

RESEARCH ARTICLE

Automatic Blood Cell Detection Based on Advanced YOLOv5s Network

YINGGANG HE 

Chengyi College, Jimei University, Xiamen, Fujian 363000, China

e-mail: heyinggang@jmu.edu.cn


This work was supported by the Education Scientific Research Project of Young Teachers in Fujian Province under Grant JAT191161.

ABSTRACT There is a great demand for automatic detection and classification of blood cells (BCs) in clinical medical diagnoses. Traditional methods, such as hematology analyzer and manual counting were laborious, time intensive, and limited by analysts' professional experience and knowledge. In this paper, the one-stage network based upon improved YOLOv5s is provided to detect BCs. First, the Transformer and bidirectional feature pyramid network (BiFPN) are introduced into the backbone network and neck network for refining the adaptive features, respectively. Second, Convolutional Block Attention Module (CBAM) is added to neck network outputs to strengthen the key features in space and channel. In addition, an Efficient Intersection over Union (EIoU) was introduced to improve model accuracy regarding localization and performance. The improvements are embedded into the YOLOv5s model and termed YOLOv5s-TRBC. The experiments on the blood cell dataset (BCCD) show that in the three types of BCs detections, the mean average precision (mAP) of the method proposed reached 93.5%. Furthermore, comparative experiments demonstrate that the model could perform favorably against the counterparts with respect to mAP rate, and the model's Giga Floating-point Operations Per Second (GFLOPs) is reduced to 1/6 of YOLOv5, which provides a potential solution for future computer-aid diagnostic systems.

INDEX TERMS Blood cell detection, YOLOv5s, BiFPN, convolutional block attention module, transformer.

I. INTRODUCTION

Complete blood count (CBC) is known as full blood examination (FBE) or full blood count (FBC), which is a common medical diagnostic examination that provides the percentage of cells in the blood. The human blood is composed of plasma and cellular components that contain thrombocytes (platelets), erythrocytes (or red blood cells (RBCs)), and leukocytes (or white blood cells (WBCs)). The primary function of RBCs is delivering oxygen to and taking back CO₂ away from the tissues via blood flow. WBCs are an important component of the immune system that defends against infection and diseases. The coagulation mechanism of platelets helps blood to clot and recover wounds. CBC reports the numbers and types of RBCs, WBCs, platelets, and hemoglobin. In general, an abnormal change in the count of BCs type may be related to a type of illness [1]. So, doctors can infer and judge a person's health by analyzing various features of BCs as well as their counts.

The associate editor coordinating the review of this manuscript and approving it for publication was Ramakrishnan Srinivasan .

BCs detection and identification technology could help doctors effectively in disease diagnosis, including that of malaria, dengue, anemia, infections, leukemia, and so forth. [2]. For example, A low RBCs count means anemia [3]. Thrombocytopenia characterized by abnormally low levels of platelets, is a feature of acute leukemia and aplastic anemia. For WBCs, an abnormally high WBC count often occurs in infections and inflammation. For patients undergoing chemotherapy or radiation therapy, monitoring of BCs counts is essential for them, because these treatments lead to a decrease of the BCs production in bone marrow.

Counting of BCs was widely used for blood tests in clinical settings. The procedure included the BCs classification and detection. The traditional method regarding BCs detection is hemocytometer, hematology analyzer, and manual counting [4]. Although the CBC can be completed automatically by laboratory equipment or hematology analyzer, manual counting is essential to confirm abnormal results. However, manual counting requires high skill and experience for clinical laboratory analysts and is time-intensive, imprecise, tedious, and fallible. So, an automated, convenient, and effective system is

required. The CBC by blood smear images functions importantly to diagnose diseases and check human health status. With advancements regarding deep learning (DL) techniques, the accuracy and robustness of object detection in computer vision are more and more indispensable. Many investigators have begun exploring DL techniques to assist the automatic detection based on blood smear images.

This study aimed to develop a one-stage detector based on improved YOLOv5s to resolve the issues of BCs detection and classification for CBC whose cell density represents a challenge for computer vision. Present study aims to detect and classify WBCs, RBCs, and Platelets in a given blood smear image through DL. Transformer, BiFPN, and CBAM were integrated into YOLOv5s, so as to improve the model accuracy with a small amount of additional computation.

The main contents of this presentation contain:

- 1) Transformer encoder was put into the YOLOv5s backbone to ameliorate the network to get the global information.
- 2) The BiFPN was added to the neck and enhanced the path aggregation network (PANet). The BiFPN could improve the network to integrate richer semantic features and spatial information.
- 3) The CBAM attention was integrated into the original YOLOv5s network, which helps the network to focus on the key information.

The paper is organized as follows. Section II introduces the research background and relevant works. Section III describes the principle and structure of the proposed network for BCs detection. Section IV depicts the dataset and experimental setup. Section V evaluated and analyzed the experiments. Section VI discusses the outputs. Section VII presents the conclusions and provides an outlook for future research directions.

