Research on Blood Cell Detection Algorithm Based on Improved YOLOv7

Yaxuan Wang †, Cun Zhao †, Zaichao Zhu, Jian Jia *

School of Mathematics, Northwest University, Xi'an, China, 710127

* Corresponding Author Email: jiajian@nwu.edu.cn

† These authors contributed equally.

Abstract. In the biomedical field, the detection of blood cells in microscopic images is crucial for assisting physicians in diagnosing blood-related diseases and plays a significant role in promoting the development of medicine towards more precise and efficient treatments. Traditional manual detection methods are time-consuming and prone to errors, and existing blood cell detection technologies face significant challenges in meeting the requirements of high precision and real-time performance. In light of this, this paper, from the perspective of image recognition and with the aid of deep learning, proposes an efficient and rapid detection model based on YOLOv7. Firstly, in order to further extract global features, this paper selects the advanced Vision Transformer module to be added to the backbone network. Then, the convolutional layer of the SPPCSPC layer in the YOLOv7 backbone network is replaced with a parameterized convolution, thus avoiding the shortcoming of traditional static convolutions where all samples share one convolution kernel. To directly add the global feature information learned in the small-scale feature layer to the maximum-scale feature layer, this paper adds an upsampling module between the minimum-scale feature layer and the maximum-scale feature layer. By using the SIoU loss function instead of the CIoU loss function, the convergence speed is further accelerated, and the precision is improved. Secondly, the experimental results validate the effectiveness of the improved YOLOv7 model. Compared to the results of YOLOv7, the mAP@0.5 value of this paper's model has increased by 3.4% compared to the original YOLOv7, by 0.7% compared to the YOLOv5x with a larger number of parameters, and by 0.3% compared to the currently best-performing open-source model in this dataset-CST-YOLO.

Keywords: Blood cell detection, YOLOv7, Vision Transformer, CondConv, SIoU.

1. Introduction

1.1. Background

The research object of blood cell detection mainly includes WBC , RBC and platelets. Typically, microscopic identification is used for the detection of blood cells; however, it is time-consuming, and the fatigue and experience of the testers can affect the accuracy of the detection. Although alternative optical or electrical devices exist, they are expensive and require specialized training and knowledge. Early methods of blood cell counting and detection were primarily based on traditional image processing techniques. Patil, Sable, and Anandgaonkar [1] used gray-level threshold processing to count and detect blood cells. Currently, mainstream blood cell target detection methods can be roughly divided into two categories: two-stage detection and single-stage detection. In two-stage detection, Raina et al. [2] applied Faster R-CNN to the Blood Cell Count and Detection dataset to detect RBCs, and Najmeddine Dhieb et al. [3] proposed an automated blood cell counting framework using convolutional neural network (CNN), instance segmentation, transfer learning, and mask R-CNN techniques. Although these methods operate slowly, they offer high precision. Single-stage detection, on the other hand, provides an improvement in detection speed compared to two-stage detection, with accuracy gradually increasing. Xia et al. [4] reduced the number of convolutional layers and filter depth based on YOLOv3. Yihai Mao et al. [5] proposed the DWS-YOLO blood detector. Chenyang Shi et al. [6] proposed a lightweight blood cell detection model based on YOLOv8n, utilizing advanced lightweight strategies and the PGhostC2f design. Z Liu et al. [7] introduced the improved YOLO-BC algorithm to solve the pixel-level differences of different

categories of blood cells by combining efficient multi-scale attention and full-dimensional dynamic convolution models, thereby achieving fast and accurate identification and counting of blood cells. M Huang et al. [8] proposed blood cell target detection algorithm based on YOLOv5 addresses the issue of low average accuracy and serious miss detection due to small blood cells and serious cell adhesion in blood cell detection by target detection algorithms. However, addressing the complex features of blood cell images, resolving issues of cell misdetection and false detection, and meeting the high demand for real-time performance in medical environments remain significant challenges that current research faces.

1.2. Main work

This paper selects the YOLOv7 algorithm as the foundational network. Within the YOLOv7 backbone network, the Vision Transformer module is added, and the SPPCSPC layer is replaced with parameterized convolutions. Subsequently, the neck network is enhanced with four times upsampling, and the SIoU loss function is utilized in place of the CIoU loss function. Finally, through comparative experiments on a public blood cell detection dataset, the detection performance of the model proposed in this paper is significantly improved, demonstrating its suitability for blood cell detection tasks.

