DOI: 10.19884/j.1672-5220.202502001

RoCoNet: Rotational Contrastive Network for Semi-Supervised Cervical Cell Image Object Detection

HUANG Qiubo*, GONG Runze, CHEN Dehua
School of Computer Science and Technology, Donghua University, Shanghai 201620, China

Abstract: A semi-supervised learning framework integrating rotational invariance, contrastive learning, and adaptive hybrid thresholds, named rotational contrastive network (RoCoNet), is proposed to enhance the applicability of semi-supervised learning for medical cell datasets. Due to the unique sampling approach of cell datasets, input images often contain uncertain rotation angles, which render traditional convolution kernels ineffective in existing semi-supervised detectors. To address this challenge, rotational attention convolution is introduced, offering robustness to rotational transformations. Additionally, cross-feature contrastive loss is proposed to improve upon the contrastive loss used in supervised learning, tackling issues of poor classification performance caused by cell overlap and clustering. An adaptive hybrid threshold is also introduced to stabilize pseudo-label generation during early training. A global threshold, computed by using Gaussian mixture models (GMMs), is applied to refine the local threshold, which helps balance the quantity and quality of pseudo-labels. Experiments on the ThinPrep cytology test (TCT) dataset for cervical cytopathology show that RoCoNet achieves a mean average precision (mAP) of 31.6% with only 10% labeled data, outperforming the baseline method by 8.4% in mAP.

Keywords: semi-supervised learning; rotational invariance; contrastive learning; object detection; cervical cell
CLC number: TP183 **Document code**: A
Article ID: 1672-5220(2026)02-0082-12

Open Science Identity (OSID)

0 Introduction

With the rapid advancement of deep learning technologies, computer vision has achieved remarkable progress in the field of medical image analysis. In cancer detection, the analysis of cell images holds significant importance. However, despite the availability of large volumes of image data, the acquisition of labeled data heavily relies on experts, making the annotation process time-consuming and costly. To address the scarcity of labeled data, semi-supervised learning has garnered significant attention due to its capability to synergistically leverage limited labeled data and vast amounts of unlabeled data.

In cell image analysis, cervical cancer cell detection faces several challenges. Firstly, images are typically captured by using devices such as microscopes, which do not guarantee that cells remain in fixed orientations. Cells of the same class may appear at varying angles in different images, and the rotational, translational, and scale variations in cell images make it difficult for models to obtain stable feature representations. Secondly, under conditions of limited space, rapid proliferation, or increased interaction, cells are prone to aggregation and overlapping, leading to significant intra-class variability and increasing the difficulty of classification. Finally, in semi-supervised object detection, the confidence threshold used for pseudo-label selection significantly impacts model performance[1]. As model training progresses, the performance improves, leading to an increase in confidence scores. Overly high thresholds in the early training stages can result in insufficient pseudo-labels, while overly low thresholds in later stages can introduce errors. Therefore, overcoming these challenges and improving the accuracy and robustness of cell object detection within a semi-supervised learning framework are the primary focuses of this study.

This study proposes a semi-supervised learning framework, rotational contrastive network (RoCoNet), based on rotational convolution, contrastive learning, and adaptive hybrid thresholds. The main contributions are as follows.

1) A rotational convolution is proposed. Theoretically, continuous rotational convolutions achieve rotational invariance. However, in practice, to balance the performance and speed, a discrete solution is adopted, employing finite rotation angles (e.g., 0°, 90°, 180°, and 270°). Innovatively, this study integrates discrete rotation layers by using the efficient channel attention (ECA) module, enabling the model to exhibit a degree of rotational invariance.

2) The contrastive loss in unsupervised learning is optimized to make it suitable for semi-supervised learning, improving model classification performance, and effectively addressing intra-class dissimilarity caused

Received date: 2025-02-02
* Correspondence should be addressed to HUANG Qiubo, email: huangturbo@dhu.edu.cn
Citation: HUANG Q B, GONG R Z, CHEN D H. RoCoNet: rotational contrastive network for semi-supervised cervical cell image object detection [J]. *Journal of Donghua University (English Edition)*, 2026, 43(2): 82-93.

by cell overlap and aggregation. To mitigate the issue of insufficient positive and negative sample pairs due to limited labeled data, cross-layer calculation is employed, incorporating five feature layers of different sizes.

3) A novel pseudo-label filtering method, adaptive hybrid thresholds, is proposed. The expectation-maximization (EM) algorithm is used to fit a Gaussian mixture model (GMM) to determine a global threshold. Additionally, intra-class exponential moving averages (EMAs) are used to compute local thresholds, which refine the global threshold. The combination of global and local thresholds balances the quantity and quality of pseudo-labels.

4) The effectiveness of the proposed method is validated on a cervical cancer cell dataset. Compared to the baseline method, the proposed method achieves a mean average precision (mAP) improvement of 8.4% (note that all percentage increases of mAP mentioned in this paper are absolute improvements) when trained with only 10% labeled data.

1 Related Work

1.1 Semi-supervised learning

As a technique situated between supervised and unsupervised learning, semi-supervised learning has been widely applied across various fields in recent years. Combining a small amount of labeled data with a large amount of unlabeled data, it improves the generalization ability of models in scenarios with scarce labeled data. Due to the difficulty and high cost of obtaining labeled cell datasets, semi-supervised learning offers an effective solution. Lee[2] proposed a semi-supervised object detection method that enhances the utilization of unlabeled data through pseudo-labels and self-training strategies. While this method achieves promising results on several datasets, issues with pseudo-label quality remain, particularly when the initial model has low accuracy, as errors in pseudo-labels can propagate through subsequent training and degrade final detection performance. Tarvainen et al.[3] introduced the mean teacher model, a teacher-student architecture for semi-supervised learning. In this method, the teacher and student models share the same network architecture but use different parameter update mechanisms. The teacher model's parameters are the EMAs of the student model's parameters. While the method performs well on many datasets, its application to aerial or cell datasets is hindered by challenges such as cell rotation and overlap, resulting in suboptimal pseudo-label quality.

