Neurocomputing 710 (2027) 135398

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

CytoSet: Bethesda-order-aware set prediction for cervical cytology cell detection

Cheng He[a,1], Fufeng Chen[b,e,1], Fuhai Chen[a,*], Weiwei Guo[a], Houbing Zheng[d,2,*], Chunxu Yang[a], Zhiwei Chen[c], Qinqin Zhou[a], Bin Fang[e]

[a] School of Computer and Data Science, Fuzhou University, Fuzhou, 350108, China
[b] Department of Automation, Xiamen University, Xiamen, 361102, China
[c] School of Artificial Intelligence, Nanchang University, Nanchang, 330031, China
[d] Department of Plastic and Cosmetic Surgery, The First Affiliated Hospital of Fujian Medical University, Fuzhou, 350004, China
[e] Fuzhou Zhizhen Medical Technology Co., Ltd., Fuzhou, 350207, China

ARTICLE INFO

Communicated by N. Zeng

Keywords:
Bethesda system
Cervical cytology
Ordered representation learning
Object detection

ABSTRACT

Squamous cell detection in cervical ThinPrep Cytologic Test (TCT) images requires accurate localization and Bethesda category assignment. We represent the five squamous categories by an operational reporting order, yet standard detectors treat them as unordered labels and apply uniform classification objectives. This formulation ignores both crossing reporting boundaries and error direction: morphological overlap makes adjacent-grade confusion common, whereas distant or high-to-low errors may reduce the high-grade cellular evidence presented for downstream review. For set-prediction detectors, ordered supervision also depends on whether the matched queries provide reliable semantic representations and sufficient spatial coverage of candidate cells. We propose CytoSet, a Bethesda-order-aware adaptation of RT-DETR that conditions the query set before applying ordered learning. Dual-codebook semantic refinement encourages distinct abnormal-cell-related and non-object feature roles, and spatial grouping promotes coarse-support diversity in the initial query pool. Ordered Center Alignment and Quality Coupling (OCAQ) then aligns matched queries with ordered class centers and transfers center consistency to localization-aware classification quality. On the five-class TCT-MS benchmark, CytoSet achieves 39.08 ± 0.31 mAP, outperforming RT-DETR by 1.92 points. After localization-controlled matching, it reduces the severe cross-grade error rate from 7.01 ± 0.29 to 5.78 ± 0.24. In exploratory detector-derived whole-slide image (WSI) analysis, high-risk downgrading decreases from 20.3% to 14.5% on the internal cohort. These results show that conditioning detector queries before Bethesda-order-aware supervision improves fine-grained cytological detection and detector-derived high-grade evidence retention in exploratory WSI analysis.

1. Introduction

Cervical cytology is a central component of cervical cancer screening, and the ThinPrep Cytologic Test (TCT) is widely used to identify abnormal cells and precancerous lesions under the Bethesda System [1]. In clinical practice, cytopathologists inspect gigapixel whole-slide images (WSIs) to find diagnostically relevant abnormal cells. This process is time-consuming and subject to inter-observer variability, especially when abnormal cells are sparse, overlapping, morphologically ambiguous, or mixed with complex cytological background [2–5]. Automated cell-level detection can therefore localize suspicious cells and provide interpretable cellular evidence for WSI-level Bethesda review.

Slide-level and cell-level systems address complementary aspects of cervical cytology analysis. Many recent cervical cytology AI systems and broader pathology WSI models target slide-level endpoints, such as negative/positive classification, abnormal-cell screening, or coarse risk stratification [6–12]. These models provide a global slide assessment, whereas cell-level detection localizes and grades individual abnormalities that can be reviewed within positive slides. This work focuses on the latter task: detecting abnormal squamous cells and

* Corresponding authors.
Email addresses: cfh3c@fzu.edu.cn (F. Chen), zhenghoubing@fjmu.edu.cn (H. Zheng).
[1] These authors contributed equally to this work.
[2] Department of Plastic and Cosmetic Surgery, National Regional Medical Center, Binhai Campus of the First Affiliated Hospital, Fujian Medical University, Fuzhou, 350200, China.

https://doi.org/10.1016/j.neucom.2026.135398
Received 2 September 2026; Received in revised form 30 September 2026; Accepted 12 October 2026
Available online 13 October 2026
0925-2312/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

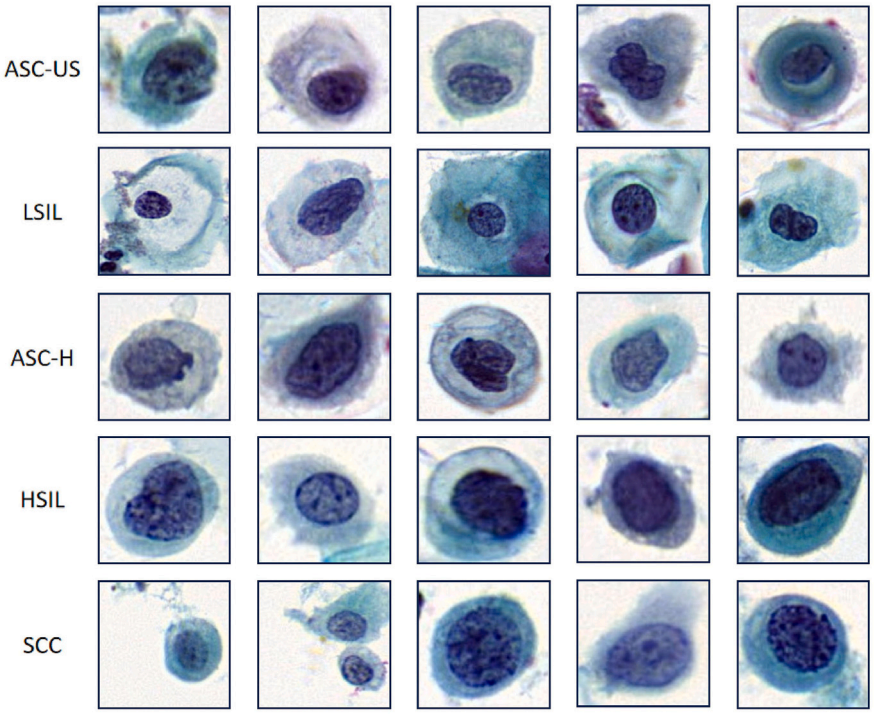


Fig. 1. Representative TCT cell examples from the five squamous Bethesda categories, ordered as ASC-US, LSIL, ASC-H, HSIL, and SCC. The matrix shows substantial intra-class variation, adjacent-grade similarity, and weak adjacent-grade boundaries in fine-grained cervical cytology cell detection.

assigning fine-grained Bethesda categories as traceable cellular evidence for downstream review.

This cell-level task poses requirements beyond generic object detection. For this study, we use the operational Bethesda reporting order ASC-US, LSIL, ASC-H, HSIL, and SCC [1]; ASC-H is placed between LSIL and HSIL because it denotes atypia for which HSIL cannot be excluded without meeting the criteria for a definitive HSIL interpretation. Rank distance therefore counts crossed reporting boundaries rather than equal clinical-risk or biological-progression intervals. Standard detectors instead treat these categories as unordered labels and do not distinguish the direction or extent of cross-grade errors. Morphological overlap makes adjacent-grade confusion frequent, whereas errors spanning several reporting boundaries represent larger ordinal deviations; high-to-low errors may additionally reduce the high-grade cellular evidence available for review. As shown in Fig. 1, weak inter-category boundaries coexist with intra-category variation. Consequently, higher aggregate detection accuracy does not necessarily imply more consistent Bethesda assignment after localization.

DETR-like detectors provide a natural query-level interface for imposing ordered supervision on one-to-one matched object representations [13]. Unlike image-level classification, detection applies category supervision only after candidate generation, query initialization, and bipartite matching. The effectiveness of ordered supervision therefore depends on the semantic reliability and spatial coverage of the upstream query pool: semantically unreliable or spatially concentrated candidates can limit the object representations that reach matching. We accordingly condition the query set before applying Bethesda-order-aware learning.

Built upon RT-DETR [14], CytoSet follows a sequential design from query-set conditioning to Bethesda-order-aware query learning. Before decoding, Dual-Codebook Semantic Refinement (DCSR) refines high-level features through weakly supervised dual codebooks, and Spatial-Grouped Query Selector (SGQS) introduces a spatial diversity prior during query initialization. Ordered Center Alignment and Quality Coupling (OCAQ), the core task-specific component, then organizes matched queries around ordered Bethesda class centers and transfers center consistency to localization-aware classification quality. We validate CytoSet through cell-level detection, localization-controlled Bethesda consistency, and WSI-level evidence analysis.

The main contributions are summarized as follows:

- We formulate fine-grained squamous TCT cell detection as Bethesda-order-aware set prediction and evaluate it with localization-controlled metrics that distinguish the extent and direction of cross-grade errors after successful localization.
- We propose a two-stage framework that conditions on high-level semantics and coarse-support diversity in the initial query set before applying its core Bethesda-order-aware query learning through OCAQ.
- We validate the framework through cell-level detection, localization-controlled Bethesda consistency, cross-cohort and external evaluation, and exploratory detector-derived WSI evidence analysis.

2. Related work

2.1. Automated cervical cell detection

Automated cervical cytology analysis has evolved from morphology-based segmentation of overlapping Pap smear cells [3,4] to deep learning systems for TCT patches cropped from WSIs [15–21]. These systems include CNN-based models, feature-pyramid detectors, YOLO-style detectors, and transformer architectures, and have substantially advanced abnormal-cell localization and recognition. Recent large-scale and hierarchical datasets further highlight the fine-grained structure of cervical cell classification [22]. However, evaluations remain dominated by AP, accuracy, or coarse abnormality recognition, and most detectors retain unordered category supervision [23,24]. Existing results therefore do not establish whether improved detection also yields more order-consistent Bethesda assignment after localization.