2. Model Introduction

2.1. YOLOv7 Algorithm Principles

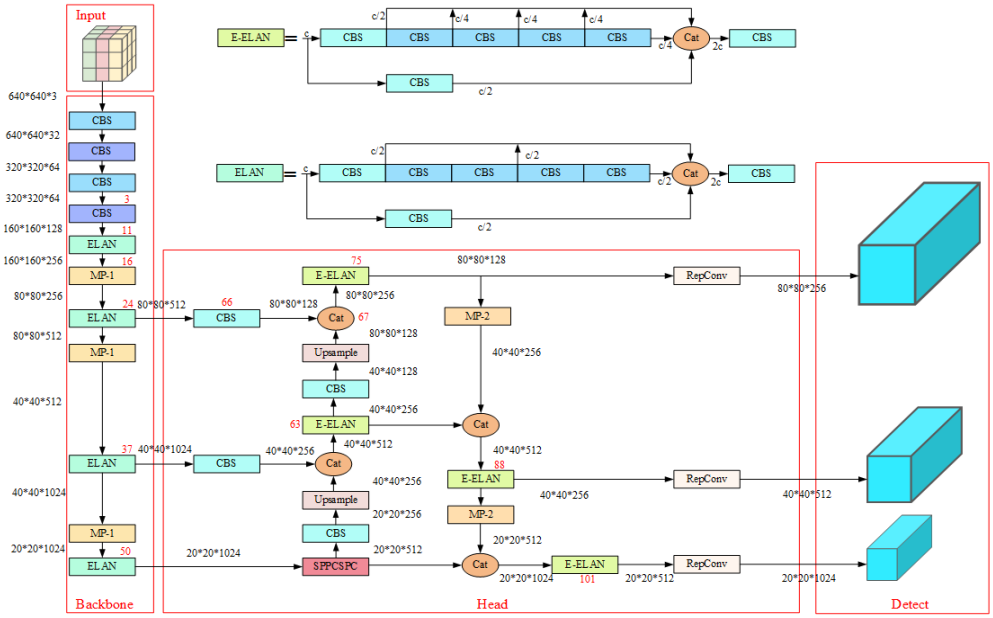

Figure 1. YOLOv7 network structure diagram

The YOLOv7 [9] network structure is primarily composed of three parts: the input end, the Backbone, and the Head. The input end is responsible for preprocessing the input images; the Backbone layer focuses on extracting feature information, with the E-ELAN module enhancing the network's learning capability and promoting the learning of more diverse features by different computational modules; the Head layer consists of the SPPCSPC layer, multiple MPCONV layers, multiple Catconv layers, and the Rep layer. The SPPCSPC layer captures multi-scale feature information through pooling operations at different scales to accommodate targets of varying sizes and accelerates the propagation of features from the main stream to the branch streams through cross-stage partial connections, thereby enhancing feature representation capabilities, and finally outputs detection results through the Rep layer. Overall, YOLOv7 optimizes the network through model structure reparameterization and dynamic label assignment. In terms of reparameterizing the model, YOLOv7 analyzes the gradient flow propagation path and reparameterization strategies for different

modules in the network, as well as utilizes reparameterized convolutions to process these modules, thus maintaining network prediction performance while reducing network complexity. YOLOv7's dynamic label assignment strategy combines the positive and negative sample allocation strategies of YOLOv5 and YOLOX, screening out more positive samples to improve recall. Finally, multiple Detect headers are used to decouple feature information, confirming the location and category of targets, which enhances the network model's feature extraction capability and the precision of object detection. The network structure is illustrated in Fig.1.

2.2. Principles of the Improved YOLOv7 Algorithm

The overall architecture of the improved YOLOv7 model proposed in this paper is shown in Fig.2 and Fig.3 with the improvements indicated by the red dashed line boxes. The model comprises three main components: the backbone network (backbone) for feature extraction from images, the neck network (neck) for fusing multi-scale image features, and the detection head (detect) for outputting the final detection results.

