CerDETR: Cell-Prior Empowered DETR for Cervical Lesion Detection

Linyun Zhou[1], Jin Chen[2], Weihan Li[2], Hengrui Lou[1], Lingxiang Jia[1⋆], Weijun Qin[3], Xiuming Zhang[4], and Zunlei Feng[2]

[1] State Key Laboratory of Blockchain and Data Security, Zhejiang University, China
[2] School of Software Technology, Zhejiang University, China
[3] EbTech Co. Ltd., China
[4] The First Affiliated Hospital, College of Medicine, Zhejiang University, China
{zhoulyaxx,chen_jin,leewh,hengrui,lingxiangjia}@zju.edu.cn, qinweijun99@163.com, {xm_zhang,zunleifeng}@zju.edu.cn

Abstract. Cervical cancer remains a leading cause of female morbidity and mortality, making early lesion detection critical. Although RCNN- and YOLO-based detectors have improved performance, they still rely on complex post-processing and are limited in global reasoning. DETR offers a new paradigm, but its direct application is challenged due to (1) large intra-class scale variation, (2) subtle inter-class differences, and (3) high inter-annotator variability. To address these challenges, we propose CerDETR, a DETR-based framework empowered by cell priors. It introduces a training-only Prior Corrector branch to capture intra-class scale and inter-class feature variations. It employs an Auto Multiscale Prior Generator (Auto MPG) to produce precise multiscale cell priors covering nuclei, cytoplasm, and clusters, an IoU and Contain-Guided (ICG) matching strategy for robust prior assignment under annotation noise, and a Prior Query Enhance module to leverage lesion characteristics and optimize prior embeddings for discriminative feature learning. Experiments on four public datasets show that CerDETR consistently outperforms RCNN-, YOLO-, and DETR-based methods, demonstrating its effectiveness, generalization, and clinical potential. Source code is available at https://github.com/imAzhou/CerDETR.

Keywords: Object detection · Cervical lesion screening · DETR

1 Introduction

As the fourth most common cancer worldwide, cervical cancer remains a major cause of female morbidity and mortality, making early detection crucial for prognosis and survival [6,28]. In recent years, deep learning have been widely applied to cervical lesion detection [13]. Methods based on RCNN [3,17,32] and YOLO-series [18,19,37] have achieved notable success, but they rely on complex post-processing and offer limited global reasoning. DETR-based detectors [2,34,38]

⋆ Corresponding author.

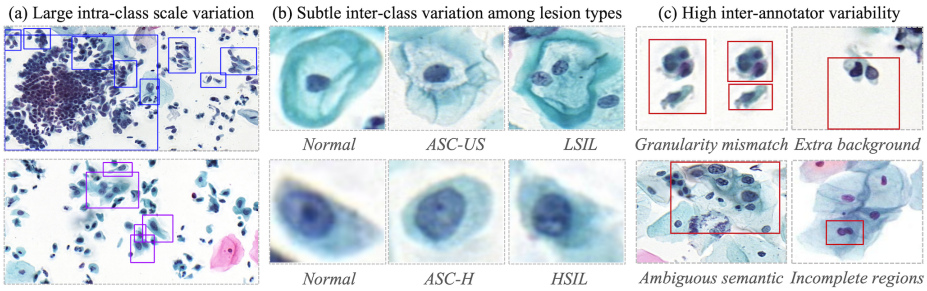

Fig. 1: The three key challenges constraining the performance of DETR-series detectors in cervical lesion detection.

have emerged as a new paradigm in object detection, reformulating it as a set prediction task that eliminates NMS while enhancing global context modeling. Recent studies [7, 15] have explored DETR-based detectors for cervical lesion detection, but their performance remains inferior to RCNN- and YOLO-series.

Directly applying DETR-series to this task faces three key challenges: (a) large intra-class scale variation, (b) subtle inter-class differences among lesion types, and (c) high inter-annotator variability due to subjectivity. As shown in Fig. 1(a), instances of the same class can span scales from single cells to large clusters, destabilizing global query matching in DETRs. Limited annotated data hinder the learning of fine-grained inter-class variations, reducing the model's ability to distinguish subtle visual differences among lesion types (Fig. 1(b)). Moreover, annotated data often contain inconsistencies and noise from annotator subjectivity, as illustrated in Fig. 1(c). These challenges motivate a dedicated DETR-based framework for cervical lesion detection, designed to improve matching stability, feature discrimination, and robustness to annotation noise.

In this paper, we present CerDETR, a novel DETR-based method tailored for cervical lesion detection. As shown in Fig. 2, a Prior Corrector branch is incorporated into the training pipeline of DETR-series models (e.g., DINO [34]). Acting as an auxiliary branch, it refines cell priors by jointly optimizing classification and localization. The branch comprises three key components: the Auto Multiscale Prior Generator (Auto MPG), the one-to-many (O2M) IoU and Contain-Guide (ICG) matching strategy, and the Prior Query Enhance module. The Auto MPG generates multiscale cell priors for nuclei, cytoplasm, and clusters without additional training, ensuring coverage of most cell instances. The ICG strategy establishes robust assignments between priors and ground truth (GT) boxes, mitigating the impact of ambiguous annotation boundaries and stabilizing detector training. The Prior Query Enhance module integrates lesion characteristics (e.g., category and scale) into prior query embeddings, improving capture of intra-class scale variations and inter-class feature differences. By jointly optimizing prior generation, matching, and embedding, the Prior Corrector branch substantially enhances DETR performance in cervical lesion detection. During

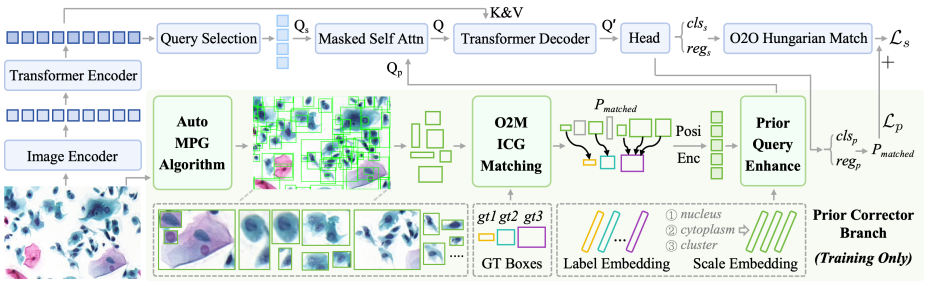

Fig. 2: Framework of the proposed CerDETR. The training-only Prior Corrector branch classifies each cell instance and refines its coordinates. The Auto MPG algorithm generates multiscale cell priors, which are assigned positive targets by the O2M ICG Matching strategy. The Prior Query Enhance module integrates query embeddings with the matched label and scale information of each prior.

inference, the auxiliary branch is removed, and only the main DETR branch predictions are used. The main contributions are summarized as follows:

(1) A novel Prior Corrector Branch is introduced as an auxiliary task for refining cell priors to guide the training of the main branch, effectively addressing the key challenges faced by DETR-based detectors in cervical lesion detection.

