UniCell: Universal Cell Nucleus Classification via Prompt Learning

Junjia Huang[1,2*], **Haofeng Li**[2*], **Xiang Wan**[2], **Guanbin Li**[1,3†]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China
[3] GuangDong Province Key Laboratory of Information Security Technology
huangjj77@mail2.sysu.edu.cn, {lhaof,wanxiang}@sribd.cn, liguanbin@mail.sysu.edu.cn

Abstract

The recognition of multi-class cell nuclei can significantly facilitate the process of histopathological diagnosis. Numerous pathological datasets are currently available, but their annotations are inconsistent. Most existing methods require individual training on each dataset to deduce the relevant labels and lack the use of common knowledge across datasets, consequently restricting the quality of recognition. In this paper, we propose a universal cell nucleus classification framework (UniCell), which employs a novel prompt learning mechanism to uniformly predict the corresponding categories of pathological images from different dataset domains. In particular, our framework adopts an end-to-end architecture for nuclei detection and classification, and utilizes flexible prediction heads for adapting various datasets. Moreover, we develop a Dynamic Prompt Module (DPM) that exploits the properties of multiple datasets to enhance features. The DPM first integrates the embeddings of datasets and semantic categories, and then employs the integrated prompts to refine image representations, efficiently harvesting the shared knowledge among the related cell types and data sources. Experimental results demonstrate that the proposed method effectively achieves the state-of-the-art results on four nucleus detection and classification benchmarks. Code and models are available at https://github.com/lhaof/UniCell

Introduction

Histopathological analysis is widely considered as the gold standard for the diagnosis and prognosis of human cancers (Wu et al. 2022; Zeiser et al. 2021). Locating and classifying the cells in histopathology images is a preliminary step in analyzing, diagnosing and grading cancerous cells. In practical applications, the procedures such as cell counting (Fridman et al. 2012), tumor grading (Fleming et al. 2012) and computer-aided diagnosis (CAD) (Saha, Mukherjee, and Chakraborty 2016) all require to identify nuclei as a fundamental task. Some studies (Abousamra et al. 2021; Huang et al. 2023b) aim to identify both the location and type of cells, while some other works (Stringer et al. 2021; Lou et al. 2022, 2023b; Ma et al. 2023; Yu et al. 2023) attempt to determine the nucleus boundaries. However, due

---

*These authors contributed equally.
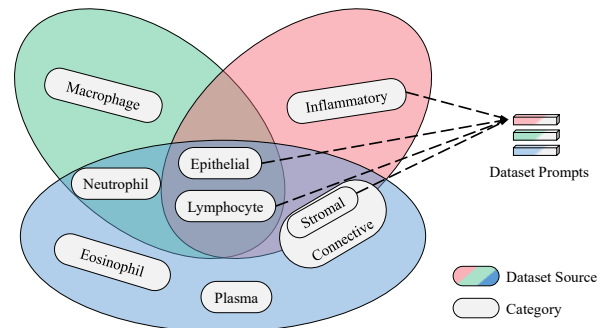
†Guanbin Li is the corresponding author.

Figure 1: The illustration of universal multi-dataset cell nucleus classification. The lack of a unified standard for annotating nucleus types hinders the efficient utilization of data and labels. For example, The Lizard dataset (Blue) has three overlapping classes (Neu., Epi., Lym.) with the MoNuSAC dataset (Green). Besides, there exists an inclusion relationship between some categories, such as connective and stromal cells. Our approach utilizes multiple datasets and their associated labels as prompts for training a unified model.

to the time-consume process of pixel-wise labeling, some datasets only provide the centroid coordinates of nuclei instead of boundaries. Thus, in this study we concentrate on locating the centroids and inferring the types of nuclei, by unifying cross-datasets labels to learn a universal model.

In recent years, numerous pathological datasets (Graham et al. 2019, 2021; Verma et al. 2021; Ryu et al. 2023) with nucleus annotations have been proposed for deep learning based computational pathology (CPath). However, due to the varying locations of collected organs and the diversity of diseases, the researchers adopt different annotating protocols for these datasets. As Figure 1 shows, the annotations from different benchmarks may partially overlap but are not exactly the same. Thus, most existing methods are limited to be trained and evaluated on each single dataset. Furthermore, the scale, the diversity and the numbers of samples in pathology benchmarks vary considerably. The labels of some datasets are insufficient for learning a robust model. Consequently, most of previous cell recognition approaches (Abousamra et al. 2021; Doan et al. 2022; Huang et al. 2022; Amgad et al. 2022) fail to adapt to or exploit the

common knowledge from multiple data sources.

To overcome the discrepancies among datasets, some recent works (Tellez et al. 2020; Mormont, Geurts, and Marée 2020; Gamper and Rajpoot 2021; Graham et al. 2023) explore more general approaches from a multi-task perspective, which could acquire ample pathological representations to extend to other benchmarks via transfer learning. However, these approaches require an additional training phase to tune a model for each single task or data source, which leads to increasing computational costs and extra training time. Some works (Qin et al. 2022; Zhang et al. 2023b) consider merging the labels from distinct sources, employing label semantics to fully utilize all annotated datasets. However, these methods are confined to classifying image patches of cell nuclei, without achieving end-to-end detection and recognition.

To harvest shared knowledge and prevent redundant training for correlated tasks, we propose an end-to-end framework for cell nucleus detection and classification, and train the single proposed model with multiple datasets concurrently. We resort to prompt learning to represent label semantics and dataset attributes, which enables jointly training a model with multiple data sources and adapting features to various cell types. Specifically, we adopt a DETR based architecture that directly predicts cell locations and categories based on candidate queries, and introduce the contrastive denoising approach to expedite convergence. To fully utilize data from various sources, we employ multiple prediction heads at the end of the proposed model, and the parameters of the model except the heads are shared by different nucleus classification tasks. To well adapt to the categorical semantics from varied sources, we devise a new dynamic prompt module. The module learns the embeddings of category prompts to construct a category memory bank, which is further utilized to yield semantic-aware dataset prompts. After that, image features are enhanced by the dataset prompts in a targeted manner, and become more adaptable to the corresponding tasks. Therefore, we propose an end-to-end model capable of making predictions on multiple data sources. With the benefit of cross-dataset samples, our approach can achieve superior results in cell detection and classification. In short, our major contributions are summarized in three folds:

- We propose UniCell, a novel end-to-end universal nucleus recognition framework, which can learn a single model from multiple datasets concurrently without additional parameters or time for transfer.