II. RELATED WORKS

In recent years, there have been various ways of CBC through blood smear images. Some researchers applied traditional machine learning (ML) algorithms and computer visual techniques for RBCs or WBCs detections.

In [6], Khodashenas et al. utilized the Otsu thresholding method to binarize images in HSV color space and annotate the white components as WBC. However, the method didn't consider the structural characters of different BCs. In [7], Acharya et al. uses a modified watershed transform to separate RBC. Then A color-based image segmentation applying the K-medoids algorithm is performed. Finally, the proposed model extracted features utilizing the region props function and fed them into the decision tree classifier. The classification rules were generated by the decision tree. Biswas and Ghoshal [8] use the Sobel filter to perform BCs edge detection. Their method can detect cells well but can't recognize the type of BCs.

Artificial neural network (ANN) belongs to ML, which is the pioneer of DL technology. Simge Çelebi and colleagues [9] implemented an ANN-based algorithm to

characterize six types of BCs. They use the color feature to seek the BCs center positions, convert RGB into Grayscale, and find the edge of cells by using Otsu thresholding. Then, possible rectangle regions of cells are predicted by these features. At last, the Artificial neural network with convolutional layers is performed to classify the type of cell in each region.

After G.E. Hinton introduced the concept of DL, researchers investigated and proposed different DL-based object detection for BCs detection and classification. Usually, the Object detectors employ convolutional neural network (CNN) to extract features. For example, [10] proposes CNN-based framework to classify BCs automatically. The framework includes convolution, max pooling, and fully connected layers.

Present detection networks could be grouped into three categories: two-stage detector based upon region proposals, one-stage detector based upon regressions, and object detector based on anchor-free [11], [12]. YOLO [13], SSD [14], and RetinaNet [15] are typical one-stage detector models. The representative two-stage detector contains Fast R-CNN [16], Faster R-CNN [17], and R-FCN [18]. Representative anchor-free object detector includes CornerNet [19], CenterNet [20], MatrixNet [21], FCOS [22], and RepPoints [23]. Two-stage detector consists of two parts: region of interest generation and candidate box regression. Furthermore, the one-stage directly detects and predicts the target without the region proposal step. Therefore, the two-stage detector achieves higher accuracy than the one-stage detector, whereas one-stage detectors have the advantage in inference speed.

Based on object detectors described above, Shakarami et al. [24] proposed an improved YOLOv3 [25], which uses EfficientNet, Dilated Convolution, and Depthwise Separable Convolution for BCs detection. They got a mean average precision (mAP) of 89.86%. Alam and Islam [4] proposed an approach utilizing various CNN architectures embedding YOLO algorithm to capture 3 classes of BCs. They achieved a mAP of 74.37% with ResNet50. Chen et al. [26] used a single shot detector (SSD) to automatically identify and calculate various BCs. They applied Resnet50 as backbone network and reached a mAP of 77.47%. These models do not fully utilize and fuse multi-layer feature maps, which limits the performance of the model. Inspired by Region-CNN [27], [28], Ruberto et al. combined the edge boxes region proposal method and knowledge-based strategy to detect RBCs and WBCs [29]. Their model improved the accuracy, while the detection speed was low.

Lee and colleagues [30] adopted VGG16 to extract features from blood smear images and introduced Region Proposal Network (RPN) to hypothesize BCs locations. CBAM [31] was put to improve the model accuracy. Experiment results show that their model has limited when BCs overlap with each other. Liu et al. [32] proposed an improved YOLOv3 with multiscale fusion and applied it to Platelet Detection. Xia et al. [33] adopted transfer learning to extract features of blood microscopic images, and then utilized Faster-RCNN network for the detection of WBCs. Data validated that the