1.2 Rotational invariance

Traditional convolutional neural networks (CNNs) often rely on data augmentation techniques, such as rotating input images, to handle rotational transformations. While these methods improve model robustness to some extent, they do not achieve true rotational invariance. Cohen et al.[4] proposed the group equivariant convolution (GEC) network, which introduces group theory into convolution operations to ensure rotational equivariance while reducing the limitations of traditional CNNs in handling rotations. However, the complexity of its structure and reliance on group structures lead to high computational overhead during training and inference. Moreover, its performance declines in scenarios involving multiple transformations, such as translation and scaling. Huang et al.[5] proposed distilled rotated kernel fusion (DRKF), which reduces computational costs through rotational convolution combined with knowledge distillation. However, the design of its rotational convolution aggregates all rotation layers with the same weights, making the model less robust to inputs at specific rotation angles. Wei et al.[6] introduced RC4 convolution, which employs predefined weight parameters to achieve better performance in fisheye lens tasks. However, this method cannot generalize to other datasets, such as the cervical cancer cell dataset used in this study.

1.3 Confidence adaptive thresholds

Most object detection algorithms (e.g., Faster R-CNN[7] and YOLOv3[8]) typically use fixed confidence thresholds during inference[9-10] to filter detection results. However, these algorithms do not account for variations in model prediction confidence across classes and iterations, which can significantly impact performance[11]. Wang et al.[12], in consistent-teacher, used the EM algorithm to fit a GMM for each class and set the threshold at the peak of the probability density. However, this method performs poorly on imbalanced samples, and using the EM algorithm for each class requires substantial computational resources. Wang et al.[13] proposed FreeMatch, which approximates global and local thresholds by using EMAs and combines them to determine the final threshold, achieving near-optimal performance with relatively high efficiency. Building on these researches, this study proposes further improvements to calculate sufficiently accurate confidence thresholds while maintaining high performance.

2 Method

In this section, we detail the overall architecture of the proposed model RoCoNet and explain how RoCoNet addresses three critical challenges in semi-supervised learning for cervical cells: rotational invariance, intra-class diversity, and pseudo-label bottlenecks.

2.1 Network architecture

The overall framework of RoCoNet is illustrated in Fig. 1. It is based on consistent-teacher[12], a state-of-the-art semi-supervised object detection method built on the RetinaNet detector. The teacher model takes weakly augmented unlabeled images as input to generate pseudo-labels, which are used to supervise the student model's training. The teacher model's parameters are updated solely through the EMA of the student model's

parameters. The student model applies strong augmentation to the same unlabeled images and weak augmentation to labeled images, learning from both to produce classification, regression, and contrastive loss terms. Inference is conducted exclusively by using the teacher model.

Given a labeled sample set $D_{\mathrm{L}} = \{x_i^{\mathrm{l}}, y_i^{\mathrm{l}}\}_{i=1}^{N}$ with N samples and an unlabeled sample set $D_{\mathrm{U}} = \{x_j^{\mathrm{u}}\}_{j=1}^{M}$ with M samples, where x^{l} represents labeled input images, x^{u} represents unlabeled input images, and y represents labels, the teacher detector f_{t} generates pseudo-labels $\hat{y}_j^{\mathrm{u}}$ as

$$\hat{y}_j^{\mathrm{u}} = f_{\mathrm{t}}(T(x^{\mathrm{u}})), \tag{1}$$

where T represents the weak augmentation (weak Aug).

The student detector f_{s} minimizes the following loss function L[12]:

$$L = \frac{1}{N}\sum_{i=1}^{N} L_{\mathrm{cls}}(f_{\mathrm{s}}(T(x_i^{\mathrm{l}})), y_i^{\mathrm{l}}) + \lambda \frac{1}{M}\sum_{j=1}^{M} L_{\mathrm{cls}}(f_{\mathrm{s}}(T'(x_j^{\mathrm{u}})), \hat{y}_j^{\mathrm{u}}) + \frac{1}{N}\sum_{i=1}^{N} L_{\mathrm{reg}}(f_{\mathrm{s}}(T(x_i^{\mathrm{l}})), y_i^{\mathrm{l}}) + \lambda \frac{1}{M}\sum_{j=1}^{M} L_{\mathrm{reg}}(f_{\mathrm{s}}(T'(x_j^{\mathrm{u}})), \hat{y}_j^{\mathrm{u}}) + L_{\mathrm{con}}^{(\mathrm{L})} + L_{\mathrm{con}}^{(\mathrm{U})}, \tag{2}$$

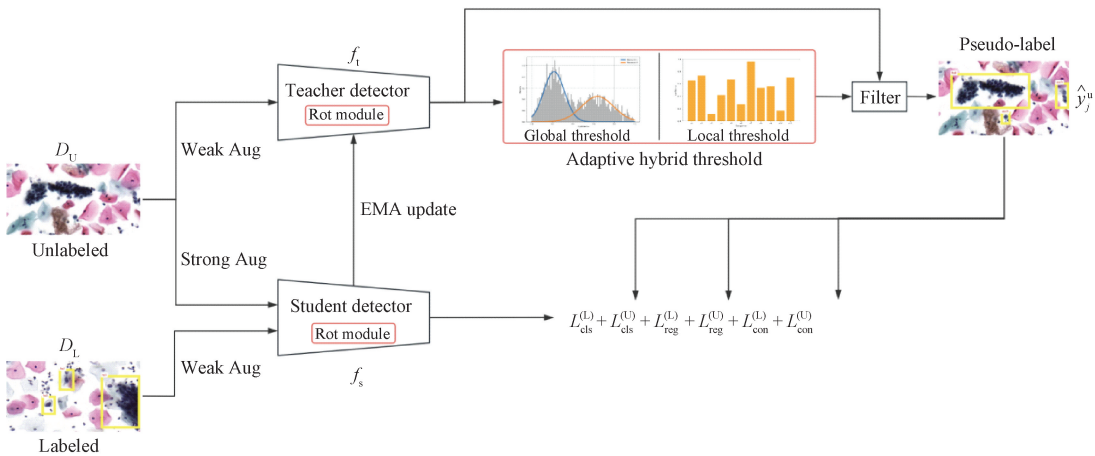
where T' represents the strong augmentation (strong Aug); L_{cls} denotes the classification loss computed using the focal loss[14]; L_{reg} represents the regression loss calculated with the generalized intersection over union (IoU) loss; superscript L indicates computations based on pseudo-labels, and superscript U signifies calculations using ground-truth labels; $L_{\mathrm{con}}^{(\mathrm{L})}$ and $L_{\mathrm{con}}^{(\mathrm{U})}$ denote contrastive loss terms, which will be detailed in Subsection 3.3; λ is a hyperparameter controlling the impact of pseudo-labels on the loss function.

The three key improvements proposed in this model are as follows.

1) Rotational attention module: enhances rotational invariance for cell detection tasks.

2) Cross-feature contrastive loss: effectively addresses intra-class dissimilarity caused by cell overlap and aggregation.