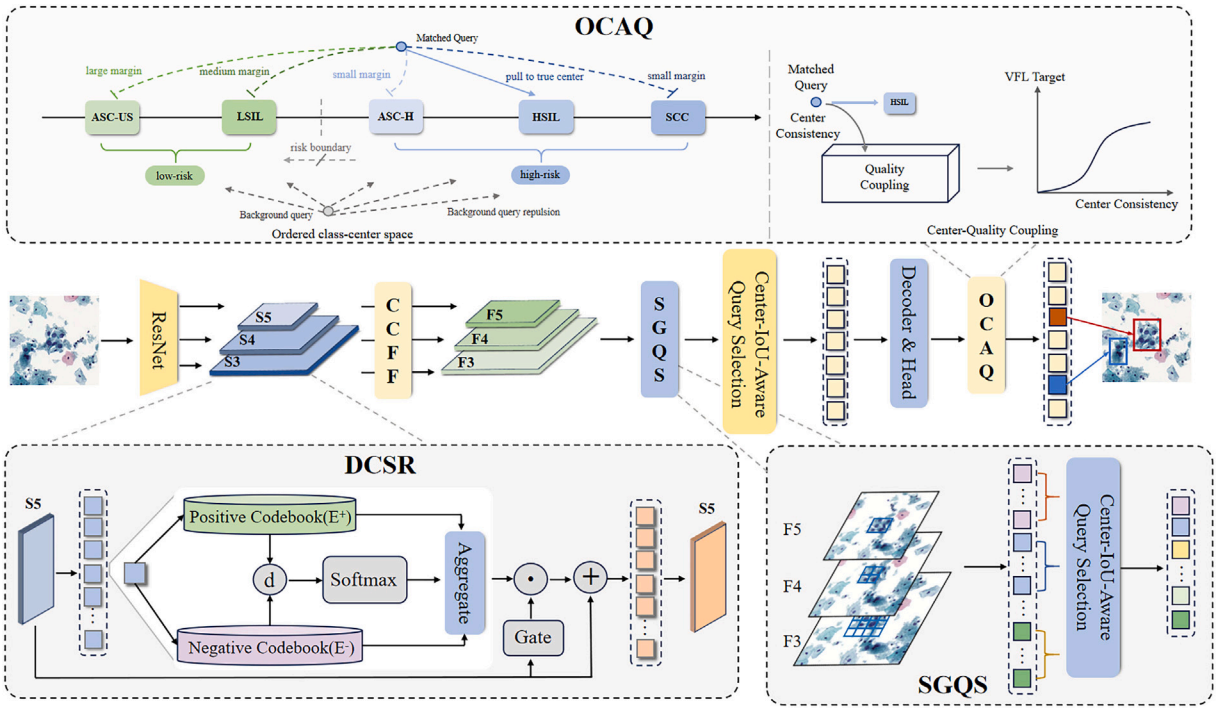


Fig. 2. Overall architecture of CytoSet. DCSR and SGQS condition the semantic reliability and spatial diversity of the query set before decoding, after which OCAQ aligns matched queries in an ordered Bethesda center space and couples center consistency with classification quality.

2.2. Query selection in DETR-like detectors

DETR-like detectors formulate object detection as end-to-end set prediction with one-to-one Hungarian assignment [13]. Later variants improve proposal generation, convergence, and decoder refinement through deformable attention, encoder-generated proposals, dense multi-scale supervision, denoising training, and query refinement or pruning [14,25–29]. At the proposal-to-query interface, we focus on cross-scale candidates that share coarse spatial support and may become over-represented in global ranking. Because Hungarian assignment is applied only after query initialization, one-to-one matching does not prevent several high-ranked encoder proposals from concentrating on the same local evidence. The issue addressed by SGQS is therefore candidate-pool over-concentration before matching, rather than duplicate assignment itself.

2.3. Fine-grained representation and prototype learning

Multiview representation learning has also been explored in biomedical classification [30]. Fine-grained recognition methods, including metric learning, prototype learning, angular-margin learning, structure-aware nuclear representation learning, and order-aware grading, have been used to address small inter-class differences, ordered labels, and large intra-class variation [10,31–40]. Learned codebooks provide a complementary means of reconstructing structured feature patterns [41,42]. Prototype- and center-based objectives encourage compact class representations, while ordinal methods distinguish errors according to label order or distance. Conventional ordinal objectives primarily reweight output errors or structure image-level embeddings. Detection introduces two additional interfaces: category supervision acts on one-to-one matched object queries, and proposal confidence also estimates localization quality. An order-aware detector must therefore connect representation geometry to the quality target used for query selection. DCSR uses weak region supervision to encourage distinct abnormal-cell-related and non-object codebook roles, whereas OCAQ structures matched decoder queries and couples center consistency to quality-aware set prediction.

3. Method

3.1. Overall architecture

Fig. 2 shows the overall framework of CytoSet. Built upon RT-DETR [14], CytoSet organizes its adaptations into two sequential levels: reliable query-set construction before decoding and Bethesda-order-aware learning of matched decoder queries. The original Hungarian matching cost and box-regression pathway remain unchanged; the center-aware quality target affects proposal ranking and supervised classification rather than bipartite assignment or box regression.

Given an input patch, the backbone extracts three feature levels, denoted as $S3$, $S4$, and $S5$. DCSR refines projected $S5$ before the cross-scale feature fusion module (CCFF) fuses it with $S3$ and $S4$ into the multi-scale encoder memory. This places semantic refinement before cross-scale aggregation while leaving the lower-level branches unchanged.

On top of the encoder memory, the proposal head predicts candidate scores at all spatial locations. Instead of directly selecting global top-K candidates, SGQS groups proposals sharing the same coarse $S5$ spatial support, retains at most one representative per group, and performs top-K selection on the reduced set. Together, DCSR and SGQS condition the semantic reliability and spatial diversity of the query set entering the decoder. OCAQ then aligns the resulting matched queries in a normalized ordered Bethesda representation space. Its Center-IoU-Aware Query Selection (CIQS) extends the IoU-aware query selection of RT-DETR by incorporating query–center consistency into the quality target used to rank encoder proposals.

3.2. Dual-codebook semantic refinement

Dual-Codebook Semantic Refinement (DCSR) is an encoder-side semantic adaptation module applied to projected $S5$. It introduces two jointly assigned codebooks whose roles are weakly supervised:

the positive branch is encouraged to capture abnormal-cell-related patterns, whereas the negative branch is encouraged to capture non-object patterns. Their differentiable reconstruction refines the main feature stream, with auxiliary commitment, utilization, and separation objectives used to discourage degenerate codebook use.

For each projected token $\mathbf{x}_i \in \mathbb{R}^C$, DCSR computes $\mathbf{u}_i = W_{\text{in}}\mathbf{x}_i$ and $\bar{\mathbf{u}}_i = \text{Norm}(\mathbf{u}_i)$. The positive and negative codebooks are concatenated as $\mathbf{E} = [\mathbf{E}^+; \mathbf{E}^-]$, with $\bar{\mathbf{e}}_k = \text{Norm}(\mathbf{e}_k)$. Each token is jointly assigned over all $K_+ + K_-$ codewords using

$$d_{i,k} = 2 - 2\langle \bar{\mathbf{u}}_i, \bar{\mathbf{e}}_k \rangle, \quad a_{i,k} = \frac{\exp(-d_{i,k}/\tau)}{\sum_{j=1}^{K_+ + K_-} \exp(-d_{i,j}/\tau)}. \tag{1}$$

The reconstruction $\mathbf{r}_i = \sum_{k=1}^{K_+ + K_-} a_{i,k}\bar{\mathbf{e}}_k$ is injected through a gated residual path:

$$\mathbf{y}_i = \mathbf{x}_i + g(\mathbf{x}_i) \odot W_{\text{out}}\mathbf{r}_i. \tag{2}$$

This residual design preserves the original feature stream while allowing codebook-guided correction. All spatial tokens attend jointly to both codebooks; region labels are used only by the auxiliary separation loss during training.

Let Ω^+ and Ω^- denote central object-core and non-object token sets. The object core reduces ambiguity from overlapping cytoplasm and imprecise box edges; non-object tokens lie outside all annotated boxes, and remaining tokens are ignored by $\mathcal{L}_{\text{sep}}$. Using $d_{i,k}^{\text{sep}} = 2 - 2\langle \text{sg}(\bar{\mathbf{u}}_i), \bar{\mathbf{e}}_k \rangle$, let d_i^{++} and d_i^{+-} be the nearest distances from a positive token to the positive and negative codebooks, with d_j^{--} and d_j^{-+} defined analogously for a non-object token. We set $\mathcal{L}_{\text{sep}} = \mathcal{L}_{\text{pull}} + \mathcal{L}_{\text{push}}$:

$$\mathcal{L}_{\text{pull}} = \frac{1}{|\Omega^+|}\sum_{i\in\Omega^+} d_i^{++} + \frac{1}{|\Omega^-|}\sum_{j\in\Omega^-} d_j^{--}. \tag{3}$$

$$\mathcal{L}_{\text{push}} = \frac{1}{|\Omega^+|}\sum_{i\in\Omega^+}\left[\beta d_i^{++} + \delta - d_i^{+-}\right]_+ + \frac{1}{|\Omega^-|}\sum_{j\in\Omega^-}\left[\beta d_j^{--} + \delta - d_j^{-+}\right]_+. \tag{4}$$

Following vector-quantized representation learning [41], let S denote all spatial tokens and $\text{sg}(\cdot)$ denote stop-gradient. VQ-style detached targets are used for commitment and utilization balancing while the reconstruction path remains differentiable. Specifically, $\mathbf{r}_i^{\text{cb}} = \sum_k \text{sg}(a_{i,k})\bar{\mathbf{e}}_k$. For utilization balancing, $a_{i,k}^{\text{div}}$ denotes the same soft assignment recomputed with $\text{sg}(\bar{\mathbf{u}}_i)$. Let $\mathcal{K}_b$ index branch $b \in \{+,-\}$. We define $\bar{p}_k^b = |S|^{-1}\sum_{i\in S} a_{i,k}^{\text{div}}$ and $\tilde{p}_k^b = (\bar{p}_k^b + \varepsilon)/\sum_{j\in\mathcal{K}_b}(\bar{p}_j^b + \varepsilon)$. The complete DCSR objective is

$$\begin{aligned}
\mathcal{L}_{\text{cb}} &= \frac{1}{|S|}\sum_i \|\text{sg}(\bar{\mathbf{u}}_i) - \mathbf{r}_i^{\text{cb}}\|_2^2, \\
\mathcal{L}_{\text{com}} &= \frac{1}{|S|}\sum_i \|\bar{\mathbf{u}}_i - \text{sg}(\mathbf{r}_i)\|_2^2, \\
\mathcal{L}_{\text{div}} &= \sum_{b\in\{+,-\}}\sum_{k\in\mathcal{K}_b} \tilde{p}_k^b \log\left(K_b\bar{p}_k^b + \varepsilon\right), \\
\mathcal{L}_{\text{DCSR}} &= \mathcal{L}_{\text{cb}} + \lambda_{\text{com}}\mathcal{L}_{\text{com}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}} + \lambda_{\text{sep}}\mathcal{L}_{\text{sep}}.
\end{aligned} \tag{5}$$

The commitment terms couple tokens and codewords, $\mathcal{L}_{\text{div}}$ balances codeword use within each branch, and $\mathcal{L}_{\text{sep}}$ is the only region-supervised term. The reconstruction remains jointly assigned across both codebooks and differentiable with respect to the detection objective.

3.3. Spatial-grouped query selector

In set prediction, multi-scale candidates sharing the same coarse spatial support can enter the global ranking independently and become over-represented in the initial query pool. Spatial-Grouped Query Selector (SGQS) imposes a spatial diversity prior by introducing local cross-scale competition before global top-K selection, allowing each coarse group to contribute one representative.