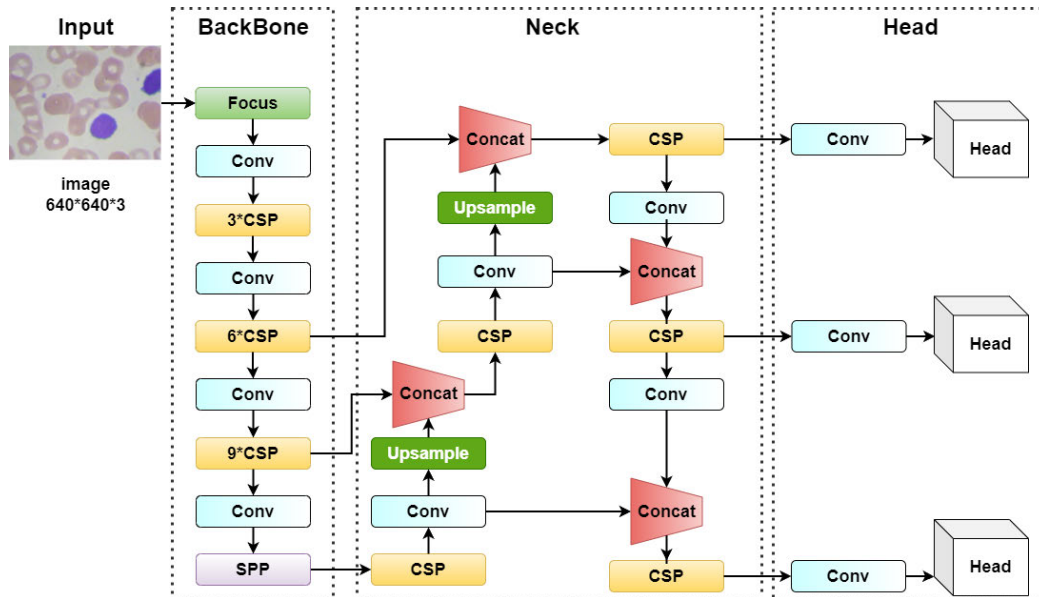


FIGURE 1. Model structure of YOLOv5. Conv, CSP, SPP, and concat are the convolution layer, cross stage partial structures, spatial pyramid pooling, and concatenation respectively.

method that proposed achieved over 90% precision. However, compared to other one-stage networks, the Faster-RCNN has huge parameters and will lower the speed of detection. Habibzadeh et al. [34] designed a CNN with architectures of LeNet5, to solve the problem of classification of normal WBC. Their experimental results indicate that convolutional neural network improved recognition accuracy even for low-quality images.

Attention mechanism [35] which simulates the human brain to focus on the importance of information, has been known to be an effective approach to advance model performance. Huang et al. [36] added CBAM to YOLOv5 framework's backbone network and bidirectional feature pyramid network (BiFPN) to the neck network and improved the BCs detection accuracy by 89.9%. Meanwhile, Gu and Sun [37] introduced Transformer [35] encoder block and CBAM attention into YOLOv5 frameworks. The proposed method improves the network's performance of distinguishing BCs in cell-dense areas and achieves a high accuracy.

Small object detection is often a challenge for many object detection models. In terms of improving the detection effect of small targets, TPH-YOLO [57] combines Transformer, CBAM, and YOLOv5, and achieves good results in remote sensing image target detection. Liang et al. [58] introduced transformer and BiFPN into the YOLOv5 to enhance the multi-scale feature fusion and improve the recognition accuracy of small objects.

III. METHODS

This section provides an accurate description regarding implementations of the proposed networks. The network was developed based on the framework of YOLOv5s and improved the mAP of the BCs detection task.

A. YOLO ARCHITECTURE

YOLO [13] is a one-stage detector algorithm, which achieved a good balance between accuracy and execution time and was used widely in many industrial scenarios and research. It has been developed into many versions, such as YOLOv2 [38], YOLOv3 [25], and YOLOv4 [39] et al. The latest version is YOLOv7.

The YOLOv3 achieved an outstanding performance improvement owing to adopting Darknet-53 structure as the backbone network to obtain features, and Feature Pyramid Networks for multi-scale feature fusion. YOLOv4 proposes PANet [5], spatial pyramid pooling (SPP) [40], Mish activation function, and self-adversarial training, along with other technologies to improve detection precisions. The backbone network employs CSPDarknet53, which incorporates the Cross Stage Partial Network (CSPNet) [41], and reduces the calculation amounts maintaining high precisions.

YOLOv5 was released by Ultralytics in June 2020. YOLOv5s, YOLOv5m, and other versions were developed based upon YOLOv5. The YOLOv5 structure is illustrated in Fig. 1. The whole network is divided into 3 parts: Backbone network, Head network, and Neck network. Input images are transferred to the backbone network (Backbone) to extract features. Through feature pyramid network (FPN) [42] and PANet network (Neck) to accomplish the feature map fusions from multi-layers with different scales. Finally, three branches of the output of the neck network were sent into the prediction network (Head) to predict the bounding box, category, as well as confidence.

YOLOv5 adopted spatial pyramid pooling (SPP) [40] to promote feature extractions of backbone networks. Additionally, In the neck network, YOLOv5 used a combination of FPN and PANet network structures. The FPN conveys top-down semantic information, and the PANet conveys