3) Adaptive hybrid threshold: balances the quantity and quality of pseudo-labels throughout the training process.



Rot—rotational attention.

Fig. 1 Overall framework of RoCoNet

2.2 Rotational attention module

During the collection of cell datasets, the camera is fixed directly above the sample, resulting in cells appearing at arbitrary angles within the images. Even cells of the same type may exhibit different orientations across images. However, conventional convolution operations possess only translational invariance and lack rotational invariance (i.e., the results of convolving a rotated input differ from those obtained by firstly convolving and then rotating the result). Consequently, the model may struggle to capture the features of targets when they appear at varying rotation angles in the input images.

To address this issue, we introduce a rotational invariant convolution W_{RIC} and integrate it between the backbone and feature pyramid network (FPN) layers (see the rotational attention module in Fig. 2). The module operates on the backbone's C_3, C_4, and C_5 feature layers, generating P_3, P_4, and P_5. Here, C_3, C_4, and C_5 denote the output feature maps from different stages of the backbone. The resulting P_3, P_4, and P_5 are the feature pyramid levels produced by the rotational attention module for multi-scale object detection. P_5 and P_4 are then upsampled and added to P_4 and P_3, respectively. In addition, P_5 is further downsampled to generate P_6 and P_7. The rotational attention module does not alter the original logic of the FPN layers, making it a plug-and-play module.

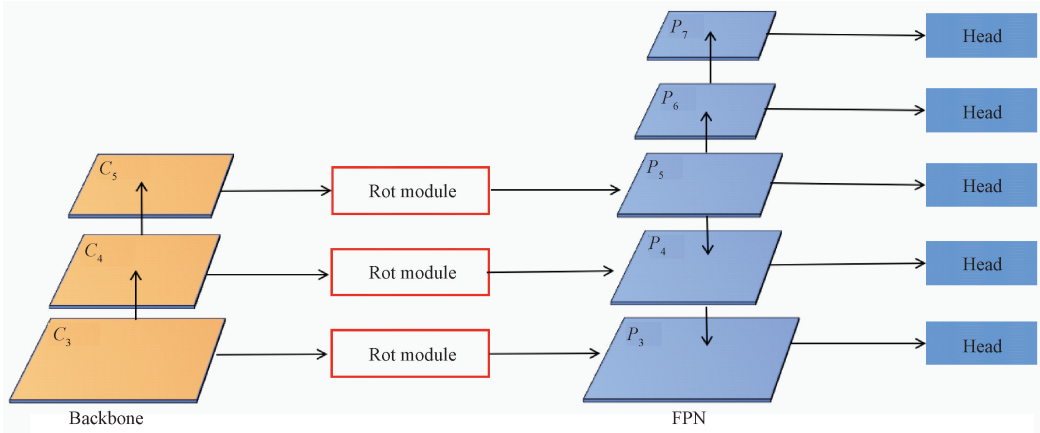


Fig. 2 Location of rotational attention module

The mechanism of W_{RIC} is as follows.

$$M_{\mathrm{out}} = M_{\mathrm{input}} \times W_{\mathrm{RIC}} = \sum_{n=0}^{N=1} (M_{\mathrm{input}} \times [\boldsymbol{L}^{\theta_n} W]),\ \theta_n = \frac{2\pi}{D} n, \quad (3)$$

where $\boldsymbol{L}^{\theta_n}$ represents the rotational transformation matrix (with a rotation angle of θ_n); W denotes the original convolution; M_{input} is the input feature; M_{out} is the output feature. The rotational invariant convolution W_{RIC} ensures that rotating the input feature by θ_n and then applying the rotational convolution is equivalent to applying the rotational convolution firstly and then rotating the output by θ_n [5].

$$[\boldsymbol{L}^{\theta_n} M_{\mathrm{input}}] \times W_{\mathrm{RIC}} = \boldsymbol{L}^{\theta_n} [M_{\mathrm{input}} \times W_{\mathrm{RIC}}], \quad (4)$$

where W_{RIC} is a continuous two-dimensional (2D) operation, enabling rotational invariance for images at arbitrary angles. However, due to computational constraints, it is not feasible to incorporate excessive rotation layers in W_{RIC}. When the number of rotations D is set to 4, the model achieves the invariance described by Eq. (4) for input features rotated at 0°, 90°, 180°, and 270°. We expect that when the input image is rotated by 90°, the 90° rotation layer will dominate. Similarly, for an input rotated by 45°, both the 0° and 90° rotation layers will contribute. By combining rotational convolution with the ECA attention module, this work achieves adaptive weighting α_n for the rotation layers.

$$M_{\mathrm{out}} = M_{\mathrm{input}} \times W_{\mathrm{RIC}} = \sum_{n=0}^{N=1} \alpha_n (M_{\mathrm{input}} \times [\boldsymbol{L}^{\theta_n} W]),$$
$$\theta_n = \frac{2\pi}{D} n. \quad (5)$$

As shown in Fig. 3, we duplicate the input feature into four copies, each of which is convolved with a different rotational state of the kernel (i. e., $\boldsymbol{L}^{\theta_n} W$ in Eq. (3)). The outputs are then rearranged along the channel dimension to obtain the rotational attention feature, ensuring that the ECA module spans across each rotational convolution.

The ECA module is an efficient channel attention mechanism, and its core idea is to capture the inter-channel correlations through local cross-channel interactions, thereby obtaining the weight for each channel. The ECA module firstly applies global average pooling (GAP) to the input feature map, producing a new feature map $\boldsymbol{Y}$ with dimensions $1 \times 1 \times C$ (where C is the number of channels in the rotational attention feature). Then, the feature map $\boldsymbol{Y}$ is passed through a one-dimensional (1D) convolutional layer with a kernel size of $1 \times k$. The purpose of this operation is to capture the local dependencies between different channels along the channel dimension. The value of k is adapted based on the number of input channels.

$$k = \left| \frac{\log_2 C}{2} + \frac{1}{2} \right|_{\mathrm{odd}}, \quad (6)$$

where $\left| \frac{\log_2 C}{2} + \frac{1}{2} \right|_{\mathrm{odd}}$ represents the nearest odd number to $\left| \frac{\log_2 C}{2} + \frac{1}{2} \right|$. The output feature map $\boldsymbol{Z}$ is passed through a sigmoid activation function, mapping its values to a range between 0 and 1, resulting in a weight coefficient vector $\boldsymbol{A}$. Each element in $\boldsymbol{A}$ represents the weight coefficient α_n for the corresponding channel. Finally, the original input feature map rotational attention feature is multiplied by the corresponding elements in $\boldsymbol{A}$, achieving adaptive weighting between channels.