(2) A training-free Auto MPG algorithm is applied to generate multiscale cell priors, providing spatially precise and comprehensive coverage of cell instances.

(3) An ICG matching strategy is utilized to establish robust assignments between cell priors and GTs, alleviating ambiguous boundaries inherent to inter-annotator variability and stabilizing detector training.

(4) A Prior Query Enhance module is designed to embed lesion characteristics into prior queries, improving the modeling of intra-class scale variation and inter-class feature discrimination.

Extensive experiments on three cervical lesion datasets with distinct staining demonstrate that CerDETR consistently outperforms RCNN- and YOLO-series, achieving SOTA performance. Evaluation on a blood cell dataset further confirms its strong generalization, highlighting CerDETR's potential as a robust, high-performance tool for automated detection across diverse cellular analyses.

2 Related Work

Deep learning–based object detection has been widely applied to automated cervical screening, with numerous studies exploring ways to improve lesion localization efficiency and consistency [13]. Early methods [1, 20] were primarily based on RCNN two-stage frameworks, such as Faster R-CNN [24] and Sparse R-CNN [27], which have long served as strong baselines. Recent works have further refined RCNN-based models for cervical lesion detection [3,12,17,32]. While sequential cell localization and fine-grained classification remain effective, these methods still present inherent limitations. For instance, DPD-Net [3] uses dual-

path features with proposal relation modules and contrastive loss for abnormal cell detection, but cannot distinguish subtypes and requires nucleus center annotations, increasing annotation burden. DCC-MSI [17] and MA-Det [32] enhance multi-scale and spatial feature fusion but remain limited by the local receptive field of convolution, preventing capture of long-range dependencies.

With the rise of YOLO-based one-stage detectors [8,16,29–31], several methods have adopted YOLO models for cervical lesion detection to achieve faster inference and improved performance [18, 19, 37]. AFE-Net [18] introduces an adaptive feature module to mitigate imbalances caused by large shape and size variations in single cells and large clusters, while MLF-SNet [19] employs a hypergraph-based network to capture spatial correlations among visual features. However, both approaches remain constrained by the local receptive field of convolution, limiting their ability to differentiate abnormal subtypes. UC-YOLOX [37] integrates a Vision Transformer block into the backbone to capture long-range dependencies, yet its detection decoder still relies on convolution and linear layers, restricting its effectiveness in abnormal cell detection.

Since its introduction, DETR [2] has inspired extensive research in object detection [10,22,26,33,34,38]. It replaces heuristic anchors with set-based matching and leverages attention queries to capture global contextual dependencies, offering potential for modeling long-range relationships. Recently, some studies [7,15] have applied DETR-series detectors to cervical lesion detection, achieving preliminary progress. For example, to address DETR's convergence difficulties under limited training data, Fei *et al*. [7] propose synthetic data augmentation and train a separate classifier to compensate for DETR's limitations in abnormal cell subtype classification. GAA-DETR [15] designs an adaptive gaze kernel and a gaze-guided assignment module to accelerate DETR's focus on critical features. Nevertheless, their performance remains inferior to RCNN- and YOLO-series detectors. Subtle morphological differences among cervical lesion subtypes, coupled with limited annotated data and inter-observer noise, challenge DETR's training stability and its ability to capture fine-grained features.

This paper proposes a cell-prior empowered DETR framework, where a prior corrector branch guides training and helps DETR accurately classify lesion subtypes, outperforming conventional RCNN- and YOLO-series methods.

3 Method

This section first overviews the CerDETR framework (Sec. 3.1), followed by the Auto MPG algorithm (Sec. 3.2), the O2M ICG matching strategy (Sec. 3.3), and the Prior Query Enhance module (Sec. 3.4).

3.1 Framework Overview

This section describes the overall CerDETR framework. As shown in Fig. 2, the architecture incorporates an auxiliary Prior Corrector branch to enhance the main branch's learning. The main branch follows the standard DETR pipeline [34,

38]: the Encoder extracts image features $I_{\text{enc}} \in \mathbb{R}^{L\times d}$, where L and d denote the feature sequence length and feature dimension, respectively. I_{enc} is processed by the Query Selection module to generate standard queries $Q_s \in \mathbb{R}^{m\times d}$ for both training and inference, while prior queries $Q_p \in \mathbb{R}^{n\times d}$ from the Prior Corrector are concatenated with Q_s only during training. Here m and n denote the numbers of selected and prior queries, respectively. The combined set then undergoes Masked Self-Attention to produce the complete query set $Q \in \mathbb{R}^{(m+n)\times d}$:

$$Q = \text{SelfAttn}(Q_p, Q_s, \mathcal{M}), \quad \text{where } \mathcal{M} = \begin{pmatrix} 0 & 0 \\ \mathbf{1} & 0 \end{pmatrix}, \tag{1}$$

where the $\mathbf{1}$ block in mask $\mathcal{M}$ prevents Q_s from incorporating information from Q_p during self-attention, ensuring the main branch remains robust during inference when Q_p is absent. The complete query set Q is fed into the Transformer Decoder to perform cross-attention over I_{enc}, yielding the updated query embeddings Q'_s and Q'_p (jointly denoted as Q' in Fig. 2). Both of them are processed by a shared prediction head to produce classification (*cls*) and box regression (*reg*) outputs. The Q'_s outputs, cls_s and res_s, are matched via O2O Hungarian to compute the main loss $\mathcal{L}_s$, while the cls_p and res_p are paired with the ICG-matched priors P_{matched} to compute the auxiliary loss $\mathcal{L}_p$. The shared main branch weights $\mathcal{W}_{\text{main}}$ are then optimized via gradient backpropagation from both the main and auxiliary branches:

$$\text{Update}(\mathcal{W}_{\text{main}}) \propto \nabla_{\mathcal{W}_{\text{main}}}\mathcal{L}_s + \nabla_{\mathcal{W}_{\text{main}}}(\lambda\mathcal{L}_p), \tag{2}$$

where λ controls the impact of $\mathcal{L}_p$. Through shared transformer decoder and task head weights, the auxiliary loss $\mathcal{L}p$ contributes additional gradients for updating $\mathcal{W}_{\text{main}}$. During inference, the auxiliary Prior Corrector branch is discarded, and final predictions are produced by the main branch without NMS.