For each coarse location u on the $S5$ grid, SGQS collects the associated candidates from three feature scales:

$$\mathcal{G}_u = \{(5,u)\} \cup \{(4,v) \mid v \in \Phi_4(u)\} \cup \{(3,v) \mid v \in \Phi_3(u)\}, \tag{6}$$

where $\Phi_4(u)$ and $\Phi_3(u)$ contain the four $S4$ and sixteen $S3$ locations mapped to the same coarse support, respectively. Thus, each group contains one $S5$, four $S4$, and sixteen $S3$ candidates. For each candidate (s,v), the proposal score is $\rho_{s,v} = \max_c \ell_{s,v,c}$, where $\ell_{s,v,c}$ denotes the encoder proposal confidence supervised by the center-IoU-aware quality target. SGQS keeps the highest-scoring representative $g_u^\star = \arg\max_{(s,v)\in\mathcal{G}_u} \rho_{s,v}$ and applies score-guided top-K selection on the reduced set; hence, both within-group maximization and global ranking follow CIQS:

$$\mathcal{Q}_0 = \text{TopK}_K\left(\{\rho(g_u^\star)\}_{u\in S5}\right). \tag{7}$$

SGQS changes the candidate pool rather than the top-K rule. Restricting each coarse spatial group to one representative increases coarse-support diversity in the reduced pool. It supplies a spatial-coverage prior for query initialization rather than for ordinal supervision.

3.4. Ordered center alignment and quality coupling

Ordered Center Alignment and Quality Coupling (OCAQ) incorporates the operational order used in this study, ASC-US < LSIL < ASC-H < HSIL < SCC [1], into matched decoder-query supervision while retaining standard five-class outputs. The distance $|r(a) - r(b)|$ counts the number of crossed reporting boundaries rather than assuming equal clinical, morphological, or biological spacing.

OCAQ has two core blocks. Bethesda-ordered query alignment organizes matched queries around category centers and includes a one-way penalty for high-to-low assignments. Center-quality coupling then transfers center consistency to the detector's IoU-aware classification target. Category geometry and background separation serve as supporting regularization for the shared center space.

1) Bethesda-Ordered Query Alignment: Let $\mathcal{M}$ denote the index set of matched foreground queries. For supervised decoder layer ℓ, OCAQ obtains $\mathbf{z}_i^{(\ell)} = \text{Norm}(h_\phi(\mathbf{q}_i^{(\ell)})) \in \mathbb{R}^{64}$; the layer superscript is omitted below. Each class maintains a shared normalized EMA center, initialized when that class is first observed. If $\bar{\mathbf{z}}_c$ is the mean final-layer embedding of currently matched queries with label c, the center is updated as $\boldsymbol{\mu}_c \leftarrow \text{Norm}(m\boldsymbol{\mu}_c + (1-m)\bar{\mathbf{z}}_c)$. Let $\mathcal{C}_t$ denote the centers available at iteration t and $\mathcal{M}_t = \{i \in \mathcal{M} : y_i \in \mathcal{C}_t\}$. The projection and centers are shared across supervised decoder layers, whereas centers are updated only from final-layer matches.

The five categories have ranks $r(c) \in \{0,1,2,3,4\}$. For $c \in \mathcal{C}_t$, the similarity is $s_{i,c} = \langle \mathbf{z}_i, \boldsymbol{\mu}_c \rangle$, and $p_{i,c} = \text{softmax}_{c\in\mathcal{C}_t}(s_{i,c}/\tau_c)$. This center-space posterior is distinct from the detector classification posterior; they are coupled through the shared decoder representation and the center-modulated IoU-aware Varifocal Loss (VFL) target.

At the query level, OCAQ aligns each valid matched foreground query in $\mathcal{M}_t$ with the ordered category-center space using a rank-distance-aware margin:

$$\mathcal{L}_{\text{ord}} = \frac{1}{|\mathcal{M}_t|}\sum_{i\in\mathcal{M}_t}\left[(1 - s_{i,y_i}) + \frac{1}{\max(1,|\mathcal{C}_t|-1)}\sum_{c\in\mathcal{C}_t\setminus\{y_i\}}\left[m_0 + \frac{|r(y_i) - r(c)|}{R} - s_{i,y_i} + s_{i,c}\right]_+\right], \tag{8}$$

where $R = 4$. The first term attracts the query to its correct center, while the second increases the separation margin with the number of crossed category boundaries.

The margin-based objective constrains pairwise query-to-center relations. We additionally measure the expected crossed-boundary cost under the center-space posterior:

$$\mathcal{L}_{\text{cal}} = \frac{1}{|\mathcal{M}_t|} \sum_{i \in \mathcal{M}_t} \sum_{c \in \mathcal{C}_t} p_{i,c} \left| r(c) - r(y_i) \right|. \tag{9}$$

$\mathcal{L}_{\text{ord}}$ defines rank-dependent pairwise alignment, whereas $\mathcal{L}_{\text{cal}}$ applies the same crossed-boundary count to the full center posterior. Because the rank gap is absolute, posterior mass on either side of the target contributes positively rather than canceling as it could under a signed rank residual. Thus, $\mathcal{L}_{\text{cal}}$ complements the pairwise margins by constraining the complete center posterior.

Because high-to-low errors can reduce the high-grade cellular evidence available under the task definition, query alignment is complemented by direction-specific supervision for matched high-risk queries. Let

$$\begin{aligned} &\mathcal{C}_{\text{H}} = \{\text{ASC-H}, \text{HSIL}, \text{SCC}\}, \\ &\mathcal{C}_{\text{L}} = \{\text{ASC-US}, \text{LSIL}\}, \\ &\mathcal{M}_{\text{H},t} = \{i \in \mathcal{M}_t : y_i \in \mathcal{C}_{\text{H}}\} \end{aligned} \tag{10}$$

denote the high- and low-risk sets and the matched high-risk queries, respectively. The one-way downgrade regularizer is

$$\mathcal{L}_{\text{down}} = \frac{1}{|\mathcal{M}_{\text{H},t}|} \sum_{i \in \mathcal{M}_{\text{H},t}} \left[\max_{c \in \mathcal{C}_{\text{L}} \cap \mathcal{C}_t} s_{i,c} - \max_{c \in \mathcal{C}_{\text{H}} \cap \mathcal{C}_t} s_{i,c} \right]_+. \tag{11}$$

This term penalizes a high-risk query when its maximum low-risk-center similarity exceeds its maximum high-risk-center similarity.

2) Center-IoU-Aware Query Selection and Quality Coupling: RT-DETR performs IoU-aware query selection by ranking encoder proposals with classification scores whose positive targets incorporate localization quality [14]. This favors proposals that are both discriminative and well localized before top-K initialization, but it does not account for their consistency with the ordered category representation. We extend this mechanism through CIQS, using center consistency to modulate the IoU-aware quality target that supervises the proposal ranking score. For a matched foreground candidate, let $u_i = \text{IoU}(\hat{\mathbf{b}}_i, \mathbf{b}_i)$ and $c_i = (1 + s_{i,y_i})/2$. The center-IoU-aware positive target is

$$q_i^{\text{joint}} = \text{sg}\left[u_i \left((1-\eta) + \eta c_i\right)\right], \qquad \eta \in [0,1]. \tag{12}$$

The joint target is stop-gradient. Because $q_i^{\text{joint}} \le u_i$, IoU remains an upper bound on the positive classification target; center consistency can lower, but not raise, this target relative to u_i. The proposal output and each supervised decoder layer recompute this target using their own predicted boxes and query representations with a shared class-center bank.

Two auxiliary regularizers stabilize the shared center space. Query-to-center alignment does not explicitly constrain the relative geometry among category representations, so we impose an anchor-relative monotonic relation at the category level. At each supervised decoder layer, let $\mathcal{M}_c = \{i \in \mathcal{M}_t : y_i = c\}$ and $\mathcal{C}_B = \{c : |\mathcal{M}_c| > 0\}$. The differentiable category representative is

$$\mathbf{v}_c = \text{Norm}\left(\frac{1}{|\mathcal{M}_c|} \sum_{i \in \mathcal{M}_c} \mathbf{z}_i \right), \qquad c \in \mathcal{C}_B. \tag{13}$$

Define $d_{ab} = (1 - \langle \mathbf{v}_a, \mathbf{v}_b \rangle)/2$, $\Delta_{ab} = |r(a) - r(b)|$, and $\mathcal{T}_B = \{(a,b,c) \in \mathcal{C}_B^3 : a, b, c \text{ are distinct}, \Delta_{ab} < \Delta_{ac}\}$. The geometry objective is

$$\mathcal{L}_{\text{geo}} = \frac{1}{|\mathcal{T}_B|} \sum_{(a,b,c) \in \mathcal{T}_B} \left[d_{ab} - d_{ac} + m_{\text{geo}} \frac{\Delta_{ac} - \Delta_{ab}}{R} \right]_+. \tag{14}$$

For a fixed anchor, $\mathcal{L}_{\text{geo}}$ encourages a smaller representation distance for a smaller reporting-rank gap.

Unmatched queries $\mathcal{U}$ are additionally repelled from the foreground Bethesda center space:

$$\mathcal{L}_{\text{bg}} = \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} \left[\max_{c \in \mathcal{C}_t} s_{i,c} - \gamma_{\text{bg}} \right]_+. \tag{15}$$

3.5. Loss function

The training objective retains the RT-DETR box losses and Hungarian assignment, replaces the positive VFL target in supervised classification outputs with q_i^{joint}, and adds DCSR and OCAQ regularization. The detection loss is

$$\mathcal{L}_{\text{det}} = \lambda_{\text{vfl}} \mathcal{L}_{\text{vfl}}^{\text{joint}} + \lambda_{\text{box}} \mathcal{L}_1 + \lambda_{\text{giou}} \mathcal{L}_{\text{giou}} + \sum_{\ell \in \mathcal{A}} \mathcal{L}_{\text{aux}}^{(\ell)}, \tag{16}$$

where $\mathcal{L}_{\text{vfl}}^{\text{joint}}$ is supervised by q_i^{joint} using Varifocal Loss [43], $\mathcal{L}_1$ and $\mathcal{L}_{\text{giou}}$ are box regression losses, and $\mathcal{A}$ denotes the ordinary auxiliary decoder outputs. The classification term of each supervised auxiliary decoder layer uses the same center-modulated positive target.