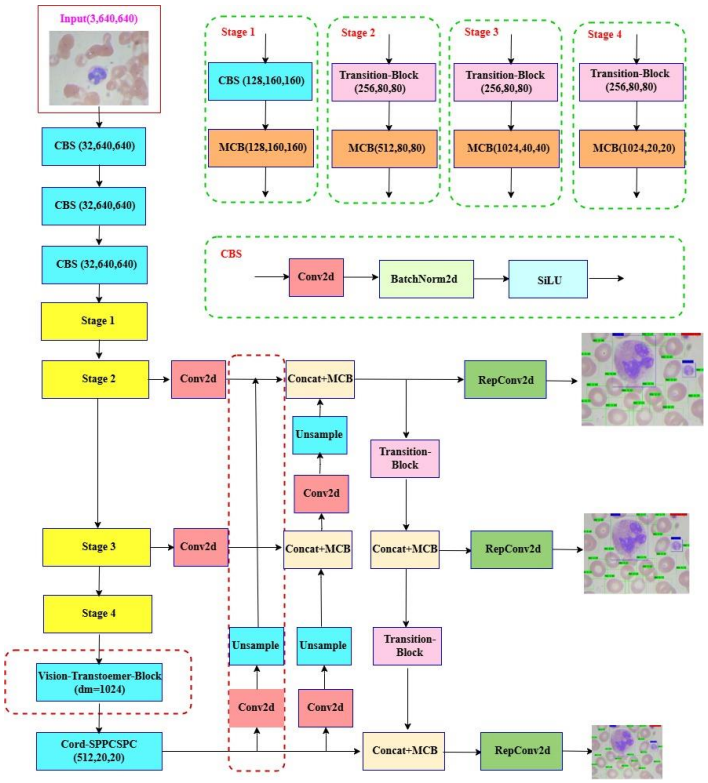

Figure 2. Improved YOLOv7 network structure diagram (a)

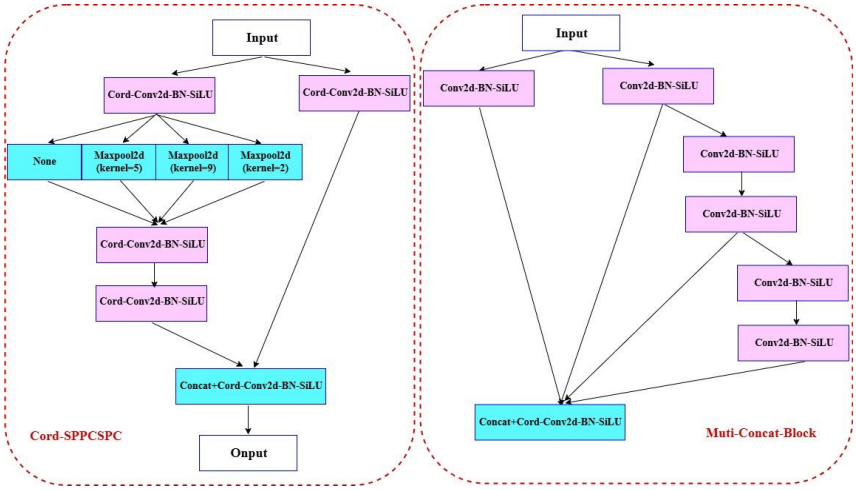

Figure 3. Improved YOLOv7 network structure diagram (b)

The main improvements of the model proposed in this paper are focused on the backbone network's output end, the neck network, and the loss function. The following sections will introduce the four improvement points of this paper, respectively.

2.2.1. Vision Transformer

The original YOLOv7 model's backbone network primarily consists of feature extraction layers composed of convolutions, normalizations, and SiLU activation functions, as well as the Multi_Concat_Block layer, which is designed to enhance the model's receptive field and feature representation capabilities, thereby capturing target information across various scales more effectively. Additionally, the Transition_Block layer is utilized for downsampling.

Convolutional Neural Network (CNN) architectures extract local features from images through locally connected convolutional kernels and significantly reduce the number of model parameters through parameter sharing, making the network more efficient and easier to train. At the same time, CNNs exhibit translation invariance: CNNs are invariant to translations in the input image, meaning that regardless of the target's position within the image, CNNs can effectively extract its features. However, CNNs lack the ability to capture global features. Therefore, this paper incorporates the Vision Transformer [10] module at the end of the backbone network to extract global features. Specifically, an 8-layer Vision Transformer module with a dimensionality of 1024 is added at the end of the backbone network, which can fully utilize the original pretrained weights to accelerate the model's convergence speed.

2.2.2. CondConv

In YOLOv7, the SPPCSPC module plays a crucial role in the network architecture, primarily being utilized for feature extraction, especially in enhancing the feature extraction network. The key characteristic of this layer is the incorporation of multiple MaxPool operations in parallel within a series of convolutions. Through the design of the SPPCSPC module, YOLOv7 can effectively mitigate issues such as image distortion caused by image processing operations. Additionally, this layer addresses the problem of redundant features that may arise during the extraction of image features by the model. This aids the model in better understanding and recognizing objects within images, thereby improving the accuracy of object detection.