3.2 Prior Generating: Auto MPG Algorithm

This part presents the Auto Multiscale Prior Generator (Auto MPG), an algorithm built upon the foundation segmentation model Cellpose [23] to generate spatially precise and multiscale cell priors. Trained on large-scale heterogeneous cell datasets with a flow field and probability map prediction objective, Cellpose is capable of segmenting category-agnostic cell instances in a zero-shot manner. However, its performance in cervical cell segmentation remains limited due to substantial scale variation, particularly in large clustered cells. As shown in Fig. 3(b), the predicted probability map remains near zero for most clusters even when the cell diameter (the key parameter of Cellpose) is increased to 240px, whereas a well-defined flow field still emerges within these regions (Fig. 3(a)). This inconsistency motivates the development of the Auto MPG algorithm.

Specifically, the input image I is processed by Cellpose in a zero-shot manner to obtain the flow field, after which Auto MPG computes its divergence and magnitude to separately delineate cell boundaries and bodies. As described in Algorithm 1, given a target diameter (denoted as dia), Cellpose predicts the

Algorithm 1 Auto Multiscale Prior Generator

Require: Input Image I; Cellpose Model M; Configuration C (contains dia); Target Type $T \in \{\text{'nucleus'}, \text{'cytoplasm'}, \text{'cluster'}\}$

Ensure: Prior Bounding Boxes $Boxes$

$Config \leftarrow C[T]$; $\quad \mathbf{F} \leftarrow M.\text{PredictFlows}(I, Config.dia)$

$Divergence \leftarrow |Div(\mathbf{F})| = |\frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y}|$

$B_{mask} \leftarrow Div(\mathbf{F}) > \text{Percentile}(Divergence, \alpha)$

$Magnitude \leftarrow ||\mathbf{F}|| = \sqrt{F_x^2 + F_y^2}$

$P_{prob} \leftarrow \text{GaussianFilter}(\text{Normalize}(Magnitude))$; $\quad P_{prob}[B_{mask}] \leftarrow 0$

$$Mask \leftarrow \begin{cases} \text{FlowTracking}(\mathbf{F}, P_{prob}, Config), & \text{if } T \in \{\text{'nucleus'}, \text{'cytoplasm'}\} \\ \text{ConnectedComponents}(P_{prob}), & \text{otherwise} \end{cases}$$

$Boxes \leftarrow \text{GetBoundingBoxes}(Mask)$

return $Boxes$

flow field $\mathbf{F} = (F_x, F_y)$, where F_x and F_y denote the pixel-wise displacements along the x- and y-axes toward the cell center. The hypothesis is that greater divergence ($|\text{Div}(\mathbf{F})|$) indicates stronger boundary likelihood, whereas higher flow magnitude ($||\mathbf{F}||$) corresponds to higher cell-body probability. Accordingly, the boundary mask $B_{mask} \in \{0, 1\}$ is obtained from the divergence:

$$B_{mask} = \mathcal{F}_{\text{binary}} \left(\frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} \right), \tag{3}$$

where $\mathcal{F}_{\text{binary}}$ denotes the α-percentile binarization operator. And the cell probability $P_{prob} \in [0, 1]$ is derived from the magnitude:

$$P_{prob} = \mathcal{F}_{\text{smooth}}(\sqrt{F_x^2 + F_y^2}), \tag{4}$$

where $\mathcal{F}_{\text{smooth}}$ applies 99-percentile normalization [23] and Gaussian smoothing to produce a continuous probability map. Setting P_{prob} to zero at B_{mask} yields a sharp and well-isolated probability map. For nucleus and cytoplasm segmentation, the refined P_{prob} is fed into Cellpose's segmentation pipeline to improve boundary clarity, whereas for large clusters, instance masks are directly obtained via connected component analysis on P_{prob}. Fig. 3(b) shows that Auto MPG probability maps yield higher cell counts at smaller scales (nucleus and cytoplasm) and more distinctly separated predictions for large clusters, which is further confirmed by the mask outputs in Fig. 3(c). In practice, the number of masks varies across images, leading to different numbers of priors n. To facilitate batch-wise computation, the priors are post-processed to ensure a consistent n for each image. Details are provided in Appendix Sec. 1.2. The training dataset is preprocessed by Auto MPG to generate bounding boxes as proposal priors for the subsequent detection training pipeline.

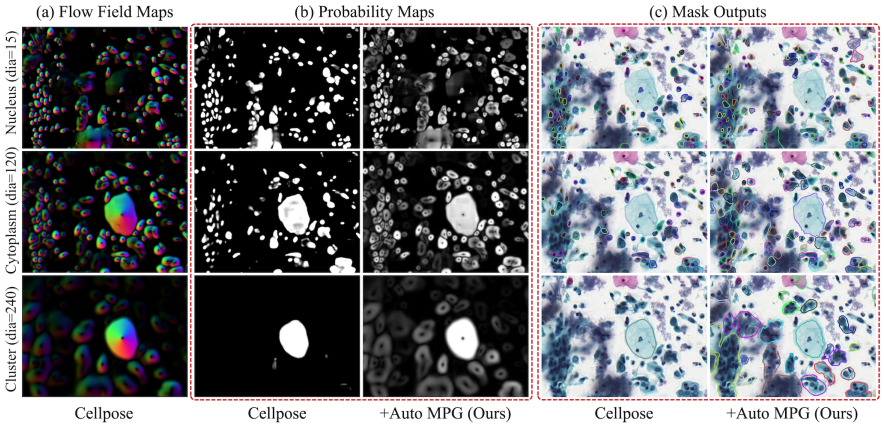

Fig. 3: Multiscale cervical cell segmentation comparison using Cellpose with and without the proposed Auto MPG algorithm. "dia" denotes cell diameter. Output masks are shown with randomly colored edges, each representing a distinct cell prior.

3.3 Prior Matching: O2M ICG Strategy

The One-to-Many (O2M) IoU and Contain-Guide (ICG) Matching Strategy aims to establish robust assignments between priors and GT boxes. The main branch adopts rigid One-to-One (O2O) Hungarian matching, enforcing a unique correspondence between each GT box and its best query, which enables end-to-end prediction without requiring NMS during inference. In contrast, the auxiliary Prior Corrector branch employs O2M ICG with flexible assignments, increasing positive sample density and alleviating the impact of annotation noise.