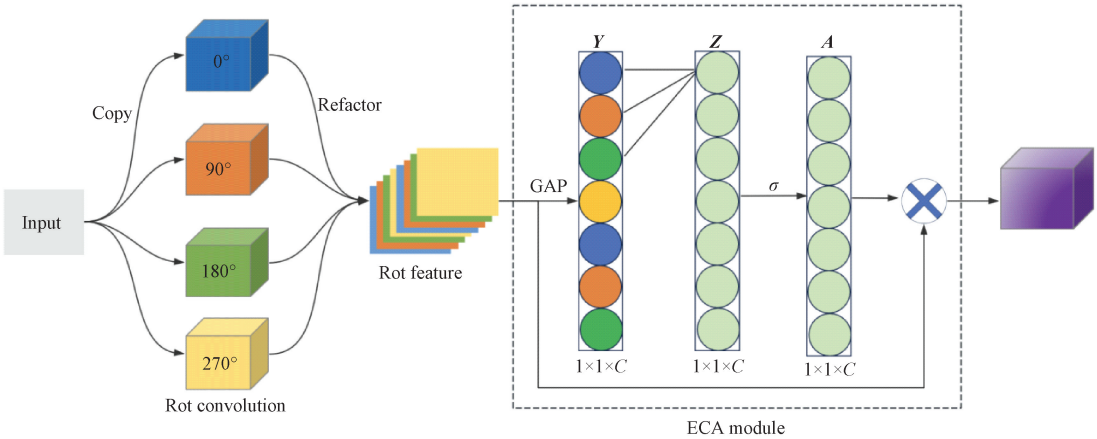


Fig. 3 Cross-layer rotational attention convolution

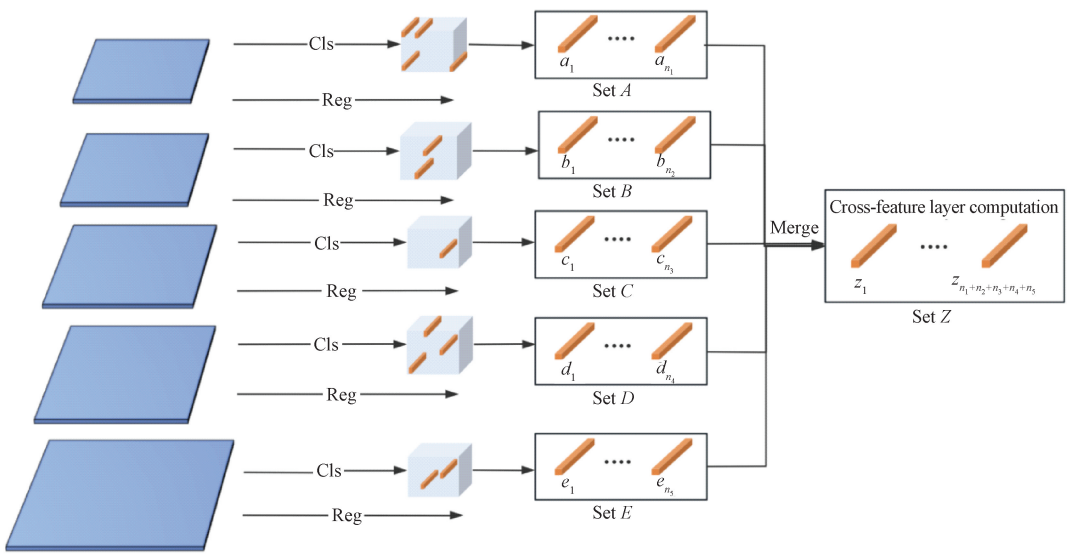
2.3 Cross-feature contrastive loss

In the cervical cancer cell dataset, due to the frequent occurrence of cell aggregation and overlap, there is a problem where similar cells within the same class have low similarity, while cells from different classes exhibit high similarity. This makes classification more challenging for the model. Therefore, this study adopts the concept of contrastive learning, using instance contrastive loss for comparing abnormal cells, thereby increasing the distance between different classes while reducing the distance within the same class. The final contrastive loss is added to the total loss function with a certain weight.

Within the same batch, if two predictions belong to the same class, they form a positive sample pair; if they belong to different classes, they form a negative sample pair. We choose the predicted bounding box feature vector from the classification branch of the head (marked as the orange vector in Fig. 4, denoted as a, b, c, d, and e for the five head layers, respectively), with a dimension of 256 × 1. For each head layer, the number of extracted feature vectors is denoted as n_1, n_2, n_3, n_4, and n_5, respectively. We incorporate all the predicted bounding box feature vectors from the five head layers into the same set Z for computation. The cross-layer contrastive loss is computed, which significantly increases the number of positive and negative sample pairs[15], thereby improving the accuracy of contrastive loss calculation. The contrastive loss[16] is

$$L_{\text{con}}^{(\text{L})} = \sum_{i \in I} \frac{-1}{|J(i)|} \sum_{j \in J(i)} \log \frac{\exp(e_i \cdot e_j / \tau)}{\sum_{a \in A(i)} \exp(e_i \cdot e_a / \tau)}, \quad (7)$$

where $L_{\text{con}}^{(\text{L})}$ is the contrastive loss for labeled images; I is the set of all predicted boxes; E is the set of all predicted feature vectors; e_i is the feature vector of the predicted box i; $J(i)$ is the set of predicted boxes with the same class as the predicted box i; $A(i)$ is the set of predicted boxes with a different class from the predicted box i; τ is the temperature parameter used to adjust the smoothness of the softmax function in the loss. The contrastive loss $L_{\text{con}}^{(\text{U})}$ for pseudo-labeled images is computed in the same manner.



Cls—classification branch; Reg—regression branch.

Fig. 4 Cross-feature contrastive loss

2.4 Adaptive hybrid threshold

In semi-supervised object detection, the confidence threshold used to filter pseudo-labels has a significant impact on model performance[1]. For each predicted bounding box of a pseudo-label (x, y, w, h, and q), the model outputs a confidence q. If the confidence is lower than the threshold, it is discarded. Generally, as the model trains, the confidence gradually increases. In the early stages of training, confidence is usually low, and if the threshold is set too high, the number of generated pseudo-labels will be insufficient. In contrast, in the later stages of training, the confidence is generally higher, and if the threshold is set too low, it will introduce too much error. To address the issue of fixed thresholds, this study adopts an adaptive hybrid threshold, allowing the model to automatically adjust the threshold during training.