The OCAQ regularization is

$$\mathcal{L}_{\text{OCAQ}} = \lambda_{\text{ord}} \mathcal{L}_{\text{ord}} + \lambda_{\text{cal}} \mathcal{L}_{\text{cal}} + \lambda_{\text{down}} \mathcal{L}_{\text{down}} + \lambda_{\text{geo}} \mathcal{L}_{\text{geo}} + \lambda_{\text{bg}} \mathcal{L}_{\text{bg}}. \tag{17}$$

It aggregates ordered alignment, asymmetric downgrade control, and supporting center-space regularization, and is summed over the supervised decoder layers; quality coupling enters the detector classification term through q_i^{joint}. The full training objective is

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_{\text{dcsr}} \mathcal{L}_{\text{DCSR}} + \mathcal{L}_{\text{OCAQ}}. \tag{18}$$

4. Experiments

To evaluate Bethesda-order-aware set prediction from cell-level detection to downstream evidence retention, we assess CytoSet at three progressive levels using a multi-scanner internal benchmark and two external datasets. First, **cell-level detection** measures standard localization and recognition accuracy against generic detectors. Second, **localization-controlled Bethesda consistency** conditions grade-error analysis on successful localization to evaluate severe grade drift, mean ordinal error, and high-risk downgrading. Finally, an exploratory **detector-derived WSI evidence-retention analysis** examines whether improved cell-level detection preserves high-grade cellular evidence for downstream slide review.

4.1. Ethics statement

This study was approved by the Ethics Committee of Fujian Maternity and Child Health Hospital (approval No. 2023KY045; April 27, 2023). The requirement for informed consent was waived by the Ethics Committee.

4.2. Implementation details

All images are resized to 640 × 640. CytoSet and RT-DETR use the official RT-DETR-R50vd 6× configuration and the same pretrained ResNet-50vd backbone. Both are trained with AdamW for 72 epochs (batch size 4; learning rates 10^{-4} and 10^{-5} for the detector and backbone) using the original schedule, EMA, and augmentation recipe; the highest-validation-mAP checkpoint from each run is evaluated. Competing detectors use official configurations and pretrained initializations with five-category prediction heads (Supplementary Table S5). All methods share the TCT-MS patient-level split, input resolution, and evaluation protocol. Supplementary Table S6 compares CytoSet and RT-DETR parameter counts and forward-inference efficiency.

Table 1
Dataset statistics. Tianchi is reported at the 40×-equivalent scale after 2× resizing from 20× WSIs.

Dataset	Mag.	WSIs	Train img.	Val img.	Test img.	Total img.	Cell boxes
0506	40×	214	10,158	1984	2056	14,198	14,838
0818	40×	231	10,734	1517	1595	13,846	16,564
TCT-MS	40×	445	**20,892**	**3501**	**3651**	**28,044**	**31,402**
Tianchi	40×	500	4948	1030	1005	6983	8635
CRIC	40×	–	2026	84	84	2194	4490

DCSR, SGQS, and OCAQ are optimized end-to-end with the RT-DETR objective. DCSR uses 64 codewords per branch, $\tau = 0.1$, $\beta = 1.2$, $\delta = 0.1$, and a central $30\% \times 30\%$ object-core region, with

$$(\lambda_{\text{com}}, \lambda_{\text{div}}, \lambda_{\text{sep}}, \lambda_{\text{dcsr}}) = (0.05, 0.1, 1.0, 0.1). \tag{19}$$

OCAQ projects 256-dimensional queries into 64-dimensional normalized embeddings and uses $m = 0.9$, $\tau_c = 0.1$, $\eta = 0.5$, $m_0 = 0.20$, $m_{\text{geo}} = 0.20$, and $\gamma_{\text{bg}} = 0.30$, with

$$(\lambda_{\text{ord}}, \lambda_{\text{cal}}, \lambda_{\text{down}}, \lambda_{\text{geo}}, \lambda_{\text{bg}}) = (0.5, 0.5, 0.2, 0.5, 0.1). \tag{20}$$

These values are selected on the TCT-MS validation split and fixed across the comparison, transfer, and ablation experiments.

For the Rank-VFL baseline, we retain the original RT-DETR varifocal loss and weight each non-target class of a matched query by $1 + |r(y_i) - r(c)|/R$; positive and unmatched-query terms remain unchanged. Rank-VFL therefore isolates output-level rank weighting without introducing class centers, query-space geometry, or center-quality coupling.

All experiments are conducted on one NVIDIA Tesla V100-SXM2 GPU with 32 GB memory. Main experiments are repeated with three random seeds, 42, 2026, and 3407, and are reported as mean ± standard deviation. Inference uses a single model without test-time augmentation or ensembling.

4.3. Datasets and splits

We construct the main benchmark from two independent internal TCT subsets, denoted as 0506 and 0818, acquired from different scanners (3DHISTECH Pannoramic SCAN and Jiangfeng KF-PRO-400, respectively) with no case overlap. Both subsets follow ThinPrep preparation with Papanicolaou staining and are scanned at 40× with 0.2438 micron-per-pixel resolution. Cell-level annotations were produced by nine cytopathologists, with no overlap in the slides assigned to different annotators. The chief cytopathologist conducted sample-based reviews; annotations that did not pass review were returned for re-review. The five foreground squamous categories are ASC-US, LSIL, ASC-H, HSIL, and SCC. For main comparison and ablation studies, the two subsets are merged as TCT-MS. Table 1 summarizes the dataset scale.

Patient-level splits are created independently for the two subsets and then merged to avoid slide-level leakage; Table 1 reports patch counts and Supplementary Section S9 reports WSI counts. Training uses 640×640 patches cropped around annotated cells with random offsets scaled to the available margins; background-only patches are excluded (Supplementary Section S10). Transfer and external evaluation use Tianchi [44] and CRIC [45]. Tianchi WSIs are acquired at 20×, resized by 2× to a 40×-equivalent scale, and split by WSI after cropping patches from 500 positive WSIs. The slides follow ThinPrep preparation with Papanicolaou staining; SCC is retained as an empty category in the five-class label space. CRIC provides conventional Pap smear images with cell-level annotations mapped to the same five squamous categories and is used to evaluate portability across preparation protocols.

4.4. Evaluation metrics

We report COCO-style mAP@[0.5:0.95], AP@0.5, and AP@0.75 [52]. On Tianchi, SCC has no ground-truth instances, and its AP is set to zero in the five-category mean. Matched SCC predictions retain their SCC rank and operational risk group in Bethesda consistency evaluation. Main TCT-MS comparisons and ablations are reported as mean ± std over three random seeds. We also provide class-wise, confusion-based, and WSI-level evidence analyses.

To evaluate Bethesda consistency after successful localization, we compute localization-controlled metrics using class-agnostic matching. All detectors are evaluated using their native inference outputs without additional confidence filtering. For each image, predictions are sorted by confidence and greedily matched one-to-one to ground-truth boxes by maximum IoU, with the same threshold $\theta = 0.50$ for all detectors. Let $\mathcal{M}_\theta = \{(g, p)\}$ denote the matched ground-truth–prediction pairs. The categories ASC-US, LSIL, ASC-H, HSIL, and SCC are assigned ordinal ranks from 0 to 4. For a matched pair (g, p) with labels y_g and y_p, $d(g, p) = |r(y_g) - r(y_p)|$ counts the crossed category boundaries.

These metrics condition Bethesda-grade evaluation on successful class-agnostic localization, reducing the confounding effect of missed localizations:

$$\text{LC-Acc} = \frac{1}{|\mathcal{M}_\theta|} \sum_{(g,p)\in\mathcal{M}_\theta} \mathbf{1}[y_g = y_p], \tag{21}$$

$$\text{SCR} = \frac{1}{|\mathcal{M}_\theta|} \sum_{(g,p)\in\mathcal{M}_\theta} \mathbf{1}[d(g, p) \geq 2], \tag{22}$$

$$\text{MOE} = \frac{1}{|\mathcal{M}_\theta|} \sum_{(g,p)\in\mathcal{M}_\theta} d(g, p). \tag{23}$$

Here, LC-Acc denotes localization-controlled accuracy, SCR denotes the severe cross-grade error rate, and MOE denotes mean ordinal error.

For category-level detection error analysis only, we operationally group ASC-H, HSIL, and SCC as high-risk classes and ASC-US and LSIL as low-risk classes. This grouping is not a patient-level clinical risk stratification. Let $\mathcal{H}$ and $\mathcal{L}$ denote the high- and low-risk label sets, $\mathcal{G}_\mathcal{H}$ denote all high-risk ground-truth cells, and $\mathcal{M}_\mathcal{H} = \{(g, p) \in \mathcal{M}_\theta : y_g \in \mathcal{H}\}$ denote matched high-risk cells. The high-risk recall (HR-Recall) and high-risk downgrading rate (HR-DG) are:

$$\text{HR-Recall} = \frac{1}{|\mathcal{G}_\mathcal{H}|} \sum_{g\in\mathcal{G}_\mathcal{H}} \mathbf{1}[\exists p : (g, p) \in \mathcal{M}_\theta,\ y_p \in \mathcal{H}], \tag{24}$$

$$\text{HR-DG} = \frac{1}{|\mathcal{M}_\mathcal{H}|} \sum_{(g,p)\in\mathcal{M}_\mathcal{H}} \mathbf{1}[y_p \in \mathcal{L}]. \tag{25}$$

HR-Recall measures whether high-risk cells are both localized and retained within the high-risk group, whereas HR-DG measures high-risk-to-low-risk downgrading after localization succeeds. SCR, MOE, and HR-DG are conditional on successful class-agnostic localization and should therefore be read together with HR-Recall, whose denominator includes all high-risk ground-truth cells and remains sensitive to missed detections.

4.5. Main results on TCT-MS

Table 3 compares standard detection and localization-controlled Bethesda metrics, with Rank-VFL included as a simple rank-aware baseline.

Table 2
Class distribution of cell-box annotations. High-risk total includes ASC-H, HSIL, and SCC; Tianchi contains no SCC annotations, with SCC retained as an empty category in the five-class label space.

Class	0506	0818	TCT-MS	Tianchi	CRIC
ASC-US	4584 (30.9%)	6041 (36.5%)	10,625 (33.8%)	2108 (24.4%)	1479 (32.9%)
LSIL	2146 (14.5%)	5273 (31.8%)	7419 (23.6%)	876 (10.1%)	1715 (38.2%)
ASC-H	1300 (8.8%)	3702 (22.3%)	5002 (15.9%)	1057 (12.2%)	402 (9.0%)
HSIL	5927 (39.9%)	1126 (6.8%)	7053 (22.5%)	4594 (53.2%)	663 (14.8%)
SCC	881 (5.9%)	422 (2.5%)	1303 (4.1%)	0 (0.0%)	231 (5.1%)
High-risk total	8108 (54.6%)	5250 (31.7%)	13,358 (42.5%)	5651 (65.4%)	1296 (28.9%)

Table 3
Main comparison on TCT-MS. Values are mean ± std over three random seeds. Bethesda consistency metrics are computed using class-agnostic matches at IoU≥0.5. All metrics except MOE are percentages; MOE is in ordinal-rank units. Bold denotes the best result.