Specifically, ICG assigns each prior to GTs using IoU and containment cues. The maximum IoU score X_i for the i-th prior P_i is computed as:

$$X_i = \max_{j \in \{0,1,\ldots,K\}} \mathcal{I}(P_i, \mathrm{GT}_j), \quad i \in \{0, 1, \ldots, n\}, \tag{5}$$

where $\mathcal{I}$ denotes the IoU function, i and j index all n priors and all K GT boxes, respectively. The subsequent assignment strategy based on score X_i. Let δ_{l} and δ_{h} denote the low and high IoU thresholds. Priors with $X_i \geq \delta_{\mathrm{h}}$ are assigned as positives to the corresponding GT box, while those with $X_i < \delta_{\mathrm{l}}$ are negatives. Priors with $\delta_{\mathrm{l}} \leq X_i < \delta_{\mathrm{h}}$ follow the Contain-Guide rule, assigned to the smallest encompassing GT box. The matching result for the i-th prior P_i is given by:

$$P^i_{\mathrm{matched}} = \begin{cases} \arg\max_{\mathrm{GT}_j}(\mathcal{I}(P_i, \mathrm{GT}_j)) & \text{if } X_i \geq \delta_h \\ \arg\min_{\mathrm{GT}_j \in \mathbb{G}_C(P_i)} \mathcal{A}(\mathrm{GT}_j) & \text{if } \delta_l \leq X_i < \delta_h \\ \mathcal{N} & \text{if } X_i < \delta_l \end{cases}, \tag{6}$$

where P^i_{matched} denotes the matched result including box coordinates and label index for P_i, $\mathcal{N}$ denotes a negative sample with all entries negative, $\mathcal{A} = \log(w \times$

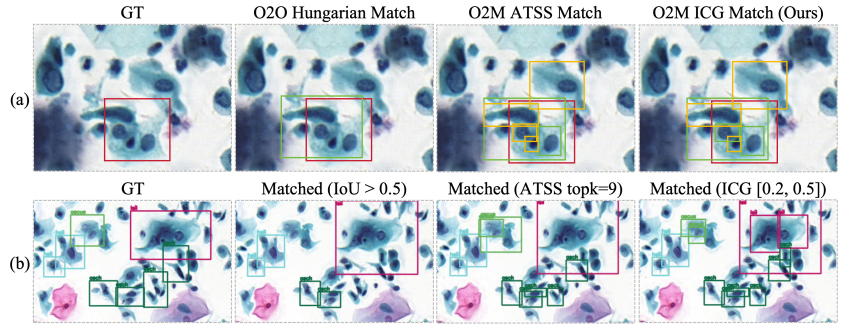

Fig. 4: (a) Conceptual illustration of O2O Hungarian, O2M ATSS, and O2M ICG matching. Green: priors matched to GT (red), yellow: unmatched negatives. (b) Compared with IoU-only and ATSS matching, ICG achieves higher GT recall.

h) represents the area of a box with width w and height h, and $\mathbb{G}_C(P_i)$ denotes the subset of GTs that contain P_i:

$$\mathbb{G}_C(P_i) = \{\mathrm{GT}_j \mid \mathcal{C}(P_i, \mathrm{GT}_j) = 1\}. \tag{7}$$

Here, $\mathcal{C} = 1$ if P_i is contained in GT_j. The Fig. 4 illustrates the differences among O2O Hungarian, O2M ATSS [35], and our O2M ICG matching. It clearly shows that ICG assigns GT boxes to more appropriate priors than Hungarian and ATSS, increasing positive prior coverage and robustness to annotation noise.

3.4 Prior Embedding: Query Enhancement

This section describes the Prior Query Enhance module, designed to improve the network's ability to handle intra-class scale variations and inter-class feature distinctions. The module is parameterized by two learnable embeddings: the Label Embedding $E_{\text{label}} \in \mathbb{R}^{C\times d}$ and the Scale Embedding $E_{\text{scale}} \in \mathbb{R}^{3\times d}$. The former encodes lesion types for C classes, including non-lesion regions, while the latter encodes three size categories: nucleus, cytoplasm, and cluster.

Leveraging the input prior P_i and its matched result P^i_{matched}, the characteristic embeddings for P_i are derived sequentially. Firstly, the prior's position embedding, $E^i_{\text{coord}} \in \mathbb{R}^{1\times d}$, is generated as:

$$E^i_{\text{coord}} = \mathcal{F}_{\text{PE}}(x_1, y_1, x_2, y_2), \tag{8}$$

where $\mathcal{F}_{\text{PE}}$ denotes the sinusoidal positional encoding, and (x_1, y_1), (x_2, y_2) are the normalized top-left and bottom-right coordinates of prior P_i. Next, the label embedding and scale embedding for prior P_i are obtained from P^i_{matched}. Recall that P^i_{matched} includes the matched label index $c \in \{0, 1, \ldots, C\}$ and box coordinates b_i. The label embedding E^i_{label} is indexed by c from E_{label}:

$$E^i_{\text{label}} = E_{\text{label}}[c], \quad E^i_{\text{label}} \in \mathbb{R}^{1\times d}. \tag{9}$$

The scale embedding E^i_{scale} is determined by the area of the matched box b_i:

$$\alpha_i = \mathcal{A}(b_i), \quad s = \begin{cases} 0, & \alpha_i < \theta_s, \\ 2, & \alpha_i > \theta_l, \\ 1, & \text{otherwise.} \end{cases} \quad E^i_{\text{scale}} = E_{\text{scale}}[s], \quad E^i_{\text{scale}} \in \mathbb{R}^{1\times d}. \tag{10}$$

Here, α_i denotes the area of box b_i, and θ_s and θ_l are predefined thresholds for nucleus and cluster scales. The scale index s is derived from α_i to select the corresponding scale embedding. For negative priors, α_i is computed from the coordinates of P_i. Finally, the enhanced prior query Q^i_p is obtained via element-wise summation of the position, label, and scale embeddings:

$$Q^i_p = E^i_{\text{coord}} + E^i_{\text{label}} + E^i_{\text{scale}}, \quad Q_p = \{Q^i_p\}^n_{i=1}. \tag{11}$$

This enhancement integrates spatial features with semantic and scale context, improving discrimination across lesion scales and categories.

4 Experiments

This section presents the dataset, metrics, and implementation details in Sec. 4.1, and reports quantitative, qualitative, and ablation analyses in Sec. 4.2.