We fit the confidence of all outputs by using the EM algorithm to model the GMM and find the confidence corresponding to the maximum probability density as the global threshold. Assuming that the confidence distribution follows a GMM with both positive and negative modes:

$$p(u) = \pi_n N(u \mid \mu_n, (\sigma_n)^2) + \pi_p N(u \mid \mu_p, (\sigma_p)^2), \quad (8)$$

where $N(x \mid \mu, \sigma^2)$ represents the Gaussian distribution; π_n, μ_n, and $(\sigma_n)^2$ represent the weights, means, and variances for the negative modes; π_p, μ_p, and $(\sigma_p)^2$ represent the weights, means, and variances for the positive modes. The global confidence threshold τ_g is the confidence at the maximum probability density for the positive mode:

$$\tau_g = \arg\max_c P(\text{pos} \mid c, \mu_p, (\sigma_p)^2). \quad (9)$$

In practice, we maintain a queue of size 1 000 to store the confidences for fitting the GMM. Considering the specificities of each class, the global threshold may not be entirely reliable. Therefore, we also compute a local threshold $\tau_1^c(t)$ for each class, as shown in Eq. (10).

$$\tau_1^c(t) = \begin{cases} \dfrac{1}{F}, & \text{if } t = 0, \\ \lambda \tau_1^c(t-1) + (1-\lambda)\dfrac{1}{n_c}\sum_{i=1}^{n_c} q_i, & \text{otherwise}, \end{cases} \quad (10)$$

where $\lambda \in (0, 1)$ is the momentum decay for the EMA; n_c is the number of predicted boxes for class c in the current batch; q_i is the confidence of the predicted box from the model's regression branch (x, y, w, h, q); F is the number of classes; t represents the batch number, and when $t = 0$, $\tau_1^c(0)$ is initialized to $1/F$.

Finally, we apply maximum normalization to the local thresholds and combine them with the global threshold to obtain the adaptive hybrid threshold $\tau_c(t)$ for the tth batch, which will be used to remove predicted boxes with confidence lower than the threshold:

$$\tau_c(t) = \text{MaxNorm}(\tau_1^c(t)) \cdot \tau_g, \quad (11)$$

where MaxNorm is the maximum normalization.

3 Experimental Results

3.1 Dataset

Our experiments were primarily conducted on the ThinPrep cytology test (TCT) dataset for cervical cytopathology[17]. This dataset contains 7 410 cervical microscopic images, which were cropped from whole slide images (WSI) obtained from the Pannoramic MIDI II digital slide scanner. The dataset includes a total of 48 587 instances, categorized into 11 classes, as shown in Fig. 5. These classes include: normal squamous cell (NSC), high-grade precancerous squamous lesion (HSIL), atypical squamous cell with mild abnormalities (ASC-US), atypical squamous cell with suspicious abnormalities (ASC-H), low-grade precancerous squamous lesion (LSIL), squamous cell carcinoma (SCC), adenocarcinoma (ADC), normal endocervical cell (NEC), inflammatory cell (IC), blood cell or debris as background interference (BC), and cell clump (CC). CC represents overlapping or aggregated cell and is a key challenge in real clinical slide detection, especially under limited data conditions.

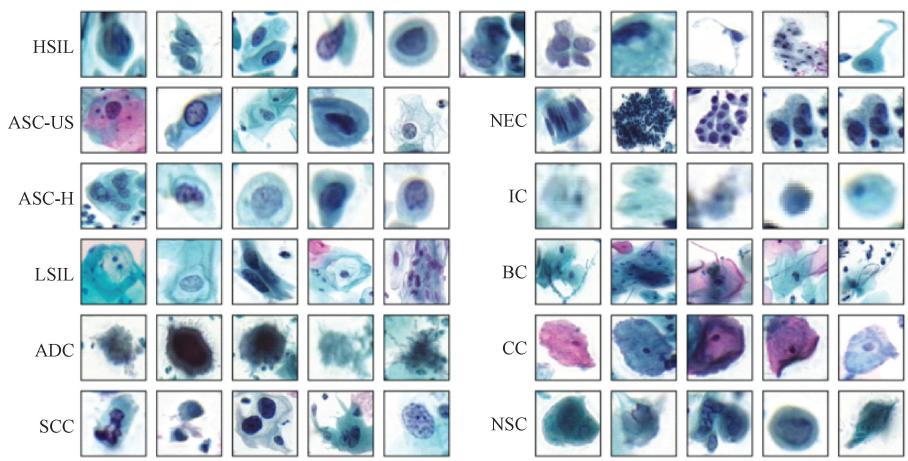


Fig. 5 TCT dataset visualization

The dataset was randomly split into a training set containing 6 666 images and a test set containing 744 images. After performing a series of preprocessing steps, such as removing labels with extremely small bounding box areas and dividing large images into four equal parts, a total of 6 926 images with 45 774 instances were used for model training, and the test set consisted of 772 images with 5 039 instances.

To evaluate the model's performance in semi-supervised learning, we used only 10% of the labeled data from the training set, and the remaining 90% of labels were removed. The unlabeled data were assigned pseudo-labels by the teacher model, and the final experimental results were evaluated on the test set.

3.2 Parameter settings

For the TCT dataset, we trained the model on an NVIDIA GeForce RTX 4090 device with a batch size of eight. During training, a total of 84 000 iterations were performed, with the first 1 000 iterations using a warm-up strategy to stabilize the training process. The learning rate was initially set to 0.005 and was gradually reduced to 0.000 5 and 0.000 05 at the 54 000th and 69 000th iterations, respectively. The above methods were built based on the MMdetection framework. The weight for unlabeled data λ_u was set to 2.

3.3 Rotational attention module

The proposed rotational attention module consists of a rotational convolution and an attention module. In Section 2, we designed a cross-layer ECA to integrate the rotational convolution. Additionally, we also tried an intra-layer ECA method (i.e., directly stacking the output of the rotational convolution along the channel dimension and then applying ECA). However, since the experimental results were inferior to the cross-layer ECA, we ultimately chose the cross-layer ECA. The comparison of experimental results for the three methods is shown in Table 1. The evaluation metrics include mAP, which represents the overall detection performance averaged over multiple IoU thresholds; AP_{50} and AP_{75}, which measure detection accuracy at IoU thresholds of 0.50 and 0.75, respectively, reflecting loose and strict localization requirements; AP_S, AP_M, and AP_L, which evaluate the detection performance on small, medium, and large objects, respectively, indicating the model's effectiveness across different object scales.