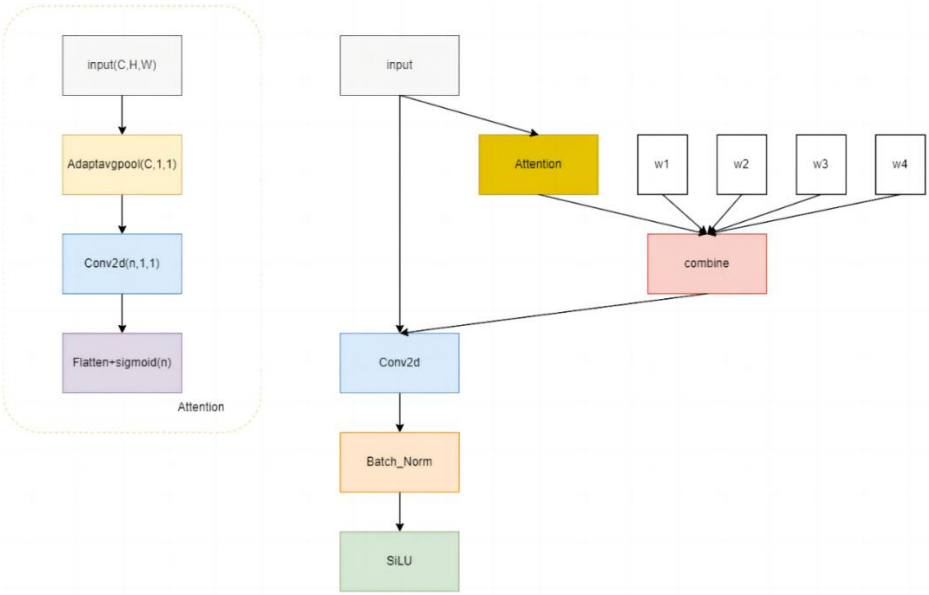

Figure 4. CondConv network structure diagram

As previously mentioned, the weights in CNN architectures are shared, which significantly reduces the number of parameters but can have an impact on model accuracy. Therefore, in this work, the article replaces the convolutional layers of this module with CondConv (parametric convolution). CondConv dynamically computes convolution kernels based on the input, thus overcoming the

limitation of traditional static convolutions where a single kernel is shared across all samples. Specifically, the convolution kernels in the CondConv layer are parameterized as a linear combination of n convolution kernels: $\alpha_1\omega_1 + \alpha_2\omega_2 + ... + \alpha_n\omega_n$, where w_i represents the convolution kernel, and α_i are the parameters learned through backpropagation. CondConv [11] was introduced by Google's team in their work published at NeurIPS 2019, and its main principle is illustrated in Fig.4.

In the specific parameter settings, this paper sets the number of convolution kernels, n, to 4. With such a design, the model's learning capability can be further enhanced.

2.2.3. Upsample

To enhance the model's ability to learn global features, Improvement Point 1 has added 8 Vision Transformer (ViT) modules. However, this module only affects the smallest medium-scale feature layer among the three output feature layers and does not have a direct impact on the largest-scale feature layer. Consequently, this paper introduces an upsampling module between the smallest and largest scale feature layers. With this module, the global feature information learned from the small-scale feature layer can be directly incorporated into the largest-scale feature layer, allowing all three output feature layers to learn the global characteristics of the image. Furthermore, since the paper employs dimensionality reduction followed by bilinear interpolation upsampling, the computational cost to the model is virtually unchanged.

2.2.4. Loss function

The loss function of the YOLO series primarily consists of three components: confidence loss, classification loss, and bounding box regression loss. The confidence loss is used to determine whether an object is present in the predicted box; the classification loss is used to assess the probability of the object in the predicted box belonging to different categories; and the bounding box regression loss measures the positional deviation between the predicted box and the ground truth box, ensuring that the model learns the target's positional information during training. In recent years, several outstanding bounding box regression loss functions have been proposed, such as GIoU [12], DIoU [13], and CIoU [13]. YOLOv7 employs the CIoU loss function, which takes into account the overlap area, normalized center, and aspect ratio between the predicted and ground truth boxes. However, CIoU does not consider the misalignment of orientation between the two boxes, which can cause oscillation of the detection box during training, leading to slow convergence and reduced accuracy of the model. Therefore, this paper uses the SIoU loss function [14] instead of CIoU. SIoU considers the angular loss between the predicted and ground truth boxes, which is the minimum angle between the central points of the predicted box and the ground truth box, as illustrated in Fig.5.

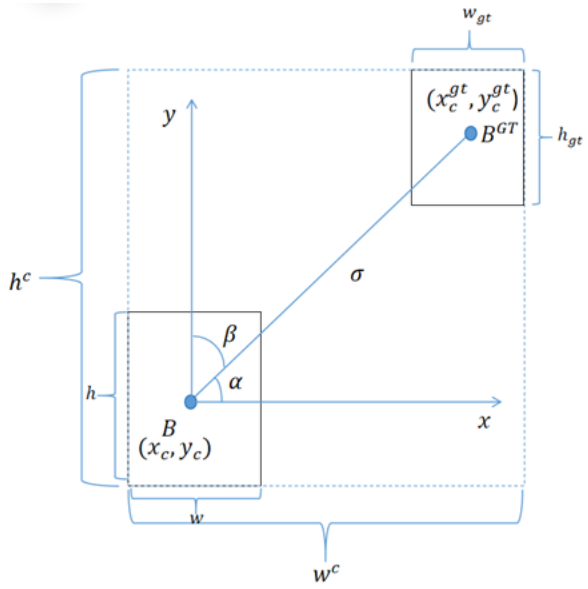

Figure 5. Angular loss in the SIoU loss function

Among them, the angular loss is expressed by Equation (1):

$$\Lambda = \sin\left(2\sin^{-1}\frac{\min\left(\left|x_c^{gt}-x_c\right|,\left|y_c^{gt}-y_c\right|\right)}{\sqrt{\left(x_c^{gt}-x_c\right)^2+\left(y_c^{gt}-y^c\right)+\dot{o}}}\right) \tag{1}$$