- We develop a new Dynamic Prompt Module integrated with UniCell. The module can adapt the representation learning to the categorical semantics from different sources using dataset and label prompts.

- We validate UniCell through extensive experiments on four public benchmarks. The results show that our method significantly outperforms the state-of-the-arts.

Related Work

Cell Nucleus Classification. Most traditional solutions to nucleus recognition are based on the Watershed algorithm and handcrafted features (Malpica et al. 1997; Plissiti, Nikou, and Charchanti 2010; Xu et al. 2016), which lack sufficient accuracy and generalization. Recently, deep-learning based approaches have achieved notable success. Some methods (Abousamra et al. 2021; Wang et al. 2022; Sugimoto et al. 2022) utilize a UNet (Ronneberger, Fischer, and Brox 2015) structure to detect cells via pixel-wise binary classification. When given the annotation of nuclei segmentation masks, some methods (Graham et al. 2019; Doan et al. 2022; Chen et al. 2023) distinguish nuclei by learning to infer the label of each pixel. Despite of impressive results, these methods are usually trained independently on a single dataset. To exploit multiple datasets, most methods (Mormont, Geurts, and Marée 2020; Wang et al. 2021; Yang et al. 2022; Huang et al. 2023a; Lou et al. 2023a) resort to pre-training or transfer learning, which requires additional training costs. Graham et al. (2023) employs multiple task-specific decoders to merge multi-task data in the model for concurrent training, enabling joint prediction with a single network. Besides, some works (Qin et al. 2022; Zhang et al. 2023b) focus on unifying the labels. Zhang et al. (2023b) formulate the nucleus classification task as a multi-label problem with missing labels. These methods are limited to classification tasks and require multi-stage training. Differently, we build an end-to-end detection and classification model trained with multiple datasets in one stage.

Prompt Learning. Prompt learning originally refers to prepending language instruction to the input text allowing the model to better understand the task. Some works (Zhou et al. 2022b,a; Khattak et al. 2023) input the prompts to the language branch of a Visual-Language model, extracting useful information from the knowledge base for downstream tasks. Zhou et al. (2022b) models a prompt's context words with learnable vectors. Jia et al. (2022) introduce a small number of trainable parameters for the visual model to improve the transfer effect. For universal training, Jain et al. (2023) utilizes the task type as a prompt to distinguish the semantic segmentation and instance segmentation tasks. Different from existing prompt learning methods, we utilize the fusion of prompt sentences to differentiate feature information and mitigate interference across various datasets.

Methodology

Overall Framework

The proposed method employs a DETR-like (Carion et al. 2020) structure, which directly predicts of the centroid positions and categories of nuclei without complicated post-refinements. As shown in Figure 2, the proposed UniCell is an end-to-end architecture that has a feature extraction backbone, a feature encoder, a dynamic prompt module, a feature decoder, and detection-classification heads (D&C Heads).

Assume that we need to solve the nuclei identification task for D different dataset sources whose sets of cell labels are not exactly the same. The dth dataset has n_d samples, and is denoted as a set of training samples, $S_d = \{(x_i^d, y_i^d)\}_{i=1}^{n_d}$. Given a pathological image $x_i^d \in \mathbb{R}^{H \times W \times 3}$ from the dth dataset, we utilize the image x_i^d, the ground-truth y_i^d and the
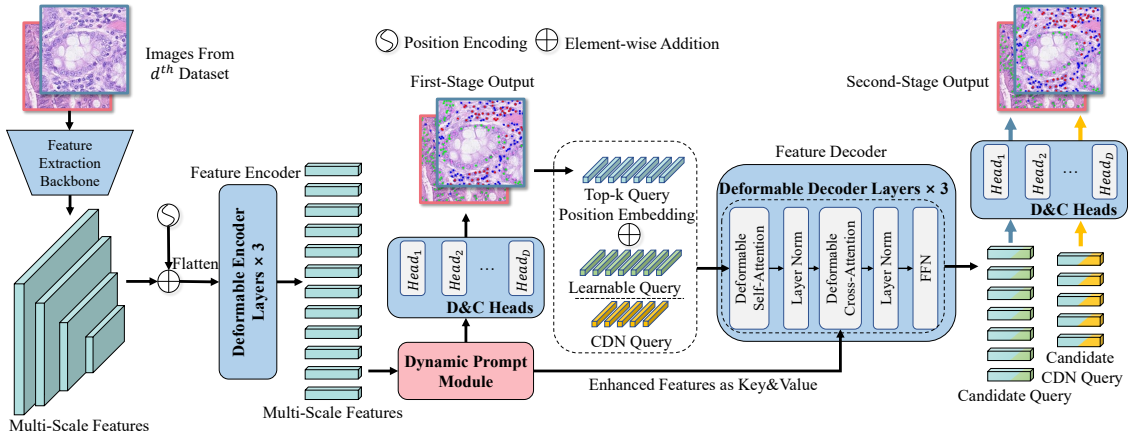
Figure 2: The framework of our proposed UniCell. The multi-scale feature and Dataset ID are inputted into the Dynamic Prompt Module for the specific dataset prompt. Note that both the deep supervision and query refinement in the decoder layer are omitted for readability.

corresponding dataset ID d as inputs. First, we use the feature extraction backbone Swin Transformer (Liu et al. 2021) to extract multi-scale features from the input image. Then the extracted features are fed into the feature encoder with their corresponding positional encodings. The feature encoder has three deformable self-attention (Zhu et al. 2021) layers. After the encoding, the flattened multi-scale features are input to the proposed dynamic prompt module for further feature enhancement. Subsequently, the enhanced representations are processed by the first-stage D&C Heads for predicting cell positions and categorical probabilities. The D&C Heads contain D different prediction heads for the D datasets. Each head consists of two parallel fully connected (FC) layers for detection and classification.