4.1 Experiment Setup

Dataset. Experiments were conducted on three cervical lesion datasets with varying staining distributions and one blood cell dataset, all publicly available.

The *CDetector* dataset [21] contains 7,410 cervical cytological images of varying sizes, with 6,666 training and 744 test images. It includes 11 abnormal cell types, such as ASC-US, LSIL, and HSIL. The *HMCHH* dataset [36] comprises 8,037 images (2,048×2,048 px) labeled only as "abnormal" cells, with roughly two abnormal cells per image on average. The *CRIC* dataset [25] contains 400 Pap smear images (1,376×1,020 px) with annotations following DPD-Net [3], each labeled with a single "abnormal" class. The *BCCD* dataset [5] is a small-scale blood cell detection dataset with 360 images (640×480 px) and three labels: RBC, WBC, and Platelets. It is noted that *CDetector* serves as the representative dataset since it is the only multi-class cervical lesion detection benchmark. More details on dataset splits and cross-validation are in Appendix Sec. 1.1.

Metric. Evaluation metrics include mean Average Precision (AP) averaged over IoU thresholds from 0.5 to 0.95, along with AP_{50} and AP_{75} (IoU thresholds = 0.5 and 0.75) and mean Average Recall (AR). All the above metrics are reported in %. Model complexity and inference speed are measured by parameters (Param), GFLOPs, and Frames Per Second (FPS, it/s).

Implementation. CerDETR uses ResNet-50 [9] as the backbone and DINO [34] for the main detection branch. The diameter settings for the three scales are (15, 120, 240), with the prior count n in the auxiliary branch set to 300. The ICG Matching uses IoU thresholds (δ_l, δ_h) of 0.2 and 0.5. Implementation is based on MMDetection [4]. More details are provided in Appendix Sec. 1.2.

Table 1: Comparison results on three cervical lesion detection datasets. **Bold** and underline denote the best and second-best results, with green values indicating their margin. – denotes unavailable results due to missing source code, and * denotes task-specific methods, while the others are generic detectors.

	Method	Year	CDetector				CRIC				HMCHH			
			AP	AP_{50}	AP_{75}	*AR*	*AP*	AP_{50}	AP_{75}	*AR*	*AP*	AP_{50}	AP_{75}	*AR*
RCNN	Faster R-CNN [24]	2015	32.2	59.7	30.5	52.0	33.6	58.7	34.7	52.5	31.5	67.4	25.0	45.2
	Sparse R-CNN [27]	2021	34.7	63.4	34.2	60.6	30.6	53.8	32.1	52.8	23.2	50.1	19.0	32.0
	*DPD-Net [3]	2024	-	-	-	-	35.2	61.9	37.3	55.3	-	-	-	-
	*DCC-MSI [17]	2025	36.1	65.3	35.4	61.3	26.3	45.8	26.8	54.2	33.8	70.1	29.4	59.8
	*MA-Det [32]	2025	35.5	66.3	-	-	30.3	44.8	-	-	-	-	-	-
	*UniCAS [12]	2026	30.5	57.8	28.2	-	-	-	-	-	-	-	-	-
YOLO	YOLOv12-l [29]	2025	35.2	61.8	36.5	58.7	32.3	56.6	33.8	54.9	37.5	68.9	36.7	63.7
	YOLOv13-l [16]	2025	33.6	59.1	34.7	60.4	29.1	51.5	30.7	54.2	36.4	67.0	36.1	64.3
	*MLF-SNet [19]	2025	29.8	51.4	-	54.4	24.8	46.4	24.7	48.1	37.0	69.4	35.1	65.7
	*UC-YOLOX [37]	2025	-	-	-	-	-	-	-	-	32.9	64.0	29.6	55.6
DETR	DETR [2]	2020	23.2	55.7	15.4	-	23.5	41.6	23.0	52.6	19.4	47.1	12.2	36.5
	DeformDETR [38]	2021	29.2	57.0	26.7	53.2	29.8	50.5	31.0	61.7	29.3	65.3	20.3	58.4
	DINO [34]	2023	30.9	56.8	30.5	59.2	36.8	57.2	42.4	65.8	35.5	69.4	32.6	65.0
	*Fei *et al.* [7]	2024	24.6	44.7	23.6	46.6	-	-	-	-	-	-	-	-
	*GAA-DETR [15]	2025	15.8	34.0	11.5	56.7	28.8	49.2	28.8	62.9	35.8	70.5	33.0	63.5
	Mr. DETR [33]	2025	33.9	63.1	32.7	62.7	35.7	59.5	39.3	64.1	37.2	64.6	37.5	65.3
	MI-DETR [22]	2025	34.3	61.7	33.1	62.0	36.2	58.7	40.0	63.7	36.7	71.2	34.6	63.6
	DEIM [10]	2025	35.4	62.4	34.5	61.5	34.0	60.1	37.8	61.6	37.7	72.1	35.5	62.4
	*CerDETR (Ours)	-	**39.1** +3.0	**68.1** +1.8	**39.8** +3.3	**66.2** +3.5	**41.4** +4.6	**64.7** +2.8	**46.2** +3.8	**70.1** +4.3	**41.7** +4.0	**77.5** +5.4	**41.8** +4.3	**71.2** +5.5

4.2 Result Analysis

Experiments were conducted on three detector series: RCNN, YOLO, and DETR. For fair comparison, both task-specific methods for cervical lesion detection and established generic detectors are evaluated. Existing benchmark results are directly adopted from the original papers, and methods with missing results are re-implemented under identical settings when source code is available. Notably, DPD-Net [3] requires nucleus center-point annotations, which are unavailable in the CDetector and HMCHH datasets, making evaluation infeasible.

Quantitative Results The quantitative results on three cervical lesion detection datasets are summarized in Tab. 1. Overall, on the representative dataset *CDetector*, recent SOTA generic DETR detectors, such as Mr. DETR [33] and DEIM [10], achieve competitive performance. Yet, the RCNN-based DCC-MSI [17] remains the top performer. CerDETR, however, not only surpasses DINO [34] and Mr. DETR [33], which also enhance DETR with a training-only auxiliary branch, but also achieves the best overall performance across all three datasets. In addition, the comparison of inference efficiency and model generalization on the *CDetector* and *BCCD* datasets is shown in Tab. 2. As shown, although CerDETR is not the fastest, it achieves superior lesion recall as indicated by

Table 2: Comparison results of inference efficiency and model generalization.