To be specific, x_c^{gt} and y_c^{gt} represent the center coordinates of the ground truth box, while x_c and y_c denote the center coordinates of the predicted box, and ε is a small constant used to avoid division by zero, which could lead to computational errors. The purpose of the angular loss is to guide the predicted box to the nearest axis, considering whether to prioritize approaching the x-axis or the y-axis based on the angle of change. In addition to the angular loss, SIoU also takes into account the distance loss Δ and the shape loss Ω, which are represented by Equation (2) and Equation (3), respectively:

$$\Delta = \frac{1}{2}\sum_{t=x,y}\left(1-e^{-\gamma\rho t}\right), \gamma = 2-\Lambda \tag{2}$$

$$\Omega = \frac{1}{2}\sum_{t=w,h}\left(1-e^{\omega t}\right)^{\theta}, \theta = 4 \tag{3}$$

$\rho_x = \left(\frac{y_c^{gt}-y_c}{w^2}\right)^2, \rho_y = \left(\frac{y_c^{gt}-y_c}{h^c}\right)^2$ where w^c and h^c represent the width and height of the predicted box, respectively. In summary, the SIoU regression loss function is given by Equation (4), and the total loss function is represented by Equation (5):

$$L_{SIoU} = 1 - IoU + \frac{(\Delta+\Omega)}{2} \tag{4}$$

$$L = W_{box}L_{box} + W_{cls}L_{cls} \tag{5}$$

3. Experimental Analysis

3.1. Dataset Source and Introduction

This paper conducted experiments on two public datasets: the BCCD dataset and the anemia detection v3i dataset, which is open-sourced by Roboflow. The paper followed a similar approach to CST-YOLO in randomly dividing the training, validation, and test sets. Specifically, the training set, validation set, and test set were divided in a ratio of 8:1:1. Fig.6 and Fig.7 respectively showcase examples of blood cell images from the BCCD dataset and the anemia detection v3i dataset.

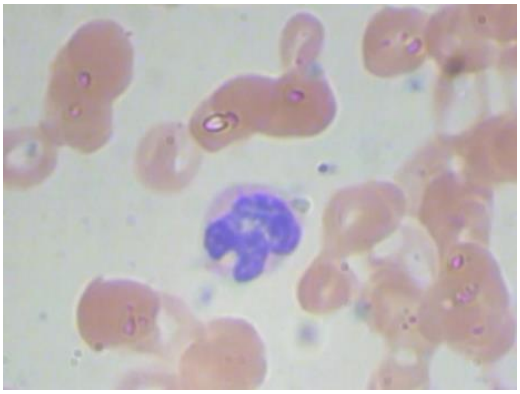

Figure 6. BCCD dataset

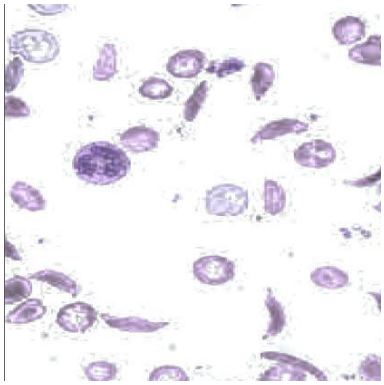

Figure 7. Anemia detection v3i dataset

3.2. Data Preprocessing

It can be observed that the BCCD dataset does not undergo offline data augmentation, whereas the anemia detection v3i dataset has already undergone prior offline data augmentation. Therefore, this paper chooses to align with CST-YOLO [15] and not employ offline data augmentation during the data preprocessing stage. However, the mosaic data augmentation method is used, which entails mixing two images at a 1:1 ratio with a 50% probability during the training process. This approach aims to better prevent model overfitting and enhance generalization capabilities. Fig.8 illustrates the results of applying mosaic data augmentation to the BCCD dataset.