For the dth dataset containing t_d cell types, its D&C Head outputs the coordinate and a t_d-element vector as the categorical confidence of each nuclues. The positions of the top k cell nuclei with the highest confidence are utilized to initialize the query position embedding that describes the locations of anchor centroids and reference points. The query position embeddings are merged with k learnable content queries. In the feature decoder, we compute deformable cross-attention between the k merged queries and the enhanced features to generate k candidate queries.

Contrastive DeNoising Query. Inspired by DINO (Zhang et al. 2023a), we adopt the Contrastive DeNoising (CDN) approach to avoid duplicate predictions and accelerate training convergence. Specifically, the original CDN method embeds ground truth (GT) labels and bounding boxes with noise to generate positive and negative noise queries, namely CDN queries. The scale of noise is controlled by two hyperparameters λ_1 and λ_2 ($\lambda_1 < \lambda_2$). During the training stage, both the CDN queries and the learnable content ones are utilized. During the inference, only the learnable content queries are deployed. Similar to the learnable content queries, we calculate cross-attention between the CDN queries and enhanced features to generate candidate CDN queries, reconstructing the GT independently.

Since we detect and classify nuclei via predicting their centroids, the pathology datasets only have the annotations of centroid locations in our setting. Thus, different from the original CDN, we only apply location-shifting noise to the coordinates of GT centroids. Formally, we denote the GTs as $y_i^d = \{(u_j, v_j), c_j\}_{j=1}^{n_c}$ with n_c annotations. For each annotation, we randomly generate a pair of positive position offset $(|\Delta u_p|, |\Delta v_p|) \in (0, \lambda_1)$ and a pair of negative one $(|\Delta u_n|, |\Delta v_n|) \in (\lambda_1, \lambda_2)$ as noise. The noisy annotation $y_{i\_\{p,n\}}^{d\_noise}$ can be formulated as:

$$y_{i\_\{p,n\}}^{d\_noise} = \{(u_j + \Delta u_{\{p,n\}}, v_j + \Delta v_{\{p,n\}}), RF(c_j, \gamma)\}_{j=1}^{n_c}, \tag{1}$$

where $RF(\cdot, \gamma)$ denotes randomly switching the GT labels to others with a probability of γ. The CDN queries are generated by embedding the locations and categories of $y_{i\_\{p,n\}}^{d\_noise}$.

As shown in Figure 2, after the decoder produces the candidate queries and candidate CDN queries, they are sent into the dth head of the second-stage prediction heads, which are similar to the first-stage ones. An independent prediction head is deployed for each dataset. Each head uses two FC layers for detection and classification. We denote the results of candidate queries output by the head as $\overline{y}_i^d = \{(\overline{u}_j, \overline{v}_j), \overline{c}_j\}_{j=1}^{n_q}$, where n_q is the number of candidate queries. The predictions of positive and negative candidate CDN queries are denoted as $\hat{y}_p^d = \{(\hat{u}_p, \hat{v}_p), \hat{c}_p\}_{p=1}^{n_c}$ and $\hat{y}_n^d = \{(\hat{u}_n, \hat{v}_n), \hat{c}_n\}_{n=1}^{n_c}$, respectively. We adopt Hungarian loss (Kuhn 1955; Carion et al. 2020) $H(\cdot)$, L1 loss $L_1(\cdot)$, and Focal loss (Lin et al. 2017) $Focal(\cdot)$ to match the predictions and GTs. The loss function for the dth dataset is formulated as:

$$\mathcal{L}^d(y_i^d, \overline{y}_i^d, \hat{y}_{\{p,n\}}^d) = H(y_i^d, \overline{y}_i^d) + L_1(y_i^d, \hat{y}_p^d) + $$
$$Focal(y_i^d, \hat{y}_p^d) + Focal(\widetilde{y}_i^d, \hat{y}_n^d) \tag{2}$$

where $\widetilde{y}_i^d$ denote empty objects and are set to $t_d + 1$, $L_1(\cdot)$ calculates the Euclidean distance between coordinates and $Focal(\cdot)$ measures the category difference. The overall loss
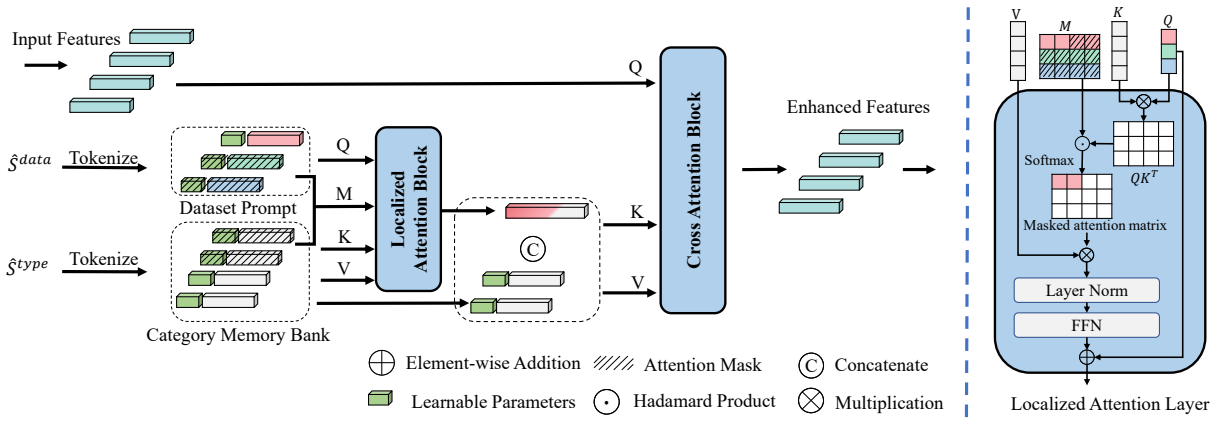
Figure 3: The proposed Dynamic Prompt Module. Both dataset prompt and category memory bank are tokenized and embedded from priori textual sequences. We update the dataset prompts with the embeddings in the category memory bank, and adopt the updated prompts to enhance input representations.

for all the datasets is $\sum_{d=1}^{D} \omega_d \mathcal{L}^d(y_i^d, \overline{y}_i^d, \hat{y}_{\{p,n\}}^d)$, where ω_d denotes a balanced factor for the dth dataset. Additionally, each feature decoder layer produces the offsets to refine the centroid coordinates and reference points of the previous layer or the decoder input. These outputs from intermediate decoder layers are sent to independent prediction heads from deep supervision.