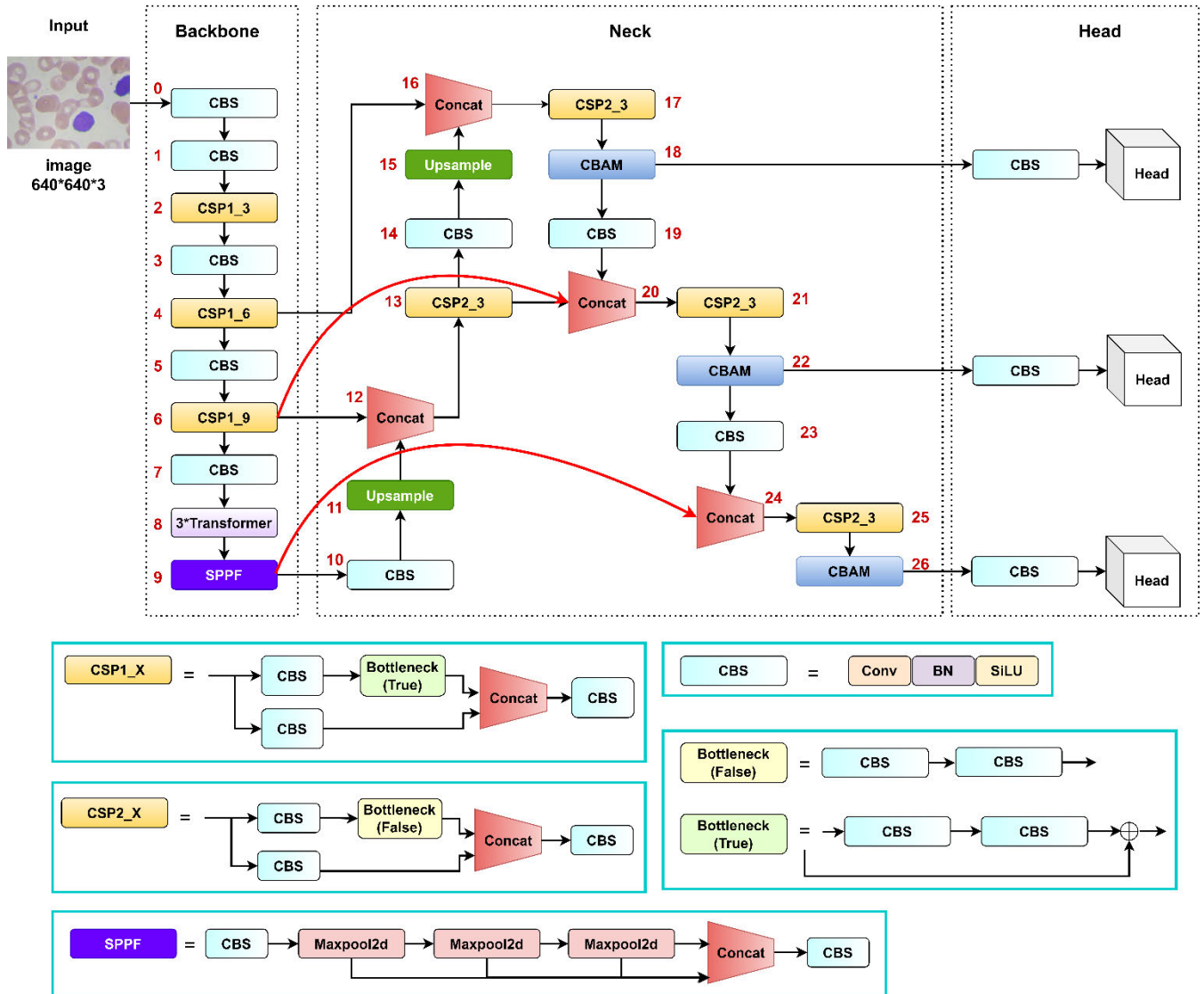


FIGURE 2. Flowchart diagram of the proposed method for BCs detection. a) Backbone with three transformer encoder blocks (block 8) at the end. b) The neck uses the structure of a bidirectional feature pyramid network (BiFPN). c) Three convolutional block attention modules (CBAM) were integrated into the neck at the end.

top-down localization information. Then, the two are concatenated to improve the effectiveness of feature fusion. Finally, the three output feature maps of Neck were fed into the Head network for the classes prediction and bounding boxes prediction separately.

The selection and optimization of the loss function is essential for models based on Neural Networks. The loss function predicts different degrees between the actual value and model predictions. The YOLOv5 loss function contains bounding box loss, classification loss, and confidence loss. CIoU loss [43] is utilized to calculate bounding box loss. Binary Cross Entropy (BCE) loss is used to obtain confidence and classification loss. Additionally, weighted non-maximum suppression (NMS) operation is performed to

filter the object detection anchor boxes and locate the target position precisely [44].

B. YOLOv5s NETWORK IMPROVEMENT

Compared to other versions of YOLOv5, the YOLOv5s was lightweight and achieved a good balance between precision and speed. So, this study proposed an improved YOLOv5s-based network for BCs detection and classification task, termed as YOLOv5s-TRBC (Transformer + BiFPN + CBAM). Current work focuses on three works, which are the backbone network and the neck network optimizations, attention mechanism embeddings, and loss function optimization. Fig. 2 depicts the YOLOv5s-TRBC network structure.

1) INTRODUCTION OF TRANSFORMER ENCODER IN BACKBONE

As shown in Fig. 2, the overall structure is built based on YOLOv5s. The original Conv located before SPP is replaced by the Transformer encoder blocks. The transformer [35] was first introduced in natural language processing (NLP) and achieved significant progress in many NLP tasks. Vision Transformer (ViT) [45] proposed by Google, is redesigned and transferred to computer vision. It splits an image into patches, linearly tokenizes each of the patches, and processes them through transformer encoders. In addition, ViT uses self-attention to integrate features across the image and learn the correlations between patches. The transformer encoder structure is shown in Fig. 3. Combining Transformer to the model, helps the backbone network to learn the relationship between objects and improves the capability of the backbone network to detect global information as well as upper-level features.