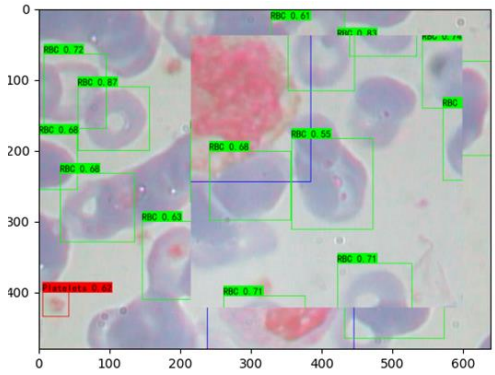

Figure 8. Mosaic data augmentation

Finally, since the categories within the two datasets are different, the BCCD dataset contains a total of three categories: WB, RBC and platelets, while the anemia detection v3i dataset adds an additional category-Sickle cell, Therefore, this paper also conducted data exploration, specifically, the article statistically analyzed the number of each type of cell in the two datasets. The results are presented in Fig.9.

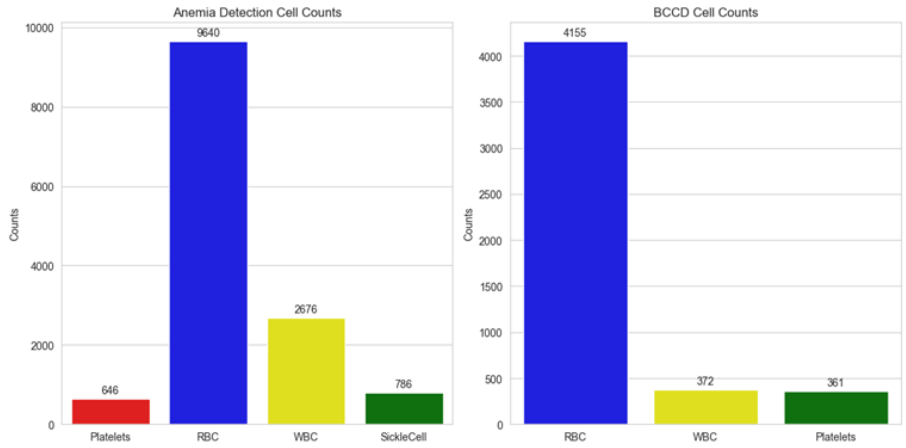

Figure 9. Blood Cell Category Statistics

It can be observed that the common characteristic of the two datasets is that red blood cells constitute the largest proportion. In the Anemia Detection dataset, white blood cells also account for a significant proportion, which is due to the fact that many labels in this dataset contain only a single white blood cell. Finally, platelets and sickle red blood cells constitute the smallest proportion. In the BCCD dataset, the proportion of white blood cells and platelets is not significantly different.

Through statistical analysis of the datasets, this paper finds that the characteristics of the two datasets align with our prior knowledge. Therefore, the datasets used in this experiment are considered reliable and interpretable.

3.3. Experimental Details

The entire process of model training, validation, and prediction was conducted on an NVIDIA RTX 2080 Ti graphics card. The training utilized the pre-trained weights provided by the YOLOv7 official repository, which were obtained from the COCO dataset. The total number of training epochs was set to 300, with an initial learning rate of 0.01 and a minimum learning rate of 0.0001. The Stochastic Gradient Descent (SGD) optimizer was employed for parameter updates, and label smoothing with a value of 0.005 was used to address the issue of class imbalance among blood cells, thereby enhancing the model's generalization capabilities. The learning rate was updated using the cosine annealing algorithm, and the training process was divided into two stages: the frozen stage (first 50 epochs) and the unfrozen stage (last 250 epochs). During the frozen stage, the parameters of the backbone network were frozen, and the input batch size was set to 8. In the unfrozen stage, the input batch size was reduced to 4.

3.4. Evaluation Metrics

Precision is a measure of the ratio of the number of cells correctly predicted (True Positives, TP) to the total number of cells predicted as positive (TP + False Positives, FP) in the task of blood cell detection. Recall refers to the proportion of cells that are truly positive and are also predicted as positive (TP) out of all the cells that should have been detected (TP + False Negatives, FN). The Average Precision (AP) is used to calculate the average accuracy of a single-class model, where N represents the total number of classes. In the blood cell detection task, the total number of classes is 3. To evaluate the performance of the object detection, this paper uses the Average Precision (AP) as a metric for each type of blood cell detected and the mean Average Precision mAP@0.5 (the average across all types) as a measure of the overall model. The formulas are as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

$$AP = \int_0^1 PRdR \tag{8}$$

$$mAP = \frac{\sum_{i=0}^{n} AP_i}{n} \tag{9}$$

3.5. Ablation Experiments

To demonstrate the effectiveness of each improvement in our model, the article has also conducted an ablation study. The results are presented in the following table.1.