Dynamic Prompt Module

As shown in Figure 3, we propose a Dynamic Prompt Module (DPM), which dynamically adapts the intermediate representations to different dataset sources, via leveraging the dataset name and its label properties as learnable prompts. The DPM takes a sequence of features as input, and outputs the prompt-adapted representations. Assume that the D datasets we use have C different cell types. The set of names of these datasets is denoted as: $S^{data} = \{Dataset\_i\}_{i=1}^{D}$, where $Dataset\_i$ is the name of the ith dataset. To compute the prompt embedding of a dataset, we construct a textual sequence (sentence) for each dataset, by adopting learnable unified tokens as Context Optimization (CoOp) (Zhou et al. 2022b) before the dataset name. The textual sequences of all the D datasets are represented as:

$$\hat{S}^{data} = \{[V_1^{data}][V_2^{data}] \cdots [V_T^{data}][Dataset\_i]\}_{i=1}^{D}, \quad (3)$$

where $\hat{S}_i^{data}$ denotes the sentence of the ith dataset and T is the number of tokens. $V_1^{data}, V_2^{data}, \cdots, V_T^{data}$ are the learnable tokens, and are shared by the D datasets. '$[*][*]\cdots$' means connecting the tokens into a textual sequence or a sentence. We convert each sentence into a lower-cased byte pair encoding (BPE) (Sennrich, Haddow, and Birch 2016) representation with a vocabulary size of 49,408. Each text sequence is encompassed with the [SOS] and [EOS] tokens, and is capped at a fixed length L^S set as the CoOp method (Zhou et al. 2022b). The sequence of size L^S is tokenized and embedded into a 256-dimensional vector. All the dataset representations form a $D \times 256$ matrix, and are also called *Dataset Prompts* as shown in Figure 3.

To compute the common representations for the C categories, we adopt a similar way by building a textual sequence for each category. These sequences are denoted as:

$$\hat{S}^{type} = \{[V_1^{type}][V_2^{type}] \cdots [V_T^{type}][Category\_i]\}_{i=1}^{C}, \quad (4)$$

where *Category_i* is the name of the ith category. The textual sequences of all C categories are also tokenized by the BPE, attached with [SOS] & [EOS] and capped at the length L^S. Consider that the category memory bank may need to restore the various patterns of cell types, and should be modeled by multiple vectors. Thus, for each category, its L^S tokens are converted into L^S 256-dimensional embeddings. Then we obtain a $C \times L^S \times 256$ matrix, called *Category Memory Bank*, which is supposed to maintain the common knowledge for identifying cell types.

We first compute the cross-attention between the dataset prompt and the category memory bank, to construct a dataset-specific prompt embedding space. Since a particular dataset may not contain all categories, performing correlation calculations on absent categories would be illogical. Therefore, we employ an attention mask for localized attention computations. We use $M = [m_{ij}]_{D \times CL^S}$ to denote the attention mask. If the category of jth prompt embedding in the memory bank is invisible to the ith dataset, $m_{ij} = -\infty$. Otherwise, $m_{ij} = 0$. The attention mask is devised as:

$$m_{ij} = \begin{cases} 0, & if \ i = d \ \& \ \lceil j/L^S \rceil \in I_d \\ -\infty, & otherwise \end{cases}, \quad (5)$$

where I_d is the set of category indices belonging to the dth dataset. $\lceil j/L^S \rceil$ returns the category index of the jth embedding in the memory bank.

To embed categorical semantics into the dataset prompts, we set the dataset embeddings to Query Q and the category memory bank to Key K & Value V, respectively. Then we perform the cross-attention operator by sending Q, K and V into a localized attention block, which contains L localized attention layer. Each layer is formulated as:

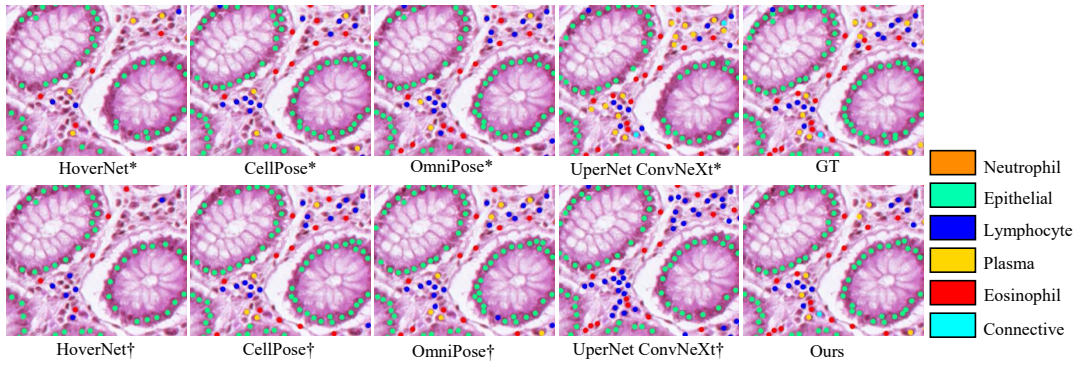$$Q_{l+1} = FFN(LN(CA(Q_l, K_l, V_l, M))) + Q_l, \quad (6)$$

Figure 4: Qualitative comparison on the Lizard dataset. Five types of cells are marked with dilated nucleus centroids in five different colors. As the results show, the category distribution of our method is the closest to that of the ground truths.

where l is the layer index, $LN(\cdot)$ denotes the layer normalization (Ba, Kiros, and Hinton 2016), $FFN(\cdot)$ denotes a feed forward network. $CA(\cdot)$ denotes a variant of cross-attention guided by an attention mask M, and is formulated as:

$$CA(Q_l, K_l, V_l, M) = softmax(M + Q_l K_l^T) V_l, \quad (7)$$

where $\odot$ is Hadamard Product. The shape of Q_l / K_l / V_l is $D \times 256$ / $CL^S \times 256$ / $CL^S \times 256$. The localized attention block outputs the updated dataset prompt that is then concatenated with its visible category embeddings in the memory bank, as shown in Fig. 3. The concatenated variables serve as Key and Value to be input into a cross attention block as Key and Value to refine the input features of DPM. The cross attention block has L cross attention layers and the input features is set to the Query for attention.