Method	CDetector						BCCD			
	AP	*AR*	Backbone	Params	GFLOPs	FPS(it/s)	*AP*	AP_{50}	AP_{75}	*AR*
DCC-MSI	<u>36.1</u>	61.3	ResNet50	107.8M	124.9	26.6	57.3	85.9	62.6	68.8
MLF-SNet	29.8	54.4	yolo11-m	33.4M	104.0	32.4	55.4	82.0	63.4	67.1
YOLOv12-l	35.2	58.7	yolo12-l	**26.4M**	**89.5**	**82.7**	59.5	88.1	65.4	70.2
DINO	30.9	59.2	ResNet50	47.6M	226.5	20.1	61.3	<u>89.5</u>	68.5	74.0
GAA-DETR	15.8	56.7	ResNet50	41.6M	115.6	28.7	53.1	81.7	64.6	72.8
Mr. DETR	33.9	<u>62.7</u>	ResNet50	53.9M	281.7	14.6	62.7	86.7	68.3	<u>75.1</u>
MI-DETR	34.3	62.0	ResNet50	76.3M	337.9	13.3	60.5	87.8	66.3	73.3
DEIM	35.4	61.5	D-FINE-l	<u>30.7M</u>	<u>90.8</u>	<u>52.3</u>	<u>63.2</u>	88.7	<u>69.6</u>	72.4
CerDETR (Ours)	**39.1**	**66.2**	ResNet50	47.9M	226.5	20.1	**65.8**	**91.6**	**72.4**	**76.6**

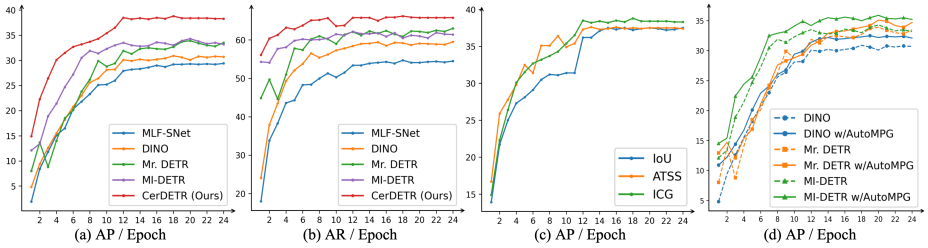

Fig. 5: Performance over 24 training epochs on *CDetector*. (a) *AP* and (b) *AR* values for strong baseline methods. (c) *AP* values with different matching strategies. (d) *AP* values for DETR methods with and without Auto MPG initialized priors.

AR, which is particularly important since practical lesion detection prioritizes minimizing missed lesions over inference speed. Generalization experiments on the *BCCD* dataset show that, although DETR-based detectors perform well, CerDETR demonstrates even stronger generalization to blood cell detection. Fig. 5(a) and (b) further show that CerDETR converges faster in terms of *AP* and maintains consistently high *AR* within only 24 training epochs. In addition, the convergence curves in Fig. 5(c) indicate that ICG yields a more stable training process compared with IoU and ATSS matching strategies. For fair comparison, Auto MPG generated prior queries are used to initialize three representative DETR-based detectors (DINO, Mr. DETR, and MI DETR). The results show that better query initialization consistently improves their performance (Fig. 5(d)). More quantitative results, including per-category performance on *CDetector* and AP decomposition analysis, are reported in Appendix Sec. 2.1.

Qualitative Results To verify whether the Label Embedding E_{label} and Scale Embedding E_{scale} effectively capture their intended semantics, their attention distributions with respect to I_{enc} were visualized separately. As shown in Fig. 6(a), the E_{label} corresponding to LSIL exhibits stronger attention around the spatial regions of its GT boxes (class label: LSIL), whereas the embedding for ASC-US

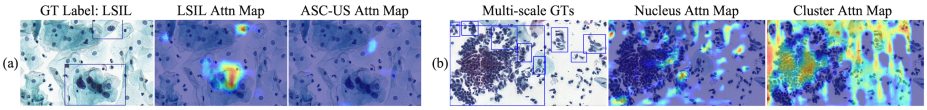

Fig. 6: Attention between label/scale embeddings and image features shows that: (a) GT-corresponding label embeddings yield sharper attention than non-corresponding ones, (b) nucleus-scale embeddings attend to smaller regions than cluster-scale ones.

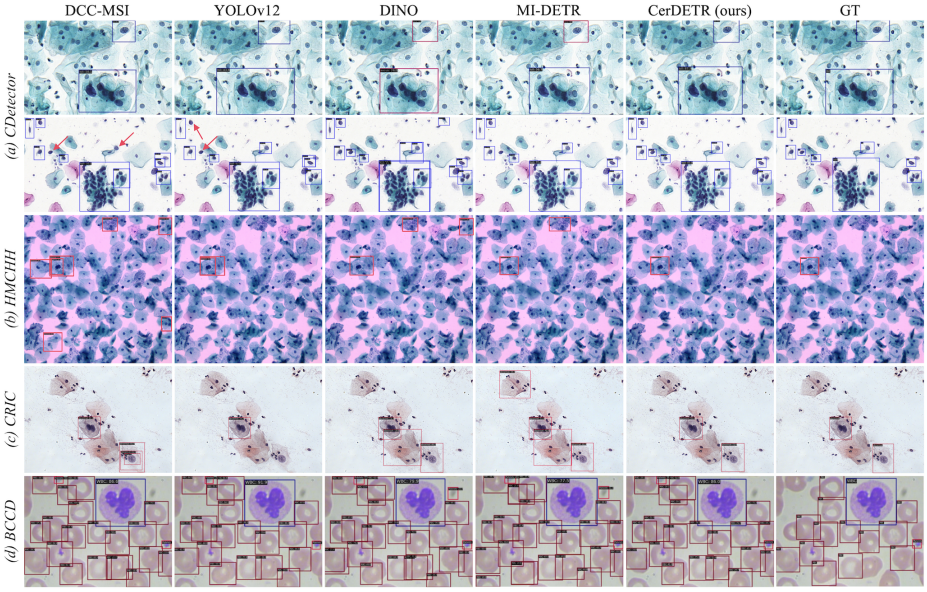

Fig. 7: Qualitative comparison of four SOTA across four datasets. Red arrows indicate lesions missed by the detectors.

does not display a similar pattern. In Fig. 6(b), even with the same category, attention maps vary with scale: embeddings for small scale nucleus exhibit compact attention regions, while those for large scale clusters distribute attention across nearly the entire image. It suggests that both E_{label} and E_{scale} have successfully learned their respective matching patterns. Moreover, Fig. 7 presents qualitative comparisons among four SOTA methods across four datasets. On *CDetector*, CerDETR shows superior performance in handling confounding categories (e.g., ASC-US and LSIL), as shown in the first row. In contrast, DCC-MSI and YOLOv12 often miss small lesions (red arrows, second row). Additional qualitative results are provided in Appendix Sec. 2.2.