Table 1. Ablation Experiment

Dataset	Model	WBC	RBC	Platelets	Average	Performance Improvement
BCCD	YOLOv7	0.977	0.829	0.883	0.896	
	+ViT	0.981	0.813	0.901	0.903	0.7%
	++CondConv	1.0	0.829	0.904	0.911	0.8%
	+++SIoU	1.0	0.825	0.941	0.922	1.1%
	++++4x_upsample	1.0	0.822	0.967	0.930	0.8%

3.6. Comparative Experiments

In the BCCD dataset, our model was compared with YOLOv5x, YOLOv7, and CST-YOLO. The following table.2 presents a performance comparison between our model and the aforementioned models. The results indicate that, in terms of the mAP@0.5 metric, our model achieves 3.4% improvement over the original YOLOv7, a 0.7% improvement over the larger-parameter YOLOv5x, and a 0.3% improvement over the currently strongest open-source model on this dataset, CST-YOLO. Additionally, our model also outperforms the three aforementioned models in the detection of platelets and white blood cells, which are of greater importance.

Table 2. Performance Comparison of Different Models on the BCCD Dataset

Dataset	Model	WBC	RBC	Platelets	Average
BCCD	YOLOv5x	0.977	**0.877**	0.915	0.923
	YOLOv7	0.977	0.829	0.883	0.896
	CST-YOLO	0.984	0.869	0.928	0.927
	Ours	**1.0**	0.822	**0.967**	**0.930**

This paper compares the prediction results of YOLOv7, CST-YOLO, and our proposed model, as shown in Fig.10.

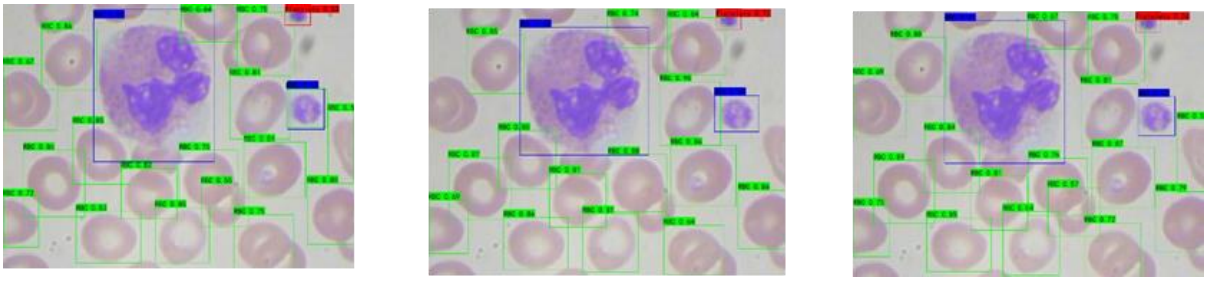

Figure 10. Detection Results of YOLOv7, CST-YOLO, and Our Proposed Model

From the above figure, it can be observed that, overall, the prediction results of the three models are quite good, especially for WBC (white blood cells) which are relatively large. However, our study found that the YOLOv7 model occasionally has overlapping predictions when predicting a single white blood cell, while the CST-YOLO and our proposed model do not exhibit this issue.

Secondly, for the Anemia Detection dataset, the article conducted a single comparative experiment, which is the comparison between YOLOv7 and our proposed model. The experimental results are shown in the following table.3.

Table 3. Performance Comparison of Different Models on the Anemia Detection Dataset

Dataset	Model	WBC	RBC	Platelets	Sickle_Cell	Average
Anemia Detection	YOLOv5x	0.989	0.811	0.863	0.665	0.832
	YOLOv7	**0.996**	0.794	0.860	0.634	0.821
	YOLOv8	0.991	**0.822**	0.859	0.561	0.808
	Ours	0.995	0.797	**0.880**	**0.763**	**0.859**

The experimental results show that our model has achieved a 3.8% improvement in the mAP@0.5 metric compared to the original YOLOv7. Moreover, in the four category metrics, our model demonstrates a significant improvement in the detection capability of sickle red blood cells, which

are the most difficult to detect. This illustrates the enhancement of our model over the original YOLOv7. Fig.11 is the detection results of the two models (Left: YOLOv7, Right: Improved Model).