Feature-Enhancing or Query-Enhancing Strategy. In our method, we consider two enhancement strategies: the Feature-Enhancing Strategy and the Query-Enhancing Strategy. The Feature-Enhancing Strategy, depicted in Fig. 3, takes the multi-scale features as the DPM input and updates the features at a pixel level. This allows the model to distinguish fine-grained details across the entire image, thereby facilitating the subsequent identification of cells. The query-enhancing strategy takes the learnable content queries as the DPM input, sends the output into the feature decoder. The strategy ensures that each content query is aware of the categorical semantics within the current dataset. Note that when setting the learnable content queries as the DPM input, the unenhanced multi-scale features are directly fed into the feature decoder as keys and values.

Experiments

Implementation Details

Dataset. We conduct experiments on four datasets: **CoN-SeP** (Graham et al. 2019) consists of 41 H&E stained image tiles from 16 colorectal adenocarcinoma whole-slide images (WSIs). Following the work (Abousamra et al. 2021), we resize the images in CoNSeP to $20\times$ and merge the labels into three types of cells: inflammatory, epithelial, and stromal. **MoNuSAC** (Verma et al. 2021) contains over 46,000 nuclei from 37 hospitals and 71 patients, sourced from four organs and categorized into four nucleus types: epithelial cells,

lymphocytes, neutrophils, and macrophages. **Lizard** (Graham et al. 2021) is collected from six datasets sources, and has nearly half a million labeled nuclei in H&E stained colon tissue. These nuclei are categorized into six types of cells: epithelial cells, connective tissue cells, lymphocytes, plasma cells, neutrophils and eosinophils. To ensure the independence among data sources, we remove the images belonging to the CoNSeP subset from the Lizard dataset. **OCELOT** (Ryu et al. 2023) consists of 400 images from six organs with the annotations of nuclei and tissues. The categories of OCELOT include background cells and tumor cells. For OCELOT, we only utilize its nuclues annotations and split the dataset into training and testing sets with a ratio of 7:3. For other datasets, we use their default/official data split. During the training, we combine all training images from four datasets and randomly sample images for model training. During the inference, we evaluate a model on the test set of each benchmark.

Evaluations. We utilize the F-score as the evaluation metrics for nucleus detection and classification tasks, following the works (Graham et al. 2019). For the detection, we compute the Euclidean distance between each predicted centroid and GT to yield a cost matrix. Then we run the Hungarian algorithm (Kuhn 1955) with the cost matrix to obtain the paired results. The paired results with a distance less than a predefined radius r are correctly detected centroids (TP_d, d for detection) while the rest of paired results and unpaired predicted centroids are overdetected predicted centroids (FP_d). The unpaired GT centroids are called misdetected GT (FN_d). The detection F-score is calculated as: $F_d = \frac{2TP_d}{2TP_d + FP_d + FN_d}$. We set the radius r as 15 pixels for OCELOT and 6 pixels for the other three datasets, following the works (Ryu et al. 2023; Sirinukunwattana et al. 2016).

For the classification task with K classes, TP_d are split into multiple subsets: correctly classified centroids of Type k (TP_c^k), incorrectly classified centroids of Type k (FP_c^k) and incorrectly classified centroids of types other than Type k (FN_c^k). The classification F-score is defined as: $F_c^k = \frac{2TP_c^k}{2(TP_c^k + FP_c^k + FN_c^k) + FP_d + FN_d}$. The average F-score of classification is $\overline{F_c} = \sum_{k=1}^{K} F_c^k$.

Hyper-parameters setting. We adopt Swin-B (Liu et al.

Methods	CoNSeP		MoNuSAC		Lizard		OCELOT	
	F_d	$\overline{F_c}$	F_d	$\overline{F_c}$	F_d	$\overline{F_c}$	F_d	$\overline{F_c}$
HoverNet (Graham et al. 2019)*	0.621	0.503	0.822	0.641	0.729	0.430	<u>0.738</u>	0.501
HoverNet (Graham et al. 2019)†	0.640	0.417	0.802	0.341	0.747	0.305	0.723	0.482
Cellpose (Stringer et al. 2021)*	0.625	0.492	<u>0.828</u>	0.684	0.783	0.450	0.715	0.502
Cellpose (Stringer et al. 2021)†	0.604	0.418	0.826	0.650	0.785	0.447	0.731	0.518
Omnipose (Cutler et al. 2022)*	0.678	0.528	0.784	0.642	0.788	0.452	0.716	0.509
Omnipose (Cutler et al. 2022)†	0.625	0.439	0.788	0.635	<u>0.799</u>	0.461	0.720	0.515
UperNet ConvNeXt (Liu et al. 2022)*	<u>0.715</u>	<u>0.595</u>	0.801	<u>0.712</u>	0.764	<u>0.474</u>	0.692	0.509
UperNet ConvNeXt (Liu et al. 2022)†	0.652	0.418	0.784	0.708	0.458	0.266	0.732	<u>0.540</u>
UniCell (Ours)	**0.762**	**0.679**	**0.847**	**0.752**	**0.813**	**0.596**	**0.751**	**0.596**

Table 1: Comparisons with existing methods on 4 datasets, CoNSeP, MoNuSAC, Lizard and OCELOT. The previous methods are trained in two manners. * denotes the results of training a model on each dataset separately, while † represents the results of training a model with all 4 datasets. For each column, the best method is in bold type and the second best method is underlined.

UniCell	CoNSeP		MoNuSAC		Lizard		OCELOT		Average	
	F_d	$\overline{F_c}$	F_d	$\overline{F_c}$	F_d	$\overline{F_c}$	F_d	$\overline{F_c}$	F_d^{avg}	$\overline{F_c}^{avg}$
w/o DPM	0.734	0.650	0.832	0.751	0.806	0.556	0.728	0.556	0.775	0.628
w/o Category MB	0.741	0.661	0.843	0.748	0.805	0.578	0.741	0.586	0.783	0.643
w/o Dataset Prompt	0.747	0.655	0.842	0.744	0.809	0.573	0.740	0.584	0.785	0.639
Query-Enhancing	0.746	0.660	0.836	0.745	0.808	0.580	0.749	0.588	0.785	0.643
Feature-Enhancing (Ours)	**0.762**	**0.679**	**0.847**	**0.752**	**0.813**	**0.596**	**0.751**	**0.596**	**0.793**	**0.656**

Table 2: Ablation study of Dynamic Prompt Module (DPM) on four benchmarks. w/o DPM, Category MB and Dataset Prompt mean the results of removing DPM, category memory bank and dataset prompts from our proposed method, respectively. Query-Enhancing is to deploy DPM to enhance learnable content queries before the feature decoder, while Feature-Enhancing (Ours) is to refine multi-scale features with DPM (Figure 2). For each column, the best method is in bold.