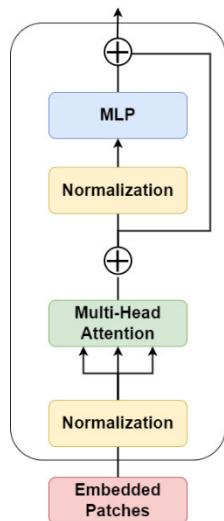


FIGURE 3. Architecture of transformer encoder [35].

2) IMPROVEMENT OF THE NECK NETWORK

The mechanism of the neck of the YOLOv5 is PANet. The PANet has a top-down pathway and an extra bottom-up path aggregation network to fuse multi-scale features. Although the PANet structure is efficient, it results in higher computational costs. Based on the PANet, structure, the BiFPN [46] adds additional interactions between output and input nodes at the same level directly and removes the nodes of single input edges to reduce computation. Therefore Bi-FPN has fewer parameters and computation than PANet, and makes the prediction network more sensitive to objects with different resolutions. Fig. 4 shows the difference between PANet and BiFPN. Accordingly, the PANet at the Neck network of YOLOv5 was replaced with BiFPN to improve the network performance as well as reduce parameters.

3) ATTENTION MECHANISM EMBEDDING

In most computer vision jobs, it is significant to obtain key features from the complex backgrounds of an image.

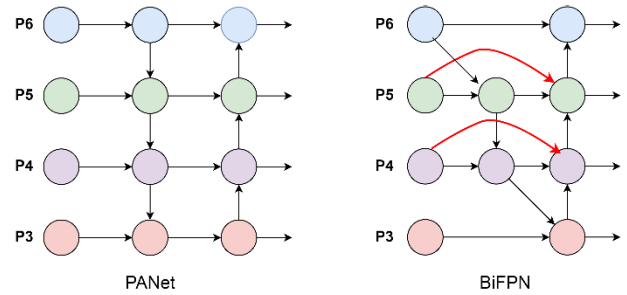


FIGURE 4. Schematic of the different feature fusion structures. PANet [5] (left) adopts a top-down pathway to fuse multi-scale feature maps (P3 – P6); BiFPN [46] (right) adds a residual connection between the original input and output node, which is shown in the figure by the red arrow.

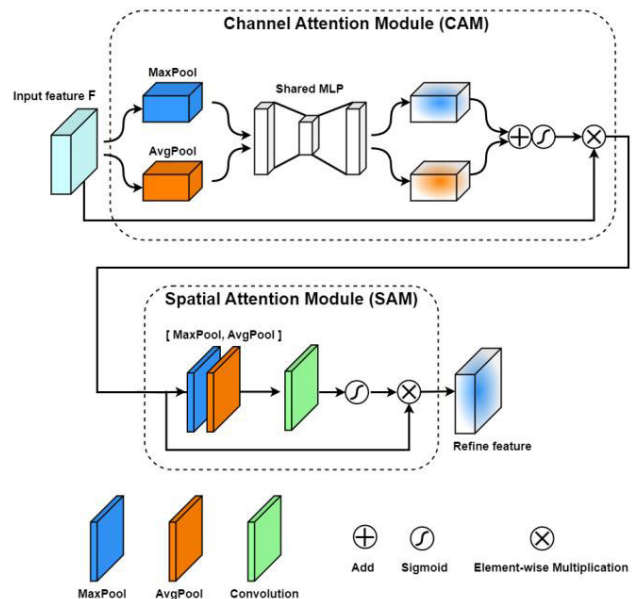


FIGURE 5. Overview of CBAM. The module includes channel attention module (CAM) and a spatial attention module (SAM) [31].

Convolutional block attention module (CBAM) [31] is a lightweight and efficient attention module, which can be easily integrated into many DL models. As shown in Fig. 5, the CBAM combines spatial attention (SA) and channel attention (CA) modules. SA is helpful for the model to capture the structure of the object. CA can help the model to focus on essential and significant colors. The CBAM attention operation is formulated as Equation (1).

$$F' = M_{CAM}(F) \otimes F, \\ F'' = M_{SAM}(F') \otimes F' \quad (1)$$

In (1), F denotes the input feature map, \otimes is an element-wise multiplication, M_{CAM} and M_{SAM} denote CA extraction operation and the spatial dimension extraction operation, respectively. Fig. 5 illustrates the CAM and SAM processes.

As given in Fig. 5, the input feature map is subject to max pooling and average pooling. Afterward, a multi-layer perception (MLP) network, element-wise summation, and

the Sigmoid activation were executed sequentially to achieve CA. The CAM operation for the attention weight M_c can be represented in Equation (2).

As provided in Fig. 5, the input feature map is subject to max pooling and average pooling. Multi-layer perception (MLP) network, element-wise summation, and the Sigmoid activation were executed sequentially to achieve CA. The CAM operation for the attention weight M_c can be represented in Equation (2).

