by LFoV-MUSE already outperforms other SOTA methods, highlighting the efficiency of MUSE. Expanding samples to LFoV samples further enriches tissue-level context, resulting in a further average classification F1 increase of 1.63. Besides, the addition of the consistency regularization term $\mathcal{L}_{cons}$ yields an overall improvement of 4.29 relative to the baseline. These results shows the importance of tissue-level context, and verify that the consistency prediction constraint based on unlabeled regions further enhances the generalization of models.

$\mathcal{L}_{image}$	$\mathcal{L}_{nuclei}$	KNN	LIN	FT
✓	✗	81.37	82.52	82.34
✗	✓	84.14	84.31	84.19
✓	✓	84.93	85.94	86.26

Table 6: Ablation study of the pretraining losses in ACC % (↑). $\mathcal{L}_{nuclei}$ highly improves the representation of nuclei.

LFoV	$\mathcal{L}_{cons}$	$F^{Tum.}$	$F^{Oth.}$	$F^{Avg.}$
✗	✗	73.48	64.52	69.00
✓	✗	74.60	66.66	70.63
✓	✓	76.37	70.20	73.29

Table 7: Ablation study of LFoV and $\mathcal{L}_{cons}$ on OCELOT in F1-score (↑). Both improve the performance.

Conclusion

In this work, we propose MUSE, a novel SSL method specifically tailored for nucleus detection and classification. Built around MUSE, we develop a complete framework comprising an encoder-decoder architecture, a pretraining strategy, and a downstream fine-tuning pipeline, enabling effective utilization of unlabeled data and LFoV across all training stages. Extensive experiments demonstrate that MUSE-pretrained models not only significantly surpass existing supervised NDC methods but also outperform generic pathology foundation models even with smaller models and fewer samples. This work highlights the critical role of task-specific pretraining in nucleus-level dense prediction tasks, provides an effective and scalable solution, and paves the way toward general-purpose NDC.

Acknowledgments

This work was supported by DAMO Academy, Alibaba Group, and the National Key R&D Program of China (No.2024YFF0728900).

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.

Assran, M.; Duval, Q.; Misra, I.; Bojanowski, P.; Vincent, P.; Rabbat, M.; LeCun, Y.; and Ballas, N. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15619–15629.

Azizi, S.; Culp, L.; Freyberg, J.; Mustafa, B.; Baur, S.; Kornblith, S.; Chen, T.; Tomasev, N.; Mitrović, J.; Strachan, P.; et al. 2023. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6): 756–779.

Bai, T.; Zhang, Z.; Zhao, C.; and Luo, X. 2021. A novel pseudo-labeling approach for cell detection based on adaptive threshold. In *International Symposium on Bioinformatics Research and Applications*, 254–265. Springer.

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.

Chai, L. T.; Xiao, J. T.; Ab Rahman, K. S.; Bakrin, I. H.; Kam, M. G.; Muhamad, W. Z. A. W.; Chee, C. L.; Wai, L. C.; and Wongsirichot, T. 2025. A Review of Current Cell Annotation Systems for Histopathology Images. *Journal of Advanced Research Design*, 135(1): 88–100.

Chen, R. J.; Ding, T.; Lu, M. Y.; Williamson, D. F.; Jaume, G.; Song, A. H.; Chen, B.; Zhang, A.; Shao, D.; Shaban, M.; et al. 2024. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3): 850–862.

Chen, X.; Xie, S.; and He, K. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9640–9649.

Corredor, G.; Wang, X.; Zhou, Y.; Lu, C.; Fu, P.; Syrigos, K.; Rimm, D. L.; Yang, M.; Romero, E.; Schalper, K. A.; et al. 2019. Spatial architecture and arrangement of tumor-infiltrating lymphocytes for predicting likelihood of recurrence in early-stage non–small cell lung cancer. *Clinical cancer research*, 25(5): 1526–1534.

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*.

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.