Table 1 Performance comparison of rotational convolution with different numbers of rotations and ECA methods

Method	Number of rotations	Accuracy/%					
		mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Rotational convolution	2	26.3	51.0	23.8	7.2	16.8	26.3
	4	26.4	52.2	23.5	10.0	16.9	25.4
	8	27.3	50.8	26.4	10.5	15.2	26.6
Rotational convolution+ intra-layer ECA	2	27.0	53.0	24.9	6.2	14.2	27.5
	4	27.2	52.4	24.4	8.0	15.0	26.3
	8	27.6	53.0	26.3	6.8	16.5	27.2
Rotational convolution+ cross-layer ECA	2	28.0	54.3	25.5	9.2	15.5	27.6
	4	28.5	54.2	27.6	7.2	16.2	27.9
	8	30.2	56.5	30.0	7.5	16.9	29.6
	16	29.7	54.5	28.8	9.49	16.2	29.5

To explore the optimal implementation of rotational convolution, we conducted experiments with a number of rotations of 2, 4, and 8. The final experimental results are shown in Table 1. In the three experimental groups, increasing the number of rotations generally leads to improved accuracy. The highest accuracy is achieved when the number of rotations is 8, as the model can achieve rotational invariance across 8 angles. However, when the number of rotations increases to 16, the model overfits due to the introduction of too many additional parameters, leading to a decrease in accuracy. Table 2 reports the baseline detection performance, where all experiments are conducted under the condition of using rotational convolution with 8 rotations. It is observed that simply adding rotational convolution improves the mAP by 4.1% compared to the baseline consistent-teacher in Table 2. When the ECA module is directly added, the mAP further improves by 0.3%. After implementing cross-layer ECA with convolution reordering, the accuracy achieves a significant improvement, reaching an mAP of 30.2%.

Table 2 Baseline detection performance of different methods under 8-rotational convolution setting

Method	Accuracy/%					
	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet[14]	18.6	36.0	17.4	6.4	9.4	18.1
Consistent-teacher[12]	23.2	45.0	23.3	7.5	13.0	24.3

3.4 Equivariance error

The ideal rotational equivariant convolution should satisfy the condition that the convolution of the rotated input is equivalent to the rotation of the convolved input, i.e., $(\boldsymbol{L}M) \times W = \boldsymbol{L}(M \times W)$. Therefore, we define the equivariance error $\Delta(\theta)$ as[6]

$$\Delta(\theta) = \frac{1}{n}\sum_{i}^{n}\frac{\mathrm{std}((\boldsymbol{L}^{\theta}M_i) \times W - \boldsymbol{L}^{\theta}(M_i \times W))}{\mathrm{std}((\boldsymbol{L}^{\theta}M_i) \times W)}, \quad (12)$$

where $\boldsymbol{L}^{\theta}$ is the rotational transformation matrix for a rotation by angle θ; M_i is the input feature; W is the convolution; std denotes the standard deviation. We use the features from the FPN layer as input and rotate them through angles from 0° to 360° to calculate $\Delta(\theta)$. Ideally, it should be 0. As shown in Fig. 6, at 0°, there is no rotation, so $\Delta(\theta)$ is 0. At 180°, $\Delta(\theta)$ for standard convolution (Conv) and GEC[4] reaches its peak, while RoCoNet maintains a lower $\Delta(\theta)$. Overall, $\Delta(\theta)$ for RoCoNet is significantly lower than that of Conv and GEC, demonstrating excellent rotational robustness.

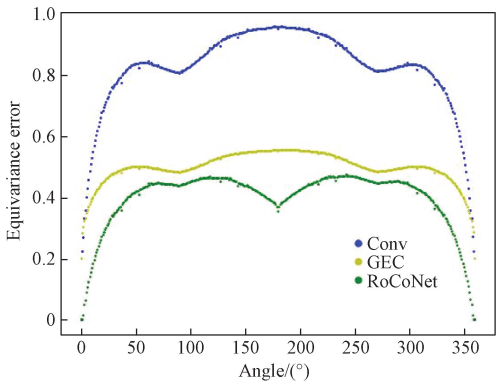


Fig. 6 Rotational convolution equivariance error

3.5 Cross-feature contrastive loss

To demonstrate the effectiveness of the cross-feature contrastive loss, we conduct experiments based on the rotational attention module (rotational convolution + cross-layer ECA) to compare the impact of cross-feature layers.

This study designs two methods for calculating the cross-feature contrastive loss.

The first method takes into account that each feature layer has a different receptive field, leading to different feature distributions. Therefore, the cross-feature contrastive loss is calculated within each feature layer, and the results of the five feature layers are then summarized.

The second method directly calculates a single contrastive loss across multiple feature layers. Since the cross-feature contrastive loss result is highly dependent on the number of positive and negative sample pairs, this method can significantly increase the number of positive and negative pairs, thus improving the calculation accuracy of the cross-feature contrastive loss.

The comparison of contrastive loss calculation methods across different feature layers, as shown in Table 3, indicates that when using rotational convolution, the mAP only improves by 0.1% (from 30.2% to 30.3%) without the cross-layer calculation. However, when the cross-layer calculation is used, the cross-feature contrastive loss significantly improves the accuracy. Compared to the mAP (30.2%) achieved with rotational convolution + cross-layer ECA with a number of rotations of 8 as shown in Table 1, the mAP improves by 0.8%.

Table 3 Comparison of contrastive loss calculation methods across different feature layers

Method	Contrastive loss	Accuracy/%					
		mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Rotational attention module	Intra-layer calculation	30.3	56.0	31.0	6.6	15.7	31.2
Rotational attention module	Cross-layer calculation	31.0	57.9	29.0	11.7	16.8	31.2

3.6 Adaptive hybrid threshold

To demonstrate the effectiveness of the adaptive hybrid threshold, this study compares it with other threshold methods, including fixed threshold, local threshold, and global threshold. The fixed threshold is set to 0.3, the local threshold uses the EM algorithm to fit GMM within each class[12], and the global threshold uses an exponential moving average[13]. The results are shown in Table 4.

The experimental results in Table 4 show that the adaptive hybrid threshold achieves higher accuracy than other threshold methods, with an mAP of 31.6%.