Method	Backbone	mAP↑	AP@0.5↑	AP@0.75↑	LC-Acc↑	SCR↓	MOE↓	HR-Recall↑	HR-DG↓
Faster R-CNN [46]	ResNet-50	35.10 ± 0.78	48.64 ± 0.72	39.36 ± 0.83	68.72 ± 0.95	8.84 ± 0.52	0.446 ± 0.015	85.06 ± 1.38	13.32 ± 0.96
RetinaNet [47]	ResNet-50	33.86 ± 0.88	47.82 ± 0.91	37.28 ± 0.92	66.42 ± 1.18	9.40 ± 0.64	0.482 ± 0.019	83.56 ± 1.76	14.74 ± 1.12
Sparse R-CNN [48]	ResNet-50	35.32 ± 0.70	48.10 ± 0.74	39.68 ± 0.80	69.18 ± 0.90	8.78 ± 0.45	0.437 ± 0.013	84.72 ± 1.31	13.10 ± 0.88
YOLOv5-M [49]	–	35.78 ± 0.59	49.94 ± 0.60	39.42 ± 0.66	69.10 ± 0.72	8.42 ± 0.39	0.430 ± 0.011	86.02 ± 1.10	12.78 ± 0.78
YOLOv8-M [50]	–	36.28 ± 0.52	50.46 ± 0.54	40.12 ± 0.61	70.08 ± 0.65	8.16 ± 0.38	0.418 ± 0.010	86.38 ± 1.03	12.22 ± 0.72
YOLO11-M [51]	–	36.52 ± 0.50	50.74 ± 0.56	40.66 ± 0.57	69.82 ± 0.69	8.08 ± 0.36	0.421 ± 0.010	85.96 ± 0.98	12.34 ± 0.69
DETR [13]	ResNet-50	31.12 ± 1.10	43.18 ± 1.02	34.42 ± 1.20	63.46 ± 1.34	10.84 ± 0.74	0.534 ± 0.023	81.48 ± 2.12	16.08 ± 1.26
Deformable DETR [25]	ResNet-50	35.82 ± 0.64	49.20 ± 0.68	40.02 ± 0.75	69.42 ± 0.82	8.48 ± 0.43	0.434 ± 0.012	85.34 ± 1.24	12.86 ± 0.86
DINO [26]	ResNet-50	36.74 ± 0.45	50.62 ± 0.50	41.08 ± 0.55	72.86 ± 0.56	7.18 ± 0.33	0.378 ± 0.010	87.62 ± 0.94	10.82 ± 0.66
RT-DETR [14]	PResNet-50vd	37.16 ± 0.36	51.18 ± 0.40	41.58 ± 0.47	74.39 ± 0.44	7.01 ± 0.29	0.360 ± 0.008	88.62 ± 0.82	10.37 ± 0.72
RT-DETR + Rank-VFL	PResNet-50vd	37.43 ± 0.39	51.55 ± 0.43	41.82 ± 0.50	75.02 ± 0.48	6.69 ± 0.31	0.349 ± 0.009	88.95 ± 0.79	9.61 ± 0.68
CytoSet	PResNet-50vd	**39.08 ± 0.31**	**53.72 ± 0.36**	**43.52 ± 0.35**	**76.62 ± 0.42**	**5.78 ± 0.24**	**0.316 ± 0.009**	**89.86 ± 0.64**	**7.82 ± 0.46**

Table 4
Per-class AP comparison at IoU thresholds of 0.5 and 0.75.

Class	RT-DETR		CytoSet	
	AP@0.5	AP@0.75	AP@0.5	AP@0.75
ASC-US	65.6 ± 0.35	58.8 ± 0.42	**66.3 ± 0.28**	**59.1 ± 0.31**
LSIL	67.5 ± 0.43	54.6 ± 0.50	**71.3 ± 0.34**	**57.2 ± 0.39**
ASC-H	34.7 ± 0.56	25.3 ± 0.63	**36.8 ± 0.47**	**26.4 ± 0.52**
HSIL	70.9 ± 0.31	63.2 ± 0.38	**74.0 ± 0.27**	**64.5 ± 0.32**
SCC	17.3 ± 0.65	6.1 ± 0.70	**20.1 ± 0.56**	**10.3 ± 0.62**

CytoSet achieved the highest performance among the compared methods. Compared with RT-DETR, it improved mAP, AP@0.5, and AP@0.75 by 1.92, 2.54, and 1.94 points, respectively, reaching 39.08 ± 0.31 mAP, 53.72 ± 0.36 AP@0.5, and 43.52 ± 0.35 AP@0.75. These results indicate improved abnormal-cell localization and fine-grained category recognition on TCT-MS.

CytoSet also reduced grade-level errors after localization. Compared with RT-DETR, SCR decreased from 7.01 ± 0.29 to 5.78 ± 0.24, MOE from 0.360 ± 0.008 to 0.316 ± 0.009, and HR-DG from 10.37 ± 0.72 to 7.82 ± 0.46. Rank-VFL provided smaller improvements, reaching 6.69 ± 0.31 SCR, 0.349 ± 0.009 MOE, and 9.61 ± 0.68 HR-DG. Together, these results show that simple rank weighting is beneficial, while CytoSet provides larger gains in both detection performance and Bethesda consistency.

4.5.1. Category-wise analysis

Table 4 reports per-class AP@0.5 and AP@0.75. CytoSet improves both metrics across all five categories, with clear gains on LSIL, HSIL, and SCC. SCC AP@0.75 increases from 6.1 to 10.3 but remains low. Its limited representation (4.1% of annotations) may contribute to the remaining difficulty, although this effect was not isolated in our experiments.

4.5.2. Localization-controlled confusion analysis

Under the class-agnostic matching protocol, Fig. 3 shows a more concentrated diagonal for CytoSet and reduced off-diagonal mass across multiple Bethesda grades and the low-/high-risk boundary, consistent with the lower SCR, MOE, and HR-DG in Table 3. Higher LR-Recall and lower LR→HR Error also indicate improvement in the opposite direction (Supplementary Table S1).

Fig. 4 presents qualitative detection results of CytoSet and competing methods on representative TCT patches.

4.6. Ablation study

Table 5 reports component-wise ablations of DCSR, SGQS, and OCAQ on RT-DETR.

Adding DCSR increased mAP from 37.16 ± 0.36 to 38.08 ± 0.44. SGQS alone primarily improved localization at AP@0.75 and provided no standalone gain in grade-consistency metrics. OCAQ produced the largest standalone changes in Bethesda consistency, reducing SCR from 7.01 ± 0.29 to 6.42 ± 0.30, MOE from 0.360 ± 0.008 to 0.339 ± 0.010, and HR-DG from 10.37 ± 0.72 to 8.94 ± 0.64. The full model achieved the best values across all reported metrics. Adding OCAQ to DCSR + SGQS reduced SCR from 6.86% to 5.78%, MOE from 0.351 to 0.316, and HR-DG from 10.02% to 7.82%, while increasing mAP from 38.88% to 39.08%. This factorial comparison identifies OCAQ as the component associated with the main ordinal-consistency gains in the conditioned-query setting.

OCAQ couples representation learning and quality supervision through shared ordered class centers. We evaluate it as a whole and separately ablate the directional penalty $\mathcal{L}_{\text{down}}$ (Table 6); these experiments do not resolve the individual contributions of the remaining internal mechanisms.

Removing $\mathcal{L}_{\text{down}}$ left mAP nearly unchanged (39.08 to 39.01) and increased HR-DG from 7.82% to 8.24%; SCR, MOE, and HR-Recall changed only slightly. Thus, $\mathcal{L}_{\text{down}}$ provides a modest direction-specific reduction in high-to-low errors.

Feature-level DCSR responses. Fig. 5 compares DCSR response maps across feature levels. In this example, the $S3$ and $S4$ responses extend beyond the annotated target, whereas the $S5$ response is more concentrated around it. This qualitative observation motivates applying DCSR

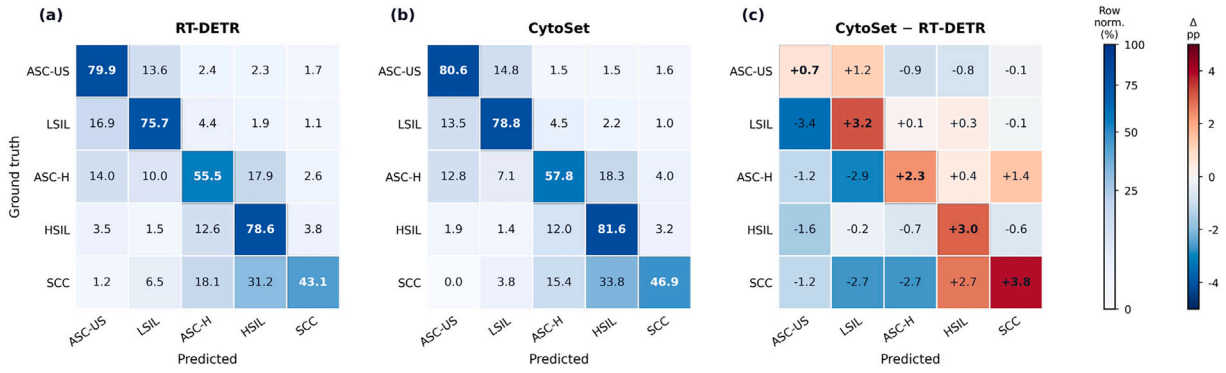


Fig. 3. Localization-controlled confusion analysis. (a) RT-DETR. (b) CytoSet. (c) CytoSet–RT-DETR difference in percentage points. Distant off-diagonal cells indicate severe cross-grade errors.

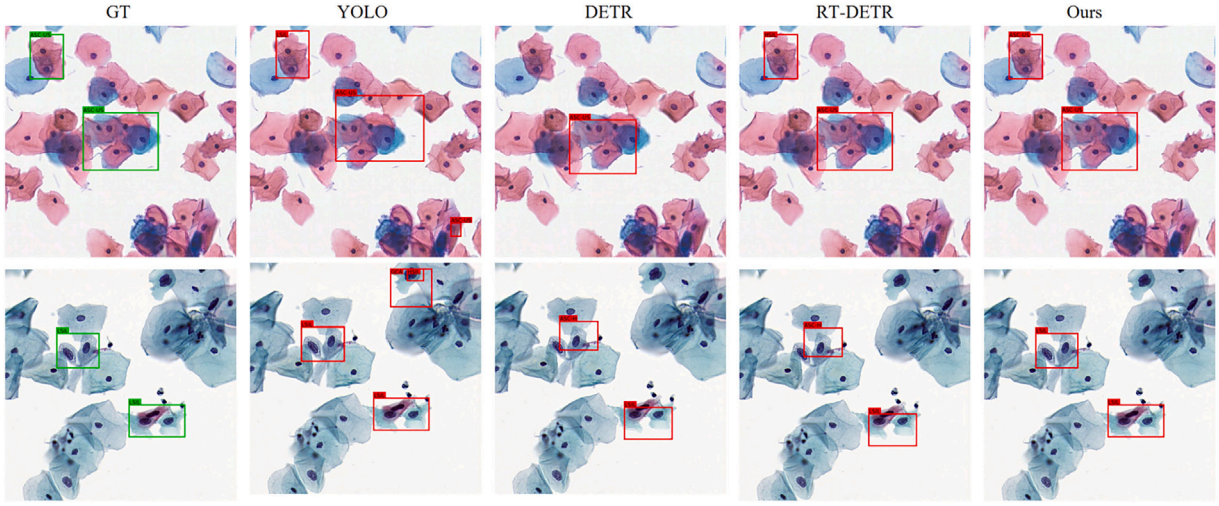


Fig. 4. Qualitative detection results on representative TCT patches. Green boxes denote ground-truth annotations and red boxes denote model predictions.