Ablation Study To evaluate the performance of the proposed CerDETR, extensive ablation experiments are conducted on the representative *CDetector* dataset. For the ablation of individual key components, the DINO [34] detector is adopted as the baseline main branch. Specifically, a simplified Prior Corrector

Table 3: Ablation study of performance variations induced by the sequential integration of the proposed key modules.

	AP	AP_{50}	AP_{75}	AR
baseline	30.9	56.8	30.5	59.2
+ Prior Corrector	33.4	60.0	31.8	61.2
+ Auto MPG	37.2	66.3	36.7	63.9
+ ICG Matching	38.4	67.5	37.8	65.4
+ Query Enhance	**39.1**	**68.1**	**39.8**	**66.2**

Table 4: Ablation study on Auto MPG algorithm by replacing it with alternative prior generators.

Generator	AP	AP_{50}	AP_{75}	AR
Random	32.1	59.5	31.4	61.4
SAM [14]	35.8	64.3	35.8	63.6
CellSAM [11]	34.2	62.8	34.7	60.5
Cellpose [23]	36.7	64.9	35.3	63.9
+Auto MPG	**39.1**	**68.1**	**39.8**	**66.2**

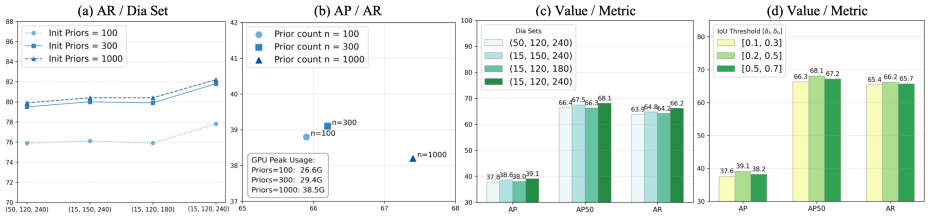

Fig. 8: Ablation experiments on key hyperparameters. (a) AR of initial priors on the training set under different Dia Sets. (b) Peak GPU usage during training and test set AP/AR for different prior counts n. (c) Performance under varying Dia Sets only. (d) Performance under varying IoU thresholds (δ_l, δ_h) only.

is first constructed by removing Auto MPG, ICG Matching, and the Prior Query Enhance Module. In this setting, prior boxes at three scales are directly generated by Cellpose [23], and samples are assigned solely via IoU matching (threshold = 0.5). The prior query embedding is composed only of the positional encoding E_{coord}. Subsequently, Auto MPG, ICG Matching, and the Prior Query Enhance module are incrementally introduced to assess their individual contributions. The ablation results in Tab. 3 demonstrate the effectiveness of each proposed key component, with Auto MPG yielding the largest performance gain.

To further investigate the contribution of Auto MPG, an ablation study is conducted by replacing it with alternative prior generators, including random initialization and zero-shot inference using SAM [14], CellSAM [11], and Cellpose [23]. All other modules in CerDETR are kept unchanged. These priors are then fed into the Prior Corrector Branch for training, and the results in Tab. 4 show that priors generated by Cellpose achieve higher performance, which is further improved by incorporating Auto MPG.

In addition, ablation experiments are conducted on key hyperparameters in CerDETR, including the diameter settings in Auto MPG, the prior count n in the Prior Corrector Branch, and the low/high IoU thresholds (δ_l, δ_h) used in the ICG Matching strategy. Fig. 8(a) reports the AR of initial priors generated by Auto MPG on the training set under four diameter settings (denoted as Dia Sets), based on the median values of GT boxes with minor perturbations. The setting (15, 120, 240) yields the highest AR regardless of the value of n. Fig. 8(b)

shows AP and AR on the test set for different n under the (15, 120, 240) setting. While $n = 1000$ yields the highest AR, it incurs higher training cost and slightly lower precision than $n = 300$ (GPU peak usage: 38.5 vs. 26.6; AP: 39.1 vs. 38.2). Fig. 8(c) and (d) illustrate the robustness of the hyperparameters Dia Sets and IoU thresholds, with fluctuations of approximately 1.5% in AP and 2% in AR.

Extended ablations in the Appendix cover ICG robustness under annotation noise (Sec. 2.2), convergence for different n values (Sec. 2.3), Dia Sets across scales (Sec. 2.4), and thresholds for Prior Scale Embedding θ_s and θ_l (Sec. 2.5).

5 Discussion and Future Work

Despite progress in cervical lesion detection, practical challenges remain. Existing annotations often contain subjective noise and sometimes group multiple cells into one box for convenience, introducing ambiguity inconsistent with actual lesion localization. More accurate cell segmentation and cell-level lesion labels could reduce this ambiguity and further enable Auto MPG, currently training-free because single-cell annotations are incomplete, to become a learnable prior generator. Moreover, existing cervical lesion detection datasets remain limited in scale, typically containing only a few thousand images, which constrains further exploration of model capacity. In the future, developing human-in-the-loop semi-automated annotation workflows based on existing public dataset annotations is expected to bring new promise to cervical lesion detection.

6 Conclusion

In this work, we present CerDETR, a training-only DETR framework for cervical lesion detection. It incorporates a Prior Corrector Branch that provides auxiliary guidance to the main branch during training and is removed at inference without any additional overhead. The branch consists of three key components: the Auto MPG algorithm, the O2M ICG matching strategy, and the Prior Query Enhancement module. Benefiting from these components, CerDETR effectively addresses large intra-class scale variation, subtle inter-class differences, and annotation noise. The quantitative analysis results indicate that CerDETR converges quickly and achieves strong performance overall. The qualitative analysis reveals that learnable lesion label and scale embeddings effectively capture their intended semantic patterns. Ablation studies emphasize that employing the ICG matching strategy mitigates annotation noise and stabilizes training. Moreover, leveraging high-quality cell priors and lesion-specific feature embeddings enables DETR to efficiently learn high-level semantic representations, significantly improving detection accuracy across diverse lesion types. Extensive experiments on three cervical lesion datasets show that CerDETR achieves SOTA performance among DETR-series models and surpasses strong RCNN- and YOLO-based baselines. Beyond this, it also shows strong potential in blood cell detection, showing broad applicability across cellular analyses.