Table 4 Performance comparison of different threshold methods

Method	Accuracy/%					
	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Fixed threshold	30.3	56.1	30.0	10.6	16.3	30.7
Local threshold	31.0	57.9	29.0	11.7	16.8	31.2
Global threshold	30.9	56.7	30.6	7.8	16.8	31.3
Adaptive hybrid threshold	31.6	57.4	30.6	12.3	19.5	31.3

In Fig. 7, we provide some qualitative analyses to illustrate the mechanism of the adaptive hybrid threshold. Figure 7(a) shows the trend of threshold changes, where the local and adaptive hybrid thresholds take the class-average threshold. Figure 7(b) shows the number of ground truth boxes change trend. All three adaptive thresholds gradually increase during training, while the fixed threshold remains unchanged. In the early stage of training, the adaptive hybrid threshold has a lower threshold compared to the fixed threshold, ensuring a sufficient number of pseudo-labels. In the later stage, the adaptive hybrid threshold has a higher threshold compared to other methods, keeping the label quantity at a lower level and preventing excessive error from being introduced into the model. Throughout the entire training process, the adaptive hybrid threshold maintains the most stable label quantity and ultimately achieves higher accuracy.

The global threshold ignores differences between classes. In the early stages, a low threshold leads to an excessive number of labels, and the resulting low-quality labels negatively affect the model's training. The fixed threshold, being set too high in the early stages, yields too few pseudo-labels, severely hindering the model's training. The local threshold has an appropriate number of labels in the early stages, but the threshold becomes too low in the later stages, causing the number of labels to become excessive. At this point, the model focuses on the quality of labels rather than their quantity, and low-quality labels interfere with training. Additionally, since the local threshold requires fitting GMM within each class, its training time increases by 10% compared to other methods. Therefore, among the three threshold methods, the adaptive hybrid threshold method is the most reasonable.

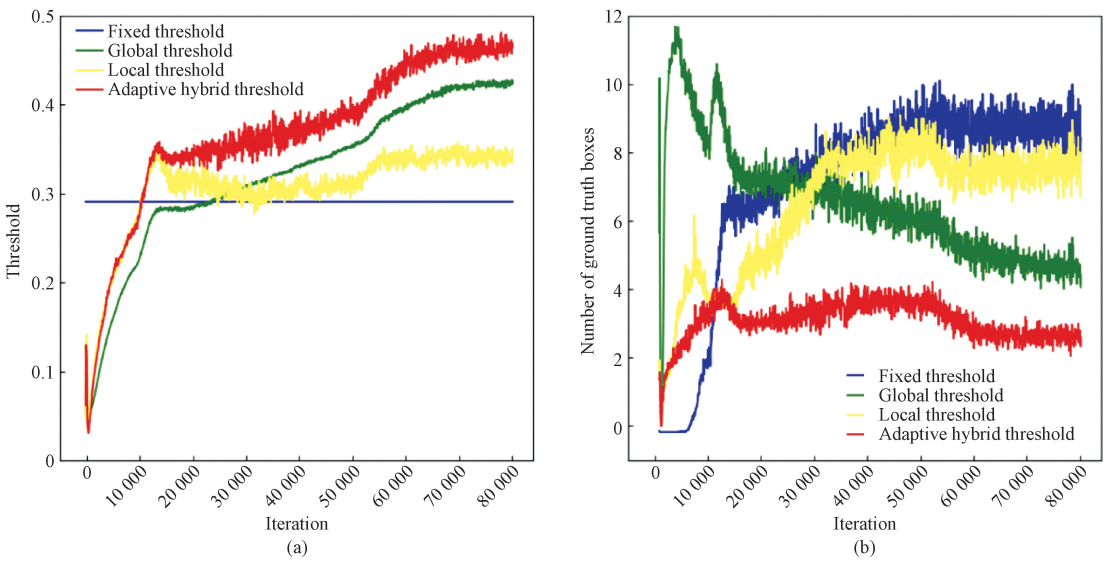


Fig. 7 Comparison of adaptive hybrid threshold with other threshold methods: (a) threshold change trend; (b) number of ground truth boxes change trend

3.7 Ablation study

To validate the effectiveness of each module, we conduct a detailed ablation study. In Table 5, the impact of each module on the model's performance is progressively demonstrated. The checkmark √ in Table 5 indicates that the module is applied. When contrastive loss and adaptive hybrid threshold are used, the model achieves an mAP of 30.0%. With the addition of rotational attention convolution and adaptive hybrid threshold, the mAP increases to 30.7%. When rotational attention convolution, contrastive loss, and adaptive hybrid threshold are all used together, the model achieves

an mAP of 31.6%.

Table 5 Ablation study on effect of different modules on detection performance

Rotational attention convolution	Contrastive loss	Adaptive hybrid threshold	mAP/%
√			30.2
	√		28.9
		√	27.3
	√	√	30.0
√	√		31.0
√		√	30.7
√	√	√	31.6

3.8 Model performance validation

To further validate the performance of our model, we firstly performed a horizontal comparison of model performance on the TCT dataset. The comparison experiment results are shown in Table 6. TCT-full denotes that the model is trained by using 100% of the available labeled data in the TCT dataset, while TCT-10% indicates that only 10% of the labeled data are randomly selected and used for training, with the remaining data treated as unlabeled. This setting follows a semi-supervised learning protocol and is designed to evaluate the model's performance under limited annotation conditions.

Table 6 Comparison experiment results on TCT dataset

Model	Dataset	mAP/%
ComparisonDetector[17]	TCT-full	26.3
Co-mining[18]	TCT-full	30.0
CO-FCOS[19]	TCT-full	33.8
RoCoNet (ours)	TCT-10%	31.6

As shown in Table 6, RoCoNet achieves an mAP of 31.6% by using only 10% of the labeled data, surpassing the performance of ComparisonDetector and Co-mining, which use 100% of the labeled data, and the mAP is only 2.2% lower than that of CO-FCOS. It is worth mentioning that under the same training parameters, CO-FCOS achieves only an mAP of 8.3% with 10% of the labeled data.

To verify the generalization ability of RoCoNet, we also evaluate it on the blood cell count and detection (BCCD) dataset. The BCCD dataset contains 364 images with a resolution of 640 × 480 pixels and includes three types of cells. It is a smaller-scale dataset and widely used as an open-source benchmark. The BCCD dataset shares similar challenges with the TCT dataset, including rotational invariance issues, making it suitable for this experiment. The experimental results are shown in Table 7. RoCoNet outperforms the previous best-performing model, PLB, by 0.3% in mAP, demonstrating its superior performance on small datasets.