Table 5
Ablation study of DCSR, SGQS, and OCAQ. All combinations are evaluated with standard detection metrics and key localization-controlled Bethesda consistency metrics. Values are mean ± std over three random seeds. All metrics except MOE are percentages; MOE is in ordinal-rank units.

DCSR	SGQS	OCAQ	mAP↑	AP@0.5↑	AP@0.75↑	SCR↓	MOE↓	HR-DG↓
×	×	×	37.16 ± 0.36	51.18 ± 0.40	41.58 ± 0.47	7.01 ± 0.29	0.360 ± 0.008	10.37 ± 0.72
✓	×	×	38.08 ± 0.44	52.74 ± 0.48	42.48 ± 0.51	6.92 ± 0.34	0.355 ± 0.011	10.18 ± 0.78
×	✓	×	37.82 ± 0.46	52.06 ± 0.51	42.76 ± 0.52	7.06 ± 0.36	0.362 ± 0.012	10.44 ± 0.82
×	×	✓	37.68 ± 0.42	51.84 ± 0.47	42.12 ± 0.49	6.42 ± 0.30	0.339 ± 0.010	8.94 ± 0.64
✓	✓	×	38.88 ± 0.39	53.52 ± 0.43	43.23 ± 0.45	6.86 ± 0.33	0.351 ± 0.011	10.02 ± 0.75
✓	×	✓	38.50 ± 0.38	52.82 ± 0.42	42.96 ± 0.44	5.96 ± 0.26	0.324 ± 0.009	8.22 ± 0.56
×	✓	✓	38.34 ± 0.43	52.58 ± 0.46	43.08 ± 0.47	6.18 ± 0.31	0.330 ± 0.010	8.67 ± 0.62
✓	✓	✓	**39.08 ± 0.31**	**53.72 ± 0.36**	**43.52 ± 0.35**	**5.78 ± 0.24**	**0.316 ± 0.009**	**7.82 ± 0.46**

Table 6
Targeted ablation of $\mathcal{L}_{\text{down}}$, with all other CytoSet components fixed. Values are mean ± std over three seeds. Bethesda metrics use class-agnostic matches at IoU≥0.5; except for MOE (rank units), all values are percentages.

Setting	mAP↑	AP@0.5↑	AP@0.75↑	SCR↓	MOE↓	HR-Recall↑	HR-DG↓
CytoSet w/o $\mathcal{L}_{\text{down}}$	39.01 ± 0.34	53.60 ± 0.39	43.44 ± 0.39	5.91 ± 0.27	0.320 ± 0.009	89.67 ± 0.67	8.24 ± 0.51
CytoSet	**39.08 ± 0.31**	**53.72 ± 0.36**	**43.52 ± 0.35**	**5.78 ± 0.24**	**0.316 ± 0.009**	**89.86 ± 0.64**	**7.82 ± 0.46**

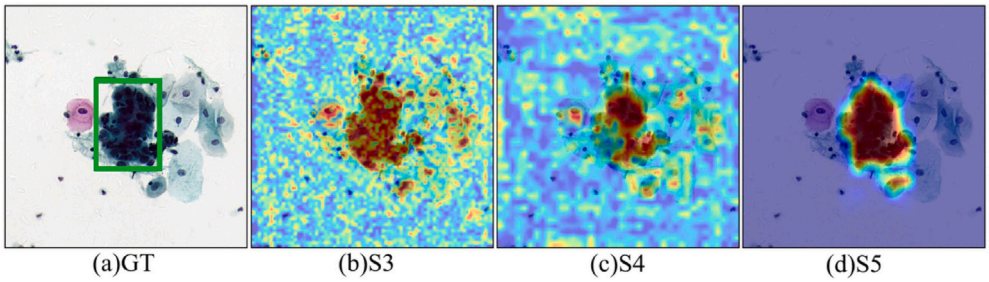


Fig. 5. Feature-level DCSR responses for a TCT patch. (a) Input patch with the annotated target outlined in green. (b)–(d) response maps from $S3$, $S4$, and $S5$, respectively; warmer colors indicate stronger responses.

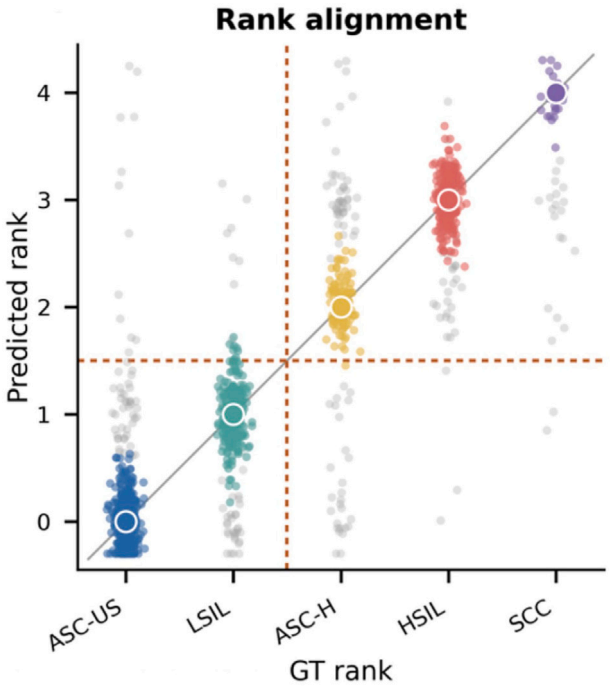


Fig. 6. Center-space rank correspondence of final-layer matched queries. Each query takes the Bethesda rank of its most cosine-similar category center. The axes show ground-truth and center-assigned ranks; colored points indicate agreement, gray points mark errors, and dashed lines mark the low-/high-risk boundary.

to projected $S5$ while leaving $S3$ and $S4$ unchanged before cross-scale fusion.

4.7. Mechanism analysis

4.7.1. Center-space rank correspondence

We examine the rank structure of matched queries directly in the OCAQ center space. For each final-layer matched query, we identify the category center with the highest cosine similarity and assign its corresponding Bethesda rank. Fig. 6 compares the ground-truth rank with this center-similarity-based rank. The correspondence is concentrated near the diagonal, with mismatches occurring predominantly between adjacent grades and relatively few long-range cross-grade deviations. Because OCAQ directly optimizes this representation, the visualization is interpreted as a mechanistic sanity check rather than an independent performance measure.

4.8. Cross-cohort and external evaluation

Table 7 separates three settings: internal cross-scanner transfer between `0506` and `0818`, in-domain replication on the external Tianchi and CRIC datasets, and external transfer from TCT-MS to Tianchi. We report standard detection metrics and localization-controlled Bethesda consistency metrics in each setting.

CytoSet outperformed RT-DETR across all evaluated settings, with higher AP and lower SCR, MOE, and HR-DG in both internal and external evaluations.

In the TCT-MS → Tianchi transfer setting, CytoSet increased mAP from 15.9 to 17.0 and reduced SCR from 27.42% to 23.78%, MOE from 1.31 to 1.18, and HR-DG from 14.08% to 12.20%.

The transfer settings were substantially harder than the in-domain settings for both models. This gap coincided with acquisition differences and case-mix shift. As shown in Table 2, the internal subsets and external datasets differed markedly in foreground category priors, with the TCT-MS → Tianchi setting reflecting a clear external domain gap. Recent work on auxiliary-domain selection and joint adaptation provides a related perspective on domain shift [53].

4.9. WSI-level Bethesda evidence retention

We conduct detector-derived evidence-retention analysis on 414 internal positive WSIs independent of detector training and 500 Tianchi positive WSIs. Sliding windows with 20% overlap are used. For each model, we select the main-experiment seed with the highest validation mAP, without using WSI labels. All detections returned by native inference and post-processing are included in highest-grade aggregation. The highest grade among detections across all windows defines the WSI prediction.

UNI [54] serves as a patch-level representation baseline. Each training patch is labeled with its highest annotated squamous Bethesda grade; the pretrained encoder is frozen and a linear five-class head is trained on these labels. The highest predicted patch grade defines the WSI prediction. In both cohorts, the reference is the highest annotated squamous grade.

Using each WSI's reference and predicted highest grades, WSI-HR-DG is the downgrading rate among high-risk reference WSIs, WSI-SCR is the proportion of evaluated WSIs for which the predicted and reference grades differ by at least two Bethesda ranks, and Risk-Group Accuracy measures agreement between the operational risk groups. These definitions follow the corresponding cell-level criteria.

Table 8 reports WSI-level high-risk downgrading, severe cross-grade error, five-class accuracy, and risk-group accuracy. Compared with RT-DETR, CytoSet reduces WSI-HR-DG from 20.3% to 14.5% internally and from 28.5% to 22.8% on Tianchi. WSI-SCR also decreases, from 9.2% to 7.0% internally and from 20.2% to 19.2% on Tianchi. Five-class accuracy increases from 76.3% to 78.5% internally and from 56.8% to 61.0%

Table 7
Cross-cohort transfer and external dataset evaluation. Matching Source and target denotes in-domain evaluation; differing datasets denote scanner or dataset transfer. All metrics except MOE are percentages; MOE is in ordinal-rank units.