2021) with ImageNet (Deng et al. 2009) pre-trained weights as the backbone. The number of both feature encoder and decoder layers is set to 3. The number of candidate queries is 900. The learnable tokens number M is 16. The number of Localized Attention Layer L is 3. AdamW optimizer is used to train UniCell with initial learning rates of $1e^{-4}$ and $1e^{-5}$ for backbone and other modules, respectively. For data argumentation, we apply the random flip, random crop and multi-scale training with sizes between 600 and 800, and infer images after resizing to 800×800. All models are trained and tested with an NVIDIA A100 (80GB) GPU. The number of training iterations is set to 160k and after training for 160k iterations, we choose the final model for evaluation. We use the SAHI (Akyon, Altinuc, and Temizel 2022) scheme to slice the images into fixed-size patches as training samples and adopt sliding-window prediction during inference.

Comparison with State-of-the-arts

In Tab. 1, we compare our proposed method with existing methods HoverNet (Graham et al. 2019), Cellpose (Stringer et al. 2021), Omnipose (Cutler et al. 2022) and UperNet with ConvNeXt (Liu et al. 2022) backbone. We try to train each of existing methods with all 4 datasets concurrently to build a single model by merging cross-sources labels into 9 categories, which is denoted as †. These models use the same training data as ours. We also train each method on each dataset separately, which is denoted by * in Tab. 1. Our

method achieves the highest mean F-score in both detection and classification tasks on all four benchmarks. Fig. 4 presents a qualitative comparison between the previous approaches and ours.

In Tab. 1, our method obtains 4.7%, 1.9%, 1.4% and 1.3% F-score in detection higher than the second best results on four datasets. The results show that our method effectively addresses detection obstacle caused by the various image sizes and nucleus densities across different datasets. Since the categories in each dataset do not entirely overlap, it may cause a more severe class imbalance issue after merging datasets. For the classification task, our method outperforms the second best method by 8.4%, 4%, 12.2% and 5.6% F-scores on four benchmarks, respectively. As the results suggest, our method with the dynamic prompt module effectively leverages the data from various sources to improve the model ability.

Ablation Study

Effectiveness of the proposed dynamic prompt module (DPM). In DPM, we utilize both the dataset prompt and category memory bank to enhance image features. To validate the efficacy of DPM, category memory bank (CMB) and dataset prompts, we examin the UniCell model after removing each of the three components, respectively. When removing dataset prompts, the category memory bank serves as keys and values to enhance multi-scale features. Compar-
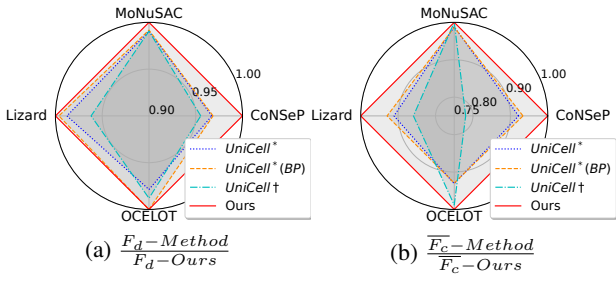
Figure 5: Different ways of using multiple datasets. 'Uni-Cell*' refers that UniCell with one (instead of four) head is trained independently on each dataset after removing DPM. 'UniCell*(BP)' denotes training 'UniCell*' from the weights pre-trained on a binary detection & classification task. 'UniCell†' uses the training data of four sources to train a UniCell* model by merging the categories of all datasets.

ing our method with 'w/o DPM' shows that DPM improves the baseline by 1.8% in F_d^{avg} and 2.8% in $\overline{F}_c^{avg}$ on the four datasets. Notably, DPM proves beneficial for the dataset with fewer samples, such as CoNSeP, and the dataset with fewer categories, such as OCELOT. These results show that DPM can well adapt the visual features to different datasets via exploiting their common semantic knowledge.

Besides, we study the proposed two deploying strategies of DPM, query-enhancing and feature-enhancing. The query-enhancing strategy deploys DPM before the feature decoder to refine content queries, which only provides instance-level information from dataset and category prompts. In contrast, the feature-enhancing strategy enhances multi-scale features at a pixel-level granularity, and the enhanced representations are also utilized to predict instance-level content queries. As shown in Tab. 2, both strategies are effective to improve the baseline w/o DPM by 1.5%-2.8% $\overline{F}_c^{avg}$. The feature-enhancing strategy is the better, exceeding the other by 1.3% in $\overline{F}_c^{avg}$. Thus, we adopt the feature-enhancing strategy in our model by default.

Effectiveness of the way to merge. In Fig. 5, we present the results of exploiting different datasets in various manners. 'UniCell*' denotes the UniCell that has one (instead of four) head and is independently trained on each dataset after removing DPM. To enable the model to employ the information of all datasets, we pretrain the UniCell* model on a binary detection and classification task using all datasets, and then finetune the model on the multi-label classification tasks using each single dataset, respectively. The resulting model is denoted as 'UniCell*(BP)' with BP for 'binary-class pretraining'. UniCell† denotes training a one-head model with all training sets after merging their categories. Fig. 5(b) shows that neither pre-training nor combined category training effectively improve the baseline of individual training (UniCell*). In contrast, our method manages to exploit the cross-sources knowledge and achieve better performance on all datasets with just one training stage.