Table 7 Comparison experiment results on BCCD dataset

Model	mAP/%
CST-YOLO[20]	62.5
TXL-PBC[21]	62.9
PLB[22]	63.7
RoCoNet (ours)	64.0

4 Conclusions

This study presents a systematic study on semi-supervised cervical cell object detection, proposing a semi-supervised learning framework, RoCoNet, based on rotational invariance, contrastive learning, and adaptive hybrid threshold. The framework achieves robustness to cell rotational transformations through rotational attention convolution, which enables discrete rotational convolutions to approximate rotational invariance. Additionally, the cross-feature contrastive loss enhances the model's ability to recognize similar cells, addressing the challenge posed by cell overlap and aggregation that can degrade model classification performance. The adaptive hybrid threshold ensures the quality and quantity of pseudo-label generation. Experimental results on the TCT dataset for cervical cytopathology demonstrate that RoCoNet can achieve an mAP of 31.6% using only 10% of the labeled data. On the BCCD dataset, it achieves an mAP of 64.0%, outperforming the state-of-the-art PLB model by 0.3%.

References

[1] CHEN B H, LI P Y, CHEN X, et al. Dense learning based semi-supervised object detection [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2022: 4805-4814.

[2] LEE D H. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks [C]//Workshop on Challenges in Representation Learning. New York: ICML, 2013: 896.

[3] TARVAINEN A, VALPOLA H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results[EB/OL]. (2017-03-06)[2025-01-02]. https://arxiv.org/abs/1703.01780v6.

[4] COHEN T S, WELLING M. Group equivariant

convolutional networks [C]//Proceedings of the 33rd International Conference on Machine Learning. New York: ACM, 2016: 2990-2999.
[5] HUANG R R, CAI J C, LI C, et al. DRKF: distilled rotated kernel fusion for efficient rotation invariant descriptors in local feature matching [C]//2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). New York: IEEE, 2023: 1885-1892.
[6] WEI X, SU S X, WEI Y, et al. Rotational convolution: rethinking convolution for downside fisheye images [J]. *IEEE Transactions on Image Processing*, 2023, 32: 4355-4364.
[7] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6): 1137-1149.
[8] FA RHADI A, REDMON J. YOLOv3: an incremental improvement[C]//Computer Vision and Pattern Recognition. Berlin: Springer, 2018, 1804: 1-6.
[9] LIU Y C, MA C Y, HE Z J, et al. Unbiased teacher for semi-supervised object detection[EB/OL]. (2021-02-18) [2025-01-02]. https://arxiv.org/abs/2102.09480v1.
[10] LIU Y C, MA C Y, KIRA Z. Unbiased teacher v2: semi-supervised object detection for anchor-free and anchor-based detectors [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2022: 9809-9818.
[11] ZHANG B W, WANG Y D, HOU W X, et al. FlexMatch: boosting semi-supervised learning with curriculum pseudo labeling [EB/OL]. (2021-10-15)[2025-01-02]. https://arxiv.org/abs/2110.08263v3.
[12] WANG X J, YANG X Y, ZHANG S L, et al. Consistent-teacher: towards reducing inconsistent pseudo-targets in semi-supervised object detection [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2023: 3240-3249.
[13] WANG Y D, CHEN H, HENG Q, et al. FreeMatch: self-adaptive thresholding for semi-supervised learning [EB/OL]. (2022-05-15) [2025-01-02]. https://arxiv.org/abs/2205.07246v3.
[14] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection [C]//2017 IEEE International Conference on Computer Vision (ICCV). New York: IEEE, 2017: 2999-3007.
[15] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations [C]//International Conference on Machine Learning. Cambridge: PMLR, 2020: 1597-1607.
[16] KHOSLA P, TETERWAK P, WANG C, et al. Supervised contrastive learning [EB/OL]. (2020-04-23)[2025-01-02]. https://arxiv.org/abs/2004.11362v5.
[17] LIANG Y X, TANG Z H, YAN M, et al. Comparison-based convolutional neural networks for cervical cell/clumps detection in the limited data scenario[EB/OL]. (2018-10-14)[2025-01-02]. https://arxiv.org/abs/1810.05952v5.
[18] WANG T C, YANG T, CAO J L, et al. Co-mining: self-supervised learning for sparsely annotated object detection[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 35(4): 2800-2808.
[19] HUANG Q B, CHEN X, GONG R Z, et al. Small objects detection for incomplete annotated cervical cancer cells dataset [C]//2023 IEEE 6th International Conference on Pattern Recognition and Artificial Intelligence (PRAI). New York: IEEE, 2023: 68-73.
[20] KANG M, TING C M, TING F F, et al. CST-yolo: a novel method for blood cell detection based on improved YOLOv7 and CNN-Swin transformer [C]//2024 IEEE International Conference on Image Processing (ICIP). New York: IEEE, 2024: 3024-3029.
[21] GAN L, LI X. TXL-PBC: a freely accessible labeled peripheral blood cell dataset[EB/OL]. (2024-07-18)[2025-01-02]. https://arxiv.org/abs/2407.13214v1.
[22] HU B, LIU Y, CHU P Z, et al. Small object detection via pixel level balancing with applications to blood cell detection[J]. *Frontiers in Physiology*, 2022, 13: 911297.

RoCoNet：一种用于宫颈细胞图像半监督目标检测的旋转对比网络

黄秋波*，龚润泽，陈德华
东华大学 计算机科学与技术学院，上海 201620

摘　要：本研究提出了一种基于旋转不变性、对比学习和自适应混合阈值的半监督学习框架——RoCoNet，旨在提升半监督学习在医学细胞数据集上的适用性。细胞数据集独特的采样方式使得输入图片的旋转角度不定，现有的半监督检测器所使用的传统卷积核难以有效工作。为解决该问题，本研究提出了一种旋转注意力卷积机制，以增强模型对旋转变换的稳健性。同时，通过对有监督学习中的对比损失进行改进，提出了跨特征对比损失，并将其运用在半监督学习中，解决了细胞重叠和聚集导致模型分类性能较差的问题。此外，对于训练早期伪标签数量不稳定的情况，我们提出了自适应混合阈值，在局部阈值的基础上使用高斯混合模型（Gaussian mixture model，GMM）计算全局阈值对其进行修正，使伪标签的生成兼顾了数量和质量。在宫颈细胞学液基薄层细胞检测（ThinPrep cytology test，TCT）数据集上的实验显示，仅使用 10% 的标签数据，RoCoNet 的平均精度（mean average precision，mAP）就能达到 31.6%，较基线方法的 mAP 高出 8.4 个百分点。

关键词：半监督学习；旋转不变性；对比学习；目标检测；宫颈细胞