Source	Target	Model	mAP↑	AP@0.5↑	AP@0.75↑	SCR↓	MOE↓	HR-DG↓
0506	0506	RT-DETR	34.0 ± 0.39	43.9 ± 0.44	38.0 ± 0.48	12.36 ± 0.31	0.59 ± 0.010	17.64 ± 0.76
		CytoSet	**35.3 ± 0.34**	**45.8 ± 0.38**	**39.2 ± 0.40**	**11.82 ± 0.26**	**0.56 ± 0.009**	**16.92 ± 0.55**
0506	0818	RT-DETR	21.2 ± 0.54	31.6 ± 0.60	23.5 ± 0.65	24.43 ± 0.44	1.15 ± 0.017	22.86 ± 0.88
		CytoSet	**22.1 ± 0.47**	**32.7 ± 0.52**	**24.4 ± 0.57**	**23.62 ± 0.38**	**1.11 ± 0.015**	**21.74 ± 0.71**
0818	0506	RT-DETR	22.0 ± 0.51	30.8 ± 0.57	24.7 ± 0.61	17.24 ± 0.41	1.09 ± 0.016	20.36 ± 0.84
		CytoSet	**23.1 ± 0.44**	**32.0 ± 0.49**	**25.8 ± 0.54**	**16.41 ± 0.35**	**1.03 ± 0.014**	**19.28 ± 0.68**
0818	0818	RT-DETR	27.8 ± 0.41	39.6 ± 0.46	29.9 ± 0.49	11.96 ± 0.30	0.58 ± 0.010	17.08 ± 0.72
		CytoSet	**28.7 ± 0.35**	**40.8 ± 0.40**	**30.9 ± 0.42**	**11.42 ± 0.25**	**0.56 ± 0.009**	**16.52 ± 0.54**
Tianchi	Tianchi	RT-DETR	28.3 ± 0.47	41.0 ± 0.53	27.0 ± 0.57	12.64 ± 0.36	0.64 ± 0.012	10.52 ± 0.79
		CytoSet	**29.9 ± 0.40**	**43.3 ± 0.46**	**28.5 ± 0.49**	**10.86 ± 0.29**	**0.56 ± 0.010**	**8.72 ± 0.60**
TCT-MS	Tianchi	RT-DETR	15.9 ± 0.60	28.5 ± 0.67	15.1 ± 0.68	27.42 ± 0.50	1.31 ± 0.020	14.08 ± 0.88
		CytoSet	**17.0 ± 0.52**	**30.3 ± 0.59**	**16.3 ± 0.61**	**23.78 ± 0.43**	**1.18 ± 0.017**	**12.20 ± 0.76**
CRIC	CRIC	RT-DETR	27.2 ± 0.50	48.8 ± 0.56	23.1 ± 0.60	11.26 ± 0.39	0.58 ± 0.013	11.10 ± 0.82
		CytoSet	**28.6 ± 0.42**	**50.7 ± 0.48**	**24.2 ± 0.51**	**9.64 ± 0.31**	**0.50 ± 0.011**	**9.35 ± 0.64**

Table 8
WSI-level Bethesda evidence retention for 414 internal and 500 Tianchi positive WSIs. Values are percentages, with counts in parentheses.

Cohort	Method	WSI-HR- DG↓	WSI-SCR↓	Accuracy↑	Risk-Group Acc.↑
Internal WSI	UNI	23.9 (33/138)	10.9 (45/414)	73.2 (303/414)	86.5 (358/414)
	RT-DETR	20.3 (28/138)	9.2 (38/414)	76.3 (316/414)	88.6 (367/414)
	CytoSet	**14.5 (20/138)**	**7.0 (29/414)**	**78.5 (325/414)**	**90.8 (376/414)**
Tianchi	UNI	29.9 (84/281)	22.4 (112/500)	60.0 (300/500)	72.6 (363/500)
	RT-DETR	28.5 (80/281)	20.2 (101/500)	56.8 (284/500)	75.2 (376/500)
	CytoSet	**22.8 (64/281)**	**19.2 (96/500)**	**61.0 (305/500)**	**76.2 (381/500)**

on Tianchi. Risk-group accuracy rises from 88.6% to 90.8% internally and from 75.2% to 76.2%, on Tianchi.

5. Conclusion

This paper formulates fine-grained TCT squamous cell detection as Bethesda-order-aware set prediction. CytoSet conditions the decoder query set through semantic refinement and spatial selection before applying ordered center alignment and quality coupling. It improves standard detection performance while reducing severe cross-grade errors, mean ordinal error, and high-risk downgrading after localization. The reporting ranks are operational surrogates for cell-level modeling rather than patient-level risks, which also depend on HPV status, screening history, and other clinical factors unavailable in this study. The results support Bethesda-order-aware query learning and indicate improved retention of detector-derived high-grade cellular evidence in the exploratory WSI analysis.

CRediT authorship contribution statement

Cheng He: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Fufeng Chen:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Fuhai Chen:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Weiwei Guo:** Writing – review & editing, Visualization, Validation, Investigation, Data curation. **Houbing Zheng:** Writing – review & editing, Validation, Investigation. **Chunxu Yang:** Writing – review & editing, Validation, Software, Investigation, Data curation. **Zhiwei Chen:** Writing – review & editing, Validation, Methodology, Formal analysis. **Qinqin Zhou:** Writing – review & editing, Validation, Resources, Data curation. **Bin Fang:** Validation, Supervision, Resources.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT (OpenAI) to assist with language refinement and manuscript-completeness checking. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Fujian Provincial Natural Science Foundation of China for Innovative Young Scientists under Grant No. 2026J008077, the Key Project of Fujian Provincial Education and Scientific Research Program for Young and Middle-aged Teachers under Grant No. JZ230006, the Engineering Research Center of Big Data Intelligence, Ministry of Education, China, and the Fujian Key Laboratory of Network Computing and Intelligent Information Processing (Fuzhou University), China.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.neucom.2026.135398.

Data availability

The internal TCT-MS dataset and the internal WSI cohort used in this study contain clinical cytology data and are not publicly available because of patient privacy, institutional data-governance, and ethical restrictions. The publicly available Tianchi cervical cytology dataset and CRIC Cervix dataset used for external evaluation can be obtained from their original repositories as cited in this article.

References

[1] R. Nayar, D.C. Wilbur, The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria, and Explanatory Notes, Springer International Publishing, 2015, https://doi.org/10.1007/978-3-319-11074-5

[2] D.F.I. Kurtycz, P.N. Staats, D.J. Chute, D. Russell, D. Pavelec, S.E. Monaco, et al., Bethesda interobserver reproducibility Study-2 (BIRST-2): Bethesda system 2014, J. Am. Soc. Cytopathol. 6 (4) (2017) 131–144, https://doi.org/10.1016/j.jasc.2017.03.003

[3] Y. Song, et al., Accurate cervical cell segmentation from overlapping clumps in PAP smear images, IEEE Trans. Med. Imag. 36 (1) (2017) 288–300, https://doi.org/10.1109/TMI.2016.2606380

[4] A. Tareef, et al., Multi-pass fast watershed for accurate segmentation of overlapping cervical cells, IEEE Trans. Med. Imag. 37 (9) (2018) 2044–2059, https://doi.org/10.1109/TMI.2018.2815013

[5] Z. Meng, Z. Zhao, B. Li, F. Su, L. Guo, A cervical histopathology dataset for computer aided diagnosis of precancerous lesions, IEEE Trans. Med. Imag. 40 (6) (2021) 1531–1541, https://doi.org/10.1109/TMI.2021.3059699

[6] J. Wang, et al., Artificial intelligence enables precision diagnosis of cervical cytology grades and cervical cancer, Nat. Commun. 15 (1) (2024) 4369, https://doi.org/10.1038/s41467-024-48705-3

[7] P. Xue, et al., Deep learning enabled liquid-based cytology model for cervical precancer and cancer detection, Nat. Commun. 16 (1) (2025) 3506, https://doi.org/10.1038/s41467-025-58883-3

[8] J. Bai, et al., AI-assisted cervical cytology precancerous screening for high-risk population in resource-limited regions using a compact microscope, Nat. Commun. 16 (1) (2025) 7429, https://doi.org/10.1038/s41467-025-62589-x

[9] Y. Zheng, et al., A graph-transformer for whole slide image classification, IEEE Trans. Med. Imag. 41 (11) (2022) 3003–3015, https://doi.org/10.1109/TMI.2022.3176598

[10] M. Yang, Z. Xie, Z. Wang, Y. Yuan, J. Zhang, Su-MICL: severity-guided multiple instance curriculum learning for histopathology image interpretable classification, IEEE Trans. Med. Imag. 41 (12) (2022) 3533–3543, https://doi.org/10.1109/TMI.2022.3188326

[11] J. Shi, L. Tang, Z. Gao, Y. Li, C. Wang, T. Gong, C. Li, H. Fu, MG-trans: multi-scale graph transformer with information bottleneck for whole slide image classification, IEEE Trans. Med. Imag. 42 (12) (2023) 3871–3883, https://doi.org/10.1109/TMI.2023.3313252

[12] C. Jin, L. Luo, H. Lin, J. Hou, H. Chen, HMIL: hierarchical multi-instance learning for fine-grained whole slide image classification, IEEE Trans. Med. Imag. 44 (4) (2025) 1796–1808, https://doi.org/10.1109/TMI.2024.3520602

[13] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: Proc. Eur. Conf. Comput. vis, 2020, pp. 213–229, https://doi.org/10.1007/978-3-030-58452-8_13

[14] Y. Zhao, et al., DETRs beat YOLOs on real-time object detection, in: Proc. IEEE/CVF Conf. Comput. vis. Pattern Recognit, 2024, pp. 16965–16974, https://doi.org/10.1109/CVPR52733.2024.01605

[15] Y. Xiang, W. Sun, C. Pan, M. Yan, Z. Yin, Y. Liang, A novel automation-assisted cervical cancer reading method based on convolutional neural network, Biocybern. Biomed. Eng. 40 (2) (2020) 611–623, https://doi.org/10.1016/j.bbe.2020.01.016

[16] L. Cao, et al., A novel attention-guided convolutional network for the detection of abnormal cervical cells in cervical cancer screening, Med. Image Anal. 73 (2021) 102197, https://doi.org/10.1016/j.media.2021.102197

[17] T. Chen, et al., A task decomposing and cell comparing method for cervical lesion cell detection, IEEE Trans. Med. Imag. 41 (9) (2022) 2432–2442, https://doi.org/10.1109/TMI.2022.3163171

[18] Y. Liang, et al., Exploring contextual relationships for cervical abnormal cell detection, IEEE J. Biomed. Health Informat. 27 (8) (2023) 4086–4097, https://doi.org/10.1109/JBHI.2023.3276919

[19] M. Fei, et al., Distillation of multi-class cervical lesion cell detection via synthesis-aided pre-training and patch-level feature alignment, Neural Netw. 178 (2024) 106405, https://doi.org/10.1016/j.neunet.2024.106405

[20] H. Jiang, Y. Zhou, Y. Lin, R.C.K. Chan, J. Liu, H. Chen, Deep learning for computational cytology: a survey, Med. Image Anal. 84 (2023) 102691, https://doi.org/10.1016/j.media.2022.102691

[21] M. Fang, B. Liao, X. Lei, F.-X. Wu, A systematic review on deep learning based methods for cervical cell image analysis, Neurocomputing 610 (2024) 128630, https://doi.org/10.1016/j.neucom.2024.128630

[22] D. Cai, et al., HiCervix: an extensive hierarchical dataset and benchmark for cervical cytology classification, IEEE Trans. Med. Imag. 43 (12) (2024) 4344–4355, https://doi.org/10.1109/TMI.2024.3419697

[23] Q. Pan, B. Yang, D. Chen, M. Wang, Enhancing model generalization of cervical fluid-based cell detection through causal feature extraction: a novel method, in: Proc. Asian Conf. Mach. Learn, 2024, pp. 1055–1070.