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ = \sigma(W_1 \left(W_0 \left(F_{avg}^c \right) \right) + \sigma(W_1 \left(W_0 \left(F_{max}^c \right) \right)) \quad (2)$$

In (2), F denotes the input feature map, σ is the sigmoid function, W_0 and W_1 are MLP weights.

SAM takes the channel-refined feature as input. The SAM performs channel-wise compression via average pooling and max pooling operations to obtain 2 feature matrices: F_{avg}^s, F_{max}^s . Then, the F_{avg}^s and F_{max}^s were concatenated and fed into the convolution layer with sigmoid activation function. The computation is mathematically expressed in Equation (3):

$$M_s(F) = \sigma(f^{7 \times 7} ([AvgPool(F); MaxPool(F)])) \\ = \sigma \left(f^{7 \times 7} \left(\left[F_{avg}^s; F_{max}^s \right] \right) \right) \quad (3)$$

where σ is sigmoid activation function, $f^{7 \times 7}$ denotes a 7×7 convolution operation.

Concerning that various types of BCs have different shapes and colors. To effectively improve the BCs detection accuracy by CA and SA on multi-scale feature maps, three CBAM modules were integrated at the output of the neck network as given in Fig. 2.

4) DETECTION HEAD AND LOSS FUNCTION

The detection network is responsible for object detection. It uses 3 scale feature detection headers to convolve feature maps generated by the neck network and outputs 3 scales of feature maps with $20 \times 20, 40 \times 40,$ and 80×80 grids respectively. The detection head with 20×20 grid feature maps has the largest receptive field and is used to predict large-size targets. The detection Head with 40×40 grid feature maps is used to detect medium-size targets. The detection head with 80×80 grid feature maps is used to detect small-size targets.

The bounding box loss of YOLOv5's detection network consists of three parts: Intersection over Union (IoU) loss [47], center distance loss, and aspect ratio loss. IoU is defined as the ratio of the intersection to union with predicted and ground truth boxes, and can be expressed as Equation (4):

$$IoU \\ = \frac{Ground\ Truth\ Bounding\ Box \cap Predicted\ Bounding\ Box}{Ground\ Truth\ Bounding\ Box \cup Predicted\ Bounding\ Box} \quad (4)$$

On the basis of IoU [47], GIoU loss [48], DIoU loss [43], CIoU loss [43] and EIoU loss [49] were extended. And, YOLOv5 adopted CIoU to obtain the IoU loss.

Although the effectiveness of CIoU and other loss functions were demonstrated in many studies. After many comparative experiments were conducted, it was found that EIoU (Efficient Intersection over Union) is the most effective among the aforementioned methods. The formula of EIoU is shown in Equation (5).

$$L_{EIoU} \\ = L_{iou} + L_{dis} + L_{asp} \\ = 1 - IoU + \frac{\rho^2(b, b^{gt})}{(w^c)^2 + (h^c)^2} + \frac{\rho^2(w, w^{gt})}{(w^c)^2} + \frac{\rho^2(h, h^{gt})}{(h^c)^2} \quad (5)$$

where, ρ represents the Euclidean distance between the prediction box and ground truth box, w^c and h^c are the width and height of the smallest rectangle covering the prediction box and the ground truth box, b and b^{gt} are the center coordinates of predicted and ground truth boxes, respectively. w and w^{gt} are width of predicted and ground truth boxes. h and h^{gt} are heights of predicted and ground truth boxes.

EIoU loss considers the overlap area, the central point, and the aspect ratio of the geometric factors, which are essentially consistent with the morphological characteristics of the BCs. Therefore, EIoU loss function was selected to detect the network in this article.

IV. EXPERIMENTS

A. DATASETS

In this study, the proposed network was trained with the BCCD dataset [50]. The dataset contains 364 BCs images validating three various classes of cells. All images are blood smear images with a resolution of 640×480 . Fig. 6 shows two samples of blood smear images in BCCD dataset.

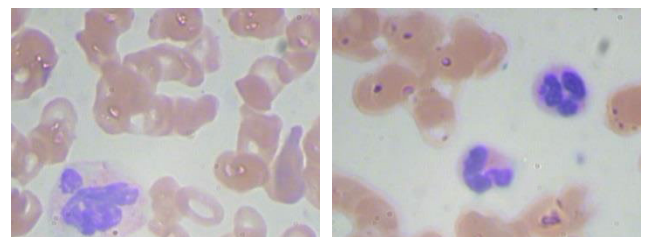


FIGURE 6. Sample blood smear Images in the BCCD dataset.

The BCCD dataset was grouped into training and testing sets with a ratio 8:2. Furthermore, K-fold cross-validation (CV) was adopted to evaluate the model reliably. Details of 3 types of BCs in two subsets are provided in Table. 1.