Acknowledgements

This work is supported by Key R&D Program of Zhejiang (2025C02071), the Fundamental Research Funds for the Central Universities (226-2025-00240), Zhejiang Provincial Natural Science Foundation of China (Grant No. LQN26F 020006), and the advanced computing resources provided by the Supercomputing Center of Hangzhou City University.

References

1. Cao, L., Yang, J., Rong, Z., Li, L., Xia, B., You, C., Lou, G., Jiang, L., Du, C., Meng, H., et al.: A novel attention-guided convolutional network for the detection of abnormal cervical cells in cervical cancer screening. Medical image analysis **73**, 102197 (2021)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
3. Chai, S., Xin, J., Wu, J., Yu, H., Liang, Z., Ma, Y., Zheng, N.: Dpd-net: Dual-path proposal discriminative network for abnormal cell detection in cervical cytology images. Biomedical Signal Processing and Control **89**, 105887 (2024)
4. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
5. Cheng, S.: BCCD (Blood Cell Count and Detection) Dataset. `https://github.com/Shenggan/BCCD_Dataset` (2019), gitHub repository
6. Cohen, P.A., Jhingran, A., Oaknin, A., Denny, L.: Cervical cancer. The Lancet **393**(10167), 169–182 (2019)
7. Fei, M., Shen, Z., Song, Z., Wang, X., Cao, M., Yao, L., Zhao, X., Wang, Q., Zhang, L.: Distillation of multi-class cervical lesion cell detection via synthesis-aided pre-training and patch-level feature alignment. Neural Networks **178**, 106405 (2024)
8. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Huang, S., Lu, Z., Cun, X., Yu, Y., Zhou, X., Shen, X.: Deim: Detr with improved matching for fast convergence. In: Proceedings of the computer vision and pattern recognition conference. pp. 15162–15171 (2025)
11. Israel, U., Marks, M., Dilip, R., Li, Q., Yu, C., Laubscher, E., Iqbal, A., Pradhan, E., Ates, A., Abt, M., et al.: Cellsam: a foundation model for cell segmentation. BioRxiv pp. 2023–11 (2025)
12. Jiang, H., Cai, J., Shen, Z., Xu, M., Fei, M., Huang, H., Wang, X., Bi, R., Shen, D., Zhang, L., et al.: Unicas: A foundation model for cervical cytology screening. Cell Reports Medicine **7**(1) (2026)
13. Jiang, P., Li, X., Shen, H., Chen, Y., Wang, L., Chen, H., Feng, J., Liu, J.: A systematic review of deep learning-based cervical cytology screening: from cell identification to whole slide image analysis. Artificial Intelligence Review **56**(Suppl 2), 2687–2758 (2023)

14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
15. Kong, Y., Peng, Z., Yin, Y., Li, Y., Cai, J., Wang, S., Wang, Q., Fang, Y., Shan, C.: Query-level alignment for end-to-end lesion detection with human gaze. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 496–506. Springer (2025)
16. Lei, M., Li, S., Wu, Y., Hu, H., Zhou, Y., Zheng, X., Ding, G., Du, S., Wu, Z., Gao, Y.: Yolov13: Real-time object detection with hypergraph-enhanced adaptive visual perception. arXiv preprint arXiv:2506.17733 (2025)
17. Li, G., Fan, X., Xu, C., Lv, P., Wang, R., Ruan, Z., Zhou, Z., Zhang, Y.: Detection of cervical cell based on multi-scale spatial information. Scientific Reports **15**(1), 3117 (2025)
18. Li, G., Li, X., Wang, Y., Gong, S., Yang, Y., Xu, C.: Detection of cervical lesion cell/clumps based on adaptive feature extraction. Bioengineering **11**(7), 686 (2024)
19. Li, J., Dong, D., Zheng, M., Zhang, J., Hang, Y., Zhang, L., Zhao, L.: High-precision mixed feature fusion network using hypergraph computation for cervical abnormal cell detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 250–259. Springer (2025)
20. Li, X., Li, Q., et al.: Detection and classification of cervical exfoliated cells based on faster r-cnn. In: 2019 IEEE 11th international conference on advanced infocomm technology (ICAIT). pp. 52–57. IEEE (2019)
21. Liang, Y., Tang, Z., Yan, M., Chen, J., Liu, Q., Xiang, Y.: Comparison detector for cervical cell/clumps detection in the limited data scenario. Neurocomputing **437**, 195–205 (2021)
22. Nan, Z., Li, X., Dai, J., Xiang, T.: Mi-detr: an object detection model with multi-time inquiries mechanism. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4703–4712 (2025)
23. Pachitariu, M., Rariden, M., Stringer, C.: Cellpose-sam: superhuman generalization for cellular segmentation. bioRxiv pp. 2025–04 (2025)
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
25. Rezende, M.T., Silva, R., Bernardo, F.d.O., Tobias, A.H., Oliveira, P.H., Machado, T.M., Costa, C.S., Medeiros, F.N., Ushizima, D.M., Carneiro, C.M., et al.: Cric searchable image database as a public platform for conventional pap smear cytology data. Scientific data **8**(1), 151 (2021)
26. Shehzadi, T., Hashmi, K.A., Liwicki, M., Stricker, D., Afzal, M.Z.: Object detection with transformers: A review. Sensors **25**(19), 6025 (2025)
27. Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al.: Sparse r-cnn: End-to-end object detection with learnable proposals. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14454–14463 (2021)
28. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)
29. Tian, Y., Ye, Q., Doermann, D.: Yolov12: Attention-centric real-time object detectors. arXiv preprint arXiv:2502.12524 (2025)

30. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al.: Yolov10: Real-time end-to-end object detection. Advances in Neural Information Processing Systems **37**, 107984–108011 (2024)
31. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7464–7475 (2023)
32. Yin, Z., Huang, Q., Chen, Y., Lu, H.: Ma-det: A discriminative morphology-aware detector for cervical lesion cell clumps. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
33. Zhang, C.B., Zhong, Y., Han, K.: Mr. detr: Instructive multi-route training for detection transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9933–9943 (2025)
34. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)
35. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9759–9768 (2020)
36. Zhang, X., Ji, J., Zhang, Q., Zheng, X., Ge, K., Hua, M., Cao, L., Wang, L.: A large annotated cervical cytology images dataset for ai models to aid cervical cancer screening. Scientific Data **12**(1), 23 (2025)
37. Zheng, H., Liu, X., Ma, D.: Uc-yolox: Enhancing urothelial carcinoma detection with an improved yolox architecture leveraging attention mechanisms. Biomedical Signal Processing and Control **110**, 108259 (2025)
38. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)