Amount of the Localized Attention Layer. Fig. 6 shows the results of using different numbers of Localized Attention
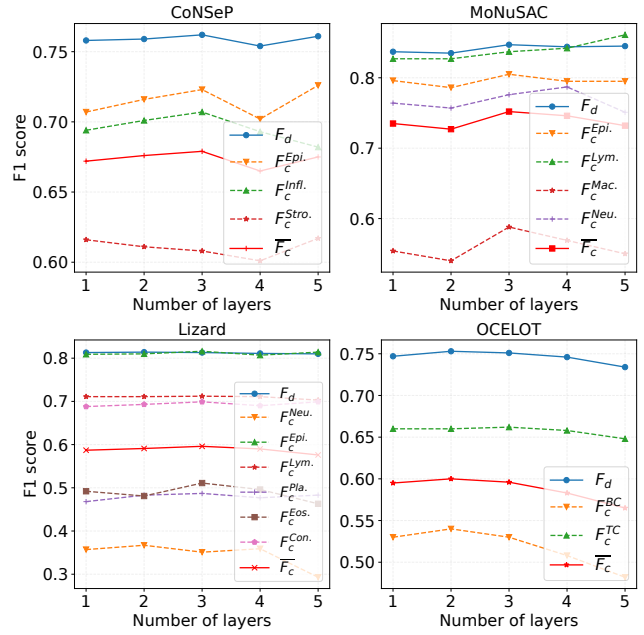

Figure 6: The effects of the number of Localized Attention Layers. $F_c^{Epi.}$, $F_c^{Infl.}$, F_c^{Stro}, $F_c^{Lym.}$, $F_c^{Mac.}$, $F_c^{Neu.}$, $F_c^{Pla.}$, $F_c^{Eos.}$, $F_c^{Con.}$, F_c^{BC} and F_c^{TC} are the F-scores of the epithelial, inflammatory, stromal, lymphocytes, macrophages, neutrophils, plasma, eosinophil, connective tissue cells, background cells and tumor cells, respectively.

Layers L. We provide the classification F-scores for each category in each dataset as well as the detection F-scores. The results show that setting L to 3 performs the best. The model performance does not vary significantly for smaller amounts of $L(L \leq 3)$. The increase of L degrades the performance of a small number of classes, such as neutrophils, macrophages and background cells.

Conclusion

In this paper, we propose an end-to-end universal cell nucleus recognition architecture (UniCell), which can locate and identify multi-category nuclei of cross-dataset pathological images using a single model. We build the overall framework based on DETR that enables the direct detection and classification of nuclei without post-processing, and introduce Contrastive DeNoising to achieve fast and robust convergence. Importantly, to adapt to multiple data sources, we develop a novel dynamic prompt module that employs dataset- and class-specific knowledge as priors to guide the feature extraction. We conduct extensive experiments to compare our UniCell with state-of-the-art methods, and validate the effectiveness of the proposed DPM on four challenging benchmarks. With the promising results, we believe that our method has great potential for cell nucleus classification and can bring new insights to the community.

Acknowledgments

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 *ICCV*, 4005–4014.

Akyon, F. C.; Altinuc, S. O.; and Temizel, A. 2022. Slicing aided hyper inference and fine-tuning for small object detection. In *ICIP*, 966–970. IEEE.

Amgad, M.; Atteya, L. A.; Hussein, H.; Mohammed, K. H.; Hafiz, E.; Elsebaie, M. A.; Alhusseiny, A. M.; AlMoslemany, M. A.; Elmatboly, A. M.; Pappalardo, P. A.; et al. 2022. NuCLS: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer. *GigaScience*, 11: giac037.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

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

Chen, S.; Ding, C.; Liu, M.; Cheng, J.; and Tao, D. 2023. CPP-net: Context-aware polygon proposal network for nucleus segmentation. *IEEE Transactions on Image Processing*, 32: 980–994.

Cutler, K. J.; Stringer, C.; Lo, T. W.; Rappez, L.; Stroustrup, N.; Brook Peterson, S.; Wiggins, P. A.; and Mougous, J. D. 2022. Omnipose: a high-precision morphology-independent solution for bacterial cell segmentation. *Nature Methods*, 19(11): 1438–1448.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. IEEE.

Doan, T. N.; Song, B.; Vuong, T. T.; Kim, K.; and Kwak, J. T. 2022. SONNET: A self-guided ordinal regression neural network for segmentation and classification of nuclei in large-scale multi-tissue histology images. *IEEE Journal of Biomedical and Health Informatics*, 26(7): 3218–3228.

Fleming, M.; Ravula, S.; Tatishchev, S. F.; and Wang, H. L. 2012. Colorectal carcinoma: Pathologic aspects. *Journal of Gastrointestinal Oncology*, 3(3): 153.

Fridman, W. H.; Pagès, F.; Sautès-Fridman, C.; and Galon, J. 2012. The immune contexture in human tumours: impact on clinical outcome. *Nature Reviews Cancer*, 12(4): 298–306.

Gamper, J.; and Rajpoot, N. 2021. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *CVPR*, 16549–16559.

Graham, S.; Jahanifar, M.; Azam, A.; Nimir, M.; Tsang, Y.-W.; Dodd, K.; Hero, E.; Sahota, H.; Tank, A.; Benes, K.; et al. 2021. Lizard: a large-scale dataset for colonic nuclear instance segmentation and classification. In *ICCV*, 684–693.

Graham, S.; Vu, Q. D.; Jahanifar, M.; Raza, S. E. A.; Minhas, F.; Snead, D.; and Rajpoot, N. 2023. One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification. *Medical Image Analysis*, 83: 102685.

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.

Huang, H.; Feng, X.; Jiang, J.; Chen, P.; and Zhou, S. 2022. Mask RCNN algorithm for nuclei detection on breast cancer histopathological images. *International Journal of Imaging Systems and Technology*, 32(1): 209–217.

Huang, J.; Li, H.; Sun, W.; Wan, X.; and Li, G. 2023a. Prompt-based grouping transformer for nucleus detection and classification. In *MICCAI*, 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.

Jain, J.; Li, J.; Chiu, M. T.; Hassani, A.; Orlov, N.; and Shi, H. 2023. Oneformer: One transformer to rule universal image segmentation. In *CVPR*, 2989–2998.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *ECCV*, 709–727. Springer.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *CVPR*, 19113–19122.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2): 83–97.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *ICCV*, 2980–2988.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 10012–10022.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *CVPR*, 11976–11986.