To improve the model's generalization capacity and robustness, data augmentation was performed by mosaic [39]. Mosaic splice four images by randomly cropping, flipping, and color gamut changes for each image, and form a new image. In addition, Mosaic enriches the background of

TABLE 1. Data profile for two subsets.

Type	Number of objects in Training Set	Number of objects in Test Set
RBC	5704	863
WBC	298	80
Platelets	378	71

the images and alleviates the problem of data imbalance of BCCD. So, the mosaic data augmentation strategy was applied in image preprocessing and enriched the input dataset for model training.

B. EXPERIMENTAL PLATFORM AND PARAMETER SETTINGS

Experiments in current research are based on Pytorch 1.12 framework and CUDA 11.6 as the parallel computing platform. The programming language is Python of version 3.9.12. The operating system of the experimental platform is Windows 10 64bit operating system, and the CPU is Intel Xeon silver 4216, and the running memory is 64GB. NVIDIA GeForce RTX3080 graphic card is used for training and testing.

The deep networks are running on a virtual environment built by Anaconda3. The training parameters related to the BCs detection model were as follows: the input image pixels were 640×640 , the batch size was 16, the number of training epochs was 300, and the initial learning rate was 0.0001. In the training phase, Adam [51] combined with a momentum optimization algorithm was adopted to train the network.

V. RESULTS

A. EVALUATION METRICS

The precision (P), recall rate (R), mAP, and GFLOPs are 4 most frequently utilized metrics to verify object detection task. So, they were adopted to measure the detection performance in this paper. The calculation formulas are expressed in Equations (6)-(9):

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$AP = \int_0^1 P(R)dR \quad (8)$$

$$mAP = \frac{1}{n} \sum_{i=1}^n AP_i \quad (9)$$

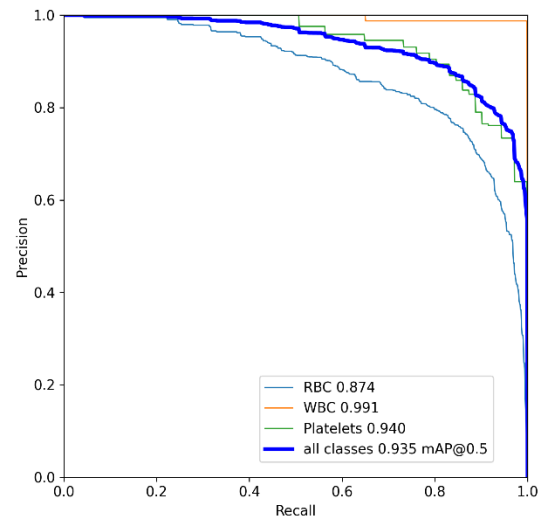
Here TP denotes true positive, FP denotes false positive, FN is the false negative, AP denotes average precision, n is the number of target categories to be detected, mAP represents the average AP value for all classified objects.

The Precision shows how many of the proposed model's predictions are correct predictions out of all the predictions made. The Recall represents the ratio of

correctly predicted positive objects to the total number of positive objects. mAP mainly assesses the recognition effect and is widely used to evaluate the detection system. $mAP@0.5$ means experiments on the mAP at the intersection of union (IoU) of 0.5. Additionally, the confusion matrix is employed to validate the model performance.

TABLE 2. BC detection performance of proposed network.

Class	RBC	WBC	Platelets	ALL
Precision (%)	80.8	98.8	82.8	87.4
Recall (%)	78.1	99.5	88.7	88.8
mAP@0.5 (%)	87.4	99.1	94.0	93.5

**FIGURE 7.** PR curves for 3 types of BCs.

B. PERFORMANCE EVALUATION

To fully utilize the training dataset, K-fold CV strategy (typically $K = 5$) was employed to validate the model reliably. The original training set was split into 5 subsets randomly. Each of sets took turns as testing set, and the remaining sets were applied as training sets. Finally, the model is evaluated via test set. The YOLOv5s-TRBC performance is presented in Table 2. The model that proposed obtained a mAP of 0.935 for the detection and identification of the three types of BCs. The precision and recall were 0.874 and 0.888, respectively. The precision-recall (PR) curve of each class is shown in Fig. 7. Horizontal and vertical coordinates are the recall and precision, respectively.

As Table 2 and Fig. 7 show, the mAP@0.5 of the WBC and platelets was higher than 0.90, influencing the model's ability to identify two types of cells effectively.

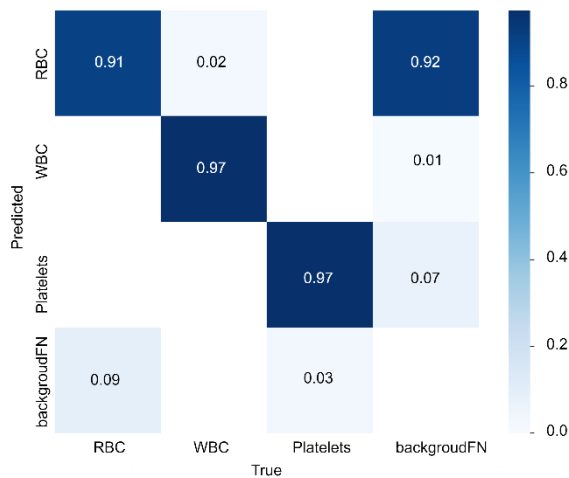


FIGURE 8. Confusion matrix on test set for YOLOv5s-TRBC.