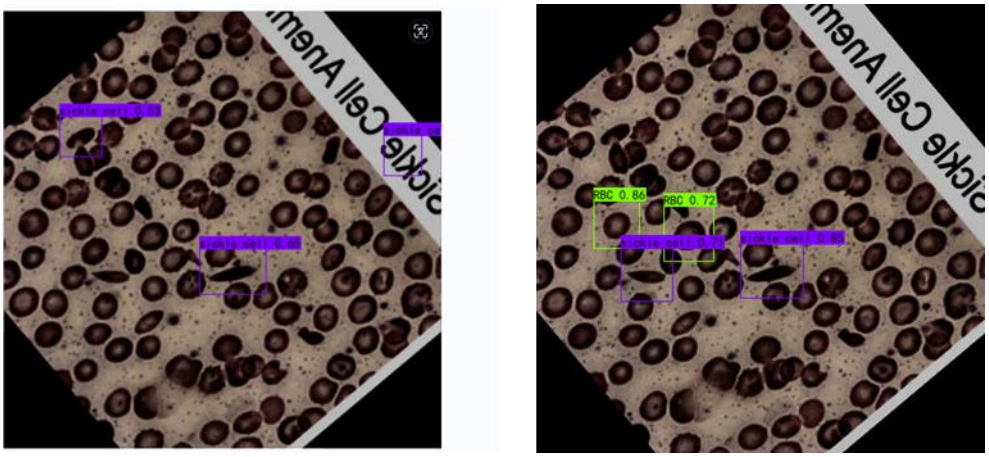

Figure 11. Detection Results Before and After Model Improvement

It can be seen that the original YOLOv7 model suffers from severe false positives and false negatives, while our proposed model does not exhibit any significant false positive issues.

4. Conclusion

The article incorporates Vision Transformer modules into the backbone network of YOLOv7, allowing us to leverage the original pre-trained weights of YOLOv7 to accelerate model convergence. The article replaces the SPPCSPC layer in the backbone network with parametric convolutions, which enables the model to dynamically adjust the convolution kernels according to the characteristics of the input data, enhancing its adaptability to various data distributions and task requirements. Furthermore,wehe article quadruple the upsampling in the neck network, which allows the global feature information learned from the small-scale feature layer to be directly added to the largest-scale feature layer, ensuring that all three output feature layers learn the global features of the image. Additionally, the article replaces the CIoU loss function with the SIoU loss function, which considers the angular loss between the predicted and ground truth boxes. Finally, through comparative experiments on public blood cell detection datasets, the article demonstrates that our method significantly improves detection performance and can be effectively applied to blood cell detection tasks.

This model provides an efficient and reliable solution for blood cell detection, offering substantial practical value in medical image analysis and clinical diagnosis. Looking forward, the article is committed to training and validating the model on a wider and more diverse range of datasets to enhance its adaptability to various biological characteristics and complexities. We also plan to explore more advanced network architectures to improve the model's performance in recognizing multiple blood cell types and handling different pathological conditions. Moreover, we will attempt to apply this model to other related fields to assess its potential in broader application scenarios.

References

[1] Patil P R, Sable G S, Anandgaonkar G. Counting of WBCs and RBCs from blood images using gray thresholding [J]. International Journal of Research in Engineering and Technology, 2014, 3 (4): 391-395.

[2] Raina R, Gondhi N K, Chaahat, et al. A systematic review on acute leukemia detection using deep learning techniques [J]. Archives of Computational Methods in Engineering, 2023, 30 (1): 251-270.

[3] Dhieb N, Ghazzai H, Besbes H, et al. An automated blood cells counting and classification framework using mask R-CNN deep learning model [C] // 2019 31st international conference on microelectronics (ICM). IEEE, 2019: 300-303.

[4] Xia T, Fu Y Q, Jin N, et al. AI-enabled microscopic blood analysis for microfluidic COVID-19 hematology [C] // 2020 5th International Conference on Computational Intelligence and Applications (ICCIA). IEEE, 2020: 98-102.

[5] Mao Y, Zhang H, Wu W, et al. DWS-YOLO: A Lightweight Detector for Blood Cell Detection [J]. Applied Artificial Intelligence, 2024, 38 (1): 2318673.

[6] Shi C, Zhu D, Zhou C, et al. Gpmb-yolo: a lightweight model for efficient blood cell detection in medical imaging [J]. Health Information Science and Systems, 2024, 12 (1): 24.

[7] Liu Z, Yuan D, Zhu G. Research on Blood Cell Detection and Counting Based on YOLO-BC Algorithm [J]. Available at SSRN 4676325, 2024.

[8] Huang M, Wang B, Wan J, et al. Improved blood cell detection method based on YOLOv5 algorithm [C] // 2023 IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, 2023, 6: 992-996.

[9] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors [C] // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 7464-7475.

[10] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv: 2010.11929, 2020.

[11] Yang B, Bender G, Le Q V, et al. Condconv: Conditionally parameterized convolutions for efficient inference [J]. Advances in neural information processing systems, 2019, 32.

[12] Rezatofighi H, Tsoi N, Gwak J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression [C] // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 658-666.

[13] Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression [C] // Proceedings of the AAAI conference on artificial intelligence. 2020, 34 (07): 12993-13000.

[14] Gevorgyan Z. SIoU loss: More powerful learning for bounding box regression [J]. arXiv preprint arXiv: 2205.12740, 2022.

[15] 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 [J]. arXiv preprint arXiv: 2306.14590, 2023.