[24] A. Khan, et al., Impact of scanner variability on lymph node segmentation in computational pathology, J. Pathol. Inform. 13 (2022) 100127, https://doi.org/10.1016/j.jpi.2022.100127

[25] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR: deformable transformers for end-to-end object detection, in: Proc. Int. Conf. Learn. Represent, 2021.

[26] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, et al., DINO: DETR with improved denoising anchor boxes for end-to-end object detection, in: Proc. Int. Conf. Learn. Represent, 2023.

[27] Y. Gao, Y. Sun, X. Ding, C. Zhao, S. Liu, EASE-DETR: easing the competition among object queries, in: Proc. IEEE/CVF Conf. Comput. vis. Pattern Recognit, 2024, pp. 17282–17291, https://doi.org/10.1109/CVPR52733.2024.01636

[28] S. Zhang, et al., Dense distinct query for end-to-end object detection, in: Proc. IEEE/CVF Conf. Comput. vis. Pattern Recognit, 2023, pp. 7329–7338, https://doi.org/10.1109/CVPR52729.2023.00708

[29] M. Ghahremani, B.R. Ernhofer, J. Wang, M. Makowski, C. Wachinger, Organ-DETR: organ detection via transformers, IEEE Trans. Med. Imag. 44 (6) (2025) 2657–2671, https://doi.org/10.1109/TMI.2025.3543581

[30] K. Yu, W. Tan, J. Ge, X. Li, Y. Wang, J. Huang, X. Chen, S. Li, N. Zeng, M-GENE: multiview genes expression network ensemble for bone metabolism-related gene classification, Neurocomputing 622 (2025) 129318.

[31] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, Adv. Neural Inf. Process. Syst. 30 (2017).

[32] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: Proc. Eur. Conf. Comput. vis, 2016, pp. 499–515, https://doi.org/10.1007/978-3-319-46478-7_31

[33] J. Deng, J. Guo, N. Xue, S. Zafeiriou, ArcFace: additive angular margin loss for deep face recognition, in: Proc. IEEE/CVF Conf. Comput. vis. Pattern Recognit, 2019, pp. 4685–4694, https://doi.org/10.1109/CVPR.2019.00482

[34] W. Lou, et al., Structure embedded nucleus classification for histopathology images, IEEE Trans. Med. Imag. 43 (9) (2024) 3149–3160, https://doi.org/10.1109/TMI.2024.3388328

[35] W. Tang, Z. Yang, Disease-grading networks with asymmetric Gaussian distribution for medical imaging, IEEE Trans. Med. Imag. 44 (11) (2025) 4457–4472, https://doi.org/10.1109/TMI.2025.3575402

[36] Y. Peng, W. Zhu, Z. Chen, M. Wang, L. Geng, K. Yu, Y. Zhou, T. Wang, D. Xiang, F. Chen, X. Chen, Automatic staging for retinopathy of prematurity with deep feature fusion and ordinal classification strategy, IEEE Trans. Med. Imag. 40 (7) (2021) 1750–1762, https://doi.org/10.1109/TMI.2021.3065753

[37] J. Ammeling, J. Ganz, F. Wilm, K. Breininger, M. Aubreville, Investigation of class separability within object detection models in histopathology, IEEE Trans. Med. Imag. 44 (8) (2025) 3162–3174, https://doi.org/10.1109/TMI.2025.3560134

[38] Y. Zhang, H. Li, Y. Gao, H. Duan, Y. Huang, Y. Zheng, Prototype correlation matching and class-relation reasoning for few-shot medical image segmentation, IEEE Trans. Med. Imag. 43 (11) (2024) 4041–4054.

[39] W. Jiang, Y. Li, Z. Liu, L. An, G. Quellec, C. Ou, SarAdapter: prioritizing attention on semantic-aware representative tokens for enhanced medical image segmentation, IEEE Trans. Med. Imag. 45 (1) (2026) 40–52.

[40] Y. Ma, Y. Gu, S. Guo, X. Qin, S. Wen, N. Shi, W. Dai, Y. Chen, Grade-skewed domain adaptation via asymmetric bi-classifier discrepancy minimization for diabetic retinopathy grading, IEEE Trans. Med. Imag. 44 (3) (2025) 1115–1126.

[41] A. van den Oord, et al., Neural discrete representation learning, Adv. Neural Inf. Process. Syst. 30 (2017).

[42] P. Wang, H. Zhang, M. Zhu, X. Jiang, J. Qin, Y. Yuan, MGIML: cancer grading with incomplete radiology–pathology data via memory learning and gradient homogenization, IEEE Trans. Med. Imag. 43 (6) (2024) 2113–2124.

[43] H. Zhang, Y. Wang, F. Dayoub, N. Sunderhauf, VarifocalNet: an IoU-aware dense object detector, in: Proc. IEEE/CVF Conf. Comput. vis. Pattern Recognit, 2021, pp. 8510–8519, https://doi.org/10.1109/CVPR46437.2021.00841

[44] Alibaba Cloud Tianchi, Digital Human Visual Challenge: Cervical Cancer Risk Intelligent Diagnosis, Tianchi Competition (ID: 231757), 2019.

[45] M.T. Rezende, et al., CRIC searchable image database as a public platform for conventional PAP smear cytology data, Sci. Data 8 (1) (2021) 151, https://doi.org/10.1038/s41597-021-00933-8

[46] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: Adv. Neural Inf. Process. Syst, vol. 28, 2015, pp. 91–99.

[47] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proc. IEEE/CVF Int. Conf. Comput. vis, 2017, pp. 2999–3007, https://doi.org/10.1109/ICCV.2017.324

[48] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, P. Luo, Sparse R-CNN: end-to-end object detection with learnable proposals, in: Proc. IEEE/CVF Conf. Comput. vis. Pattern Recognit, 2021, pp. 14449–14458, https://doi.org/10.1109/CVPR46437.2021.01422

[49] G. Jocher, et al., Ultralytics/yolov5: v7.0, 2022. https://doi.org/10.5281/zenodo.7347926. https://github.com/ultralytics/yolov5., GitHub.

[50] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YOLOv8, 2023. https://github.com/ultralytics/ultralytics., GitHub.

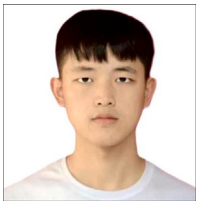
[51] Ultralytics, Ultralytics YOLO11, 2024. https://github.com/ultralytics/ultralytics., GitHub.

[52] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al., Microsoft COCO: common objects in context, in: Proc. Eur. Conf. Comput. vis, 2014, pp. 740–755, https://doi.org/10.1007/978-3-319-10602-1_48

[53] N. Zeng, Z. Li, Z. Cheng, D. Teng, P. Wu, M. Li, Auxiliary domain joint adaptation and selection for cross-domain few-shot object detection, Neurocomputing (2026) 133294.

[54] R.J. Chen, T. Ding, M.Y. Lu, D.F.K. Williamson, G. Jaume, A.H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, et al., Towards a general-purpose foundation model for computational pathology, Nat. Med. 30 (3) (2024) 850–862, https://doi.org/10.1038/s41591-024-02857-3

Author biography



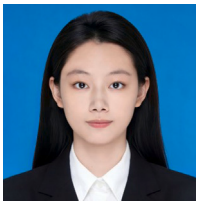
Cheng He received the B.S. degree from the School of Mathematics and Computer Science, Yuzhang Normal University, Nanchang, China, in 2024. He is currently pursuing the M.S. degree in computer science at Fuzhou University, Fuzhou, China. His research interests include medical image analysis and multimodal learning.



Fufeng Chen is the general manager of Fuzhou Zhizhen Medical Technology Co., Ltd., Fujian, China. He serves as a member of the Bioinformatics and Artificial Life Committee of the Chinese Association for Artificial Intelligence, as well as a member of the Biometrics Subcommittee of the National Information Technology Standardization Technical Committee.



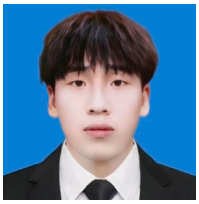
Fuhai Chen received the B.S. and Ph.D. degrees in intelligent science and technology from Xiamen University, Xiamen, China, in 2014 and 2020, respectively, through a combined M.S.–Ph.D. program, supervised by Prof. Rongrong Ji. He was formerly a postdoctoral fellow at the National University of Singapore, the University of Hong Kong, and the Hong Kong Polytechnic University. He is currently an associate professor at the College of Computer and Data Science, Fuzhou University, Fuzhou, China. His research interests include computer vision, natural language processing, large language models (LLMs), and multimodal learning.



Weiwei Guo is currently pursuing the B.S. degree in Computer Science and Technology at Fuzhou University, China. She is expected to graduate in 2027. As a core member, she has participated in multiple research projects on computational pathology and bioinformatics. Her research interests include computer vision, deep learning, and intelligent medical image analysis.



Houbing Zheng graduated from the Department of Clinical Medicine at Fujian Medical University in August 2002. He has worked in the Department of Plastic and Cosmetic Surgery at the First Affiliated Hospital of Fujian Medical University since August 2002. His research interests include 3D digital technology, artificial intelligence, wounds, scars, and his clinical work specializes in craniomaxillofacial surgery, wound repair, skin tumors, as well as aesthetic surgery including eye aesthetics, rhinoplasty, fat grafting, and facial rejuvenation.



Yang Chunxu received the B.S. degree in artificial intelligence from Hunan University of Technology and Business, China, in 2025. Yang Chunxu is currently pursuing the M.S. degree in artificial intelligence at the School of Computer and Big Data, Fuzhou University, Fuzhou, Fujian, China. His current research interests include object detection and multimodal learning.



Zhiwei Chen received the Ph.D. degree from the School of Informatics, Xiamen University, China, in 2024. He is currently a Research Fellow with the School of Artificial Intelligence, Nanchang University, China. His research interests include computer vision, weakly supervised learning, visual object localization, and multimodal learning. He has published more than 20 papers in international journals and conferences, including TPAMI, TNNLS, CVPR, and ICCV.



Qinqin Zhou received the Ph.D. degree from Xiamen University (China). He is currently working as a Lecturer with the College of Computer and Data Science, Fuzhou University, Fuzhou, China. His research falls in the fields of computer vision, machine learning, and efficient deep learning.



Bin Fang received his M.S. degree in statistics from San Diego State University, CA. He is currently working as a data analyst at Fuzhou Zhizhen Medical Technology Co., Ltd., Fujian, China.