Hörst, F.; Rempe, M.; Becker, H.; Heine, L.; Keyl, J.; and Kleesiek, J. 2025. Cellvit++: Energy-efficient and adaptive cell segmentation and classification using foundation models. *arXiv preprint arXiv:2501.05269*.

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.

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.

Liu, J.; Lichtenberg, T.; Hoadley, K. A.; Poisson, L. M.; Lazar, A. J.; Cherniack, A. D.; Kovatich, A. J.; Benz, C. C.; Levine, D. A.; Lee, A. V.; et al. 2018. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2): 400–416.

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*, 43(9): 3149–3160.

Lu, M. Y.; Chen, B.; Williamson, D. F.; Chen, R. J.; Liang, I.; Ding, T.; Jaume, G.; Odintsov, I.; Le, L. P.; Gerber, G.; et al. 2024. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3): 863–874.

Misra, I.; and Maaten, L. v. d. 2020. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6707–6717.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Page, D. B.; Broeckx, G.; Jahangir, C. A.; Verbandt, S.; Gupta, R. R.; Thagaard, J.; Khiroya, R.; Kos, Z.; Abduljabbar, K.; Acosta Haab, G.; et al. 2023. Spatial analyses of immune cell infiltration in cancer: current methods and future directions: A report of the International Immuno-Oncology Biomarker Working Group on Breast Cancer. *The Journal of pathology*, 260(5): 514–532.

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.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12179–12188.

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.

Schuiveling, M.; Liu, H.; Eek, D.; Breimer, G. E.; Suijkerbuijk, K. P.; Blokx, W. A.; and Veta, M. 2025. A novel dataset for nuclei and tissue segmentation in melanoma with baseline nuclei segmentation and tissue segmentation benchmarks. *GigaScience*, 14: giaf011.

Shui, Z.; Zheng, S.; Zhu, C.; Zhang, S.; Yu, X.; Li, H.; Li, J.; Chen, P.; and Yang, L. 2024. 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.

Su, L.; Wang, Z.; Zhu, X.; Meng, G.; Wang, M.; and Li, A. 2023. Dual consistency semi-supervised nuclei detection via global regularization and local adversarial learning. *Neurocomputing*, 529: 204–213.

Wang, S.; Rong, R.; Zhou, Q.; Yang, D. M.; Zhang, X.; Zhan, X.; Bishop, J.; Chi, Z.; Wilhelm, C. J.; Zhang, S.; et al. 2023. Deep learning of cell spatial organizations identifies clinically relevant insights in tissue images. *Nature communications*, 14(1): 7872.

Wang, X.; Yang, S.; Zhang, J.; Wang, M.; Zhang, J.; Yang, W.; Huang, J.; and Han, X. 2022a. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81: 102559.

Wang, X.; Yang, S.; Zhang, J.; Wang, M.; Zhang, J.; Yang, W.; Huang, J.; and Han, X. 2022b. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81: 102559.

Wang, X.; Zhao, J.; Marostica, E.; Yuan, W.; Jin, J.; Zhang, J.; Li, R.; Tang, H.; Wang, K.; Li, Y.; et al. 2024. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035): 970–978.

Xu, H.; Usuyama, N.; Bagga, J.; Zhang, S.; Rao, R.; Naumann, T.; Wong, C.; Gero, Z.; González, J.; Gu, Y.; et al. 2024. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015): 181–188.

Yang, X.; Song, Z.; King, I.; and Xu, Z. 2022. A survey on deep semi-supervised learning. *IEEE transactions on knowledge and data engineering*, 35(9): 8934–8954.

Yao, K.; Huang, K.; Sun, J.; and Hussain, A. 2023. Pointnunet: 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, P.; Gao, C.; Zhang, Z.; Yuan, Z.; Zhang, Q.; Zhang, P.; Du, S.; Zhou, W.; Li, Y.; and Li, S. 2025. Systematic inference of super-resolution cell spatial profiles from histology images. *Nature Communications*, 16(1): 1838.

Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; and Kong, T. 2021. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*.