Lou, W.; Li, H.; Li, G.; Han, X.; and Wan, X. 2022. Which pixel to annotate: a label-efficient nuclei segmentation framework. *IEEE Transactions on Medical Imaging*, 42(4): 947–958.

Lou, W.; Wan, X.; Li, G.; Lou, X.; Li, C.; Gao, F.; and Li, H. 2023a. Structure Embedded Nucleus Classification for Histopathology Images. *arXiv preprint arXiv:2302.11416*.

Lou, W.; Yu, X.; Liu, C.; Wan, X.; Li, G.; Liu, S.; and Li, H. 2023b. Multi-stream Cell Segmentation with Low-level

Cues for Multi-modality Images. In *Competitions in Neural Information Processing Systems*, 1–10. PMLR.

Ma, J.; Xie, R.; Ayyadhury, S.; Ge, C.; Gupta, A.; Gupta, R.; Gu, S.; Zhang, Y.; Lee, G.; Kim, J.; et al. 2023. The Multi-modality Cell Segmentation Challenge: Towards Universal Solutions. *arXiv preprint arXiv:2308.05864*.

Malpica, N.; De Solórzano, C. O.; Vaquero, J. J.; Santos, A.; Vallcorba, I.; García-Sagredo, J. M.; and Del Pozo, F. 1997. Applying watershed algorithms to the segmentation of clustered nuclei. *Cytometry: The Journal of the International Society for Analytical Cytology*, 28(4): 289–297.

Mormont, R.; Geurts, P.; and Marée, R. 2020. Multi-task pre-training of deep neural networks for digital pathology. *IEEE Journal of Biomedical and Health Informatics*, 25(2): 412–421.

Plissiti, M. E.; Nikou, C.; and Charchanti, A. 2010. Automated detection of cell nuclei in pap smear images using morphological reconstruction and clustering. *IEEE Transactions on Information Technology in Biomedicine*, 15(2): 233–241.

Qin, J.; He, Y.; Ge, J.; and Liang, Y. 2022. A multi-task feature fusion model for cervical cell classification. *IEEE Journal of Biomedical and Health Informatics*, 26(9): 4668–4678.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241. Springer.

Ryu, J.; Puche, A. V.; Shin, J.; Park, S.; Brattoli, B.; Lee, J.; Jung, W.; Cho, S. I.; Paeng, K.; Ock, C.-Y.; et al. 2023. OCELOT: Overlapped Cell on Tissue Dataset for Histopathology. In *CVPR*, 23902–23912.

Saha, M.; Mukherjee, R.; and Chakraborty, C. 2016. Computer-aided diagnosis of breast cancer using cytological images: A systematic review. *Tissue and Cell*, 48(5): 461–474.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural machine translation of rare words with subword units. *ACL*.

Sirinukunwattana, K.; Raza, S. E. A.; Tsang, Y.-W.; Snead, D. R.; Cree, I. A.; and Rajpoot, N. M. 2016. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5): 1196–1206.

Stringer, C.; Wang, T.; Michaelos, M.; and Pachitariu, M. 2021. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, 18(1): 100–106.

Sugimoto, T.; Ito, H.; Teramoto, Y.; Yoshizawa, A.; and Bise, R. 2022. Multi-class cell detection using modified self-attention. In *CVPRW*, 1855–1863.

Tellez, D.; Höppener, D.; Verhoef, C.; Grünhagen, D.; Nierop, P.; Drozdzal, M.; Laak, J.; and Ciompi, F. 2020. Extending unsupervised neural image compression with supervised multitask learning. In *Medical Imaging with Deep Learning*, 770–783. PMLR.

Verma, R.; Kumar, N.; Patil, A.; Kurian, N. C.; Rane, S.; Graham, S.; Vu, Q. D.; Zwager, M.; Raza, S. E. A.; Rajpoot,

N.; et al. 2021. MoNuSAC2020: A multi-organ nuclei segmentation and classification challenge. *IEEE Transactions on Medical Imaging*, 40(12): 3413–3423.

Wang, X.; Yang, S.; Zhang, J.; Wang, M.; Zhang, J.; Huang, J.; Yang, W.; and Han, X. 2021. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *MICCAI*, 186–195. Springer.

Wang, Z.; Zhu, X.; Li, A.; Wang, Y.; Meng, G.; and Wang, M. 2022. Global and local attentional feature alignment for domain adaptive nuclei detection in histopathology images. *Artificial Intelligence in Medicine*, 132: 102341.

Wu, Y.; Cheng, M.; Huang, S.; Pei, Z.; Zuo, Y.; Liu, J.; Yang, K.; Zhu, Q.; Zhang, J.; Hong, H.; et al. 2022. Recent advances of deep learning for computational histopathology: principles and applications. *Cancers*, 14(5): 1199.

Xu, H.; Lu, C.; Berendt, R.; Jha, N.; and Mandal, M. 2016. Automatic nuclei detection based on generalized laplacian of gaussian filters. *IEEE Journal of Biomedical and Health Informatics*, 21(3): 826–837.

Yang, P.; Yin, X.; Lu, H.; Hu, Z.; Zhang, X.; Jiang, R.; and Lv, H. 2022. CS-CO: A Hybrid Self-Supervised Visual Representation Learning Method for H&E-stained Histopathological Images. *Medical Image Analysis*, 81: 102539.

Yu, X.; Li, G.; Lou, W.; Liu, S.; Wan, X.; Chen, Y.; and Li, H. 2023. Diffusion-based data augmentation for nuclei image segmentation. In *MICCAI*, 592–602. Springer.

Zeiser, F. A.; da Costa, C. A.; Roehe, A. V.; da Rosa Righi, R.; and Marques, N. M. C. 2021. Breast cancer intelligent analysis of histopathological data: A systematic review. *Applied Soft Computing*, 113: 107886.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2023a. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *ICLR*.

Zhang, W.; Zhang, J.; Wang, X.; Yang, S.; Huang, J.; Yang, W.; Wang, W.; and Han, X. 2023b. Merging nucleus datasets by correlation-based cross-training. *Medical Image Analysis*, 84: 102705.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *CVPR*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable detr: Deformable transformers for end-to-end object detection. *ICLR*.