Fig. 8 provides the Confusion matrix of the model for 3 types of BCs. It is clear that YOLOv5s-TRBC better distinguishes the BCs types in the testing sets. The prediction correct rates of RBCs, WBCs, and platelets are 91%, 97%, and 97%, respectively.

Fig. 9 displays the detection situations of the improved model. The images arranged in the first row are the detection results by YOLOv5 model. Images arranged in the next row are the detection results by YOLOv5s-TRBC. Images listed in the first column are the detection results of all three types of BCs. The images listed in the next column are WBCs detection results. The last columns list the platelets detection result images. The detection effect in Fig. 9 clearly shows that the YOLOv5s-TRBC detects almost all of the RBCs, WBCs, and Platelets. However, as shown in Fig. 10, due to the larger count of RBCs in the blood, and the phenomenon of cellular overlap between RBCs, the problem of missed detection occurs in both models occasionally. Compared to YOLOv5 model, the improved YOLOv5s-TRBC model performs well in the detection of the Platelets and WBCs. For small target detection, YOLOv5 misses a lot of small targets, while YOLOv5-TRBC detects more small targets.

In order to further verify the effectiveness of the proposed model, the heatmaps of the model detection were drawn by gradient-weighted class activation mapping (Grad-CAM) in a visual way. The heatmap can clearly show the regions of interest of the network by highlighting them in red. In Figure 10, three heatmaps were generated by layer 18th, layer 22th, and layer 26th at the neck network of the proposed model, respectively. As depicted in Figure 10, the brighter regions in the heatmaps exhibit the models ability to capture and localize the relevant features associated with three types of blood cells.

Fig. 11 displays the model's train/validation precision, train/validation recall performance, train/validation loss, and mAP. As you can see intuitively from Fig. 11, after 200 epochs, the loss reaches the lowest value and tends to balance.

C. DETECTION PERFORMANCE COMPARISON

To evaluate the proposed model performance, a variety of recent related works and state-of-the-art object detection models were selected to conduct experiments, including Faster R-CNN [17], CenterNet [20], YOLOv3 [25], YOLOv4 [39], YOLOv5, YOLOv7, and YOLOv8. Faster R-CNN and CenterNet are classical models for object detection and classification. Also, the proposed model was compared with some latest methods listed in the literature [24], [36], [37], [52], [53], [54].

All experiments were conducted under similar conditions, and we compared the classification performance upon the testing set. Table 3 displays the YOLOv5s-TRBC improved model's detection performance. As shown in Table 3, the YOLOv5s-TRBC model has a high performance in terms of Recall and mAP@0.5, compared to other detection models. Moreover, the GFLOPs of the proposed model is 17.0 less than YOLOv3, YOLOv4, YOLOv5, et al.

From the mAP@0.5, the comprehensive effect of the proposed model is much better than other comparative models. Compared to YOLOv5, Although the Precision of YOLOv5s-TRBC is reduced with a slight probability, the computation amount is only 1/6 of the YOLOv5. Compared to the latest SOTA object detection methods, the mAP@0.5 of YOLOv5s-TRBC (93.50%) is higher than that of YOLOv7 (91.20%) and YOLOv8 (92.20%). YOLOv5s-TRBC (87.40%) is 3.1% higher than YOLOv7 (84.30%) and YOLOv8 (84.3%) in terms of precision, but YOLOv5s-TRBC (88.8.%) is less than YOLOv7 (89.30%) and YOLOv8 (91.00%) in terms of recall.

Collectively, the YOLOv5s-TRBC model improved the rate of correct detection and showed great potential for application in the field of biomedicine.

D. ABLATION STUDIES

To explore the relative contributions of various modules in the proposed network, ablation experiments were made on the proposed model to verify the contribution of each improvement. Table 4 shows performance comparisons between ablation studies, including standard YOLOv5s, YOLOv5s with transformer, YOLOv5s with Transformer and BiFPN, YOLOv5s with BiFPN, YOLOv5s with Transformer and CBAM, YOLOv5s with Transformer and BiFPN and CBAM. As shown in Table 4, it can be observed that all designs of the proposed model could increase the mAP. The inclusion of Transformer, BiFPN, and CBAM has played a positive role in the model accuracy improvement.

(a) The replacing CSP Bottleneck blocks in the original version of YOLOv5s backbone network with transformer module increased precisions by 4.4% and mAP by 0.1% while reducing the number of GFLOPs. Compared to the original YOLOv5s, the improved model has reduced GFLOPs from 15.8 to 15.6 due to the transformer's excellent parameter compression scheme. The global contextual feature learning ability of the Transformer module can help to improve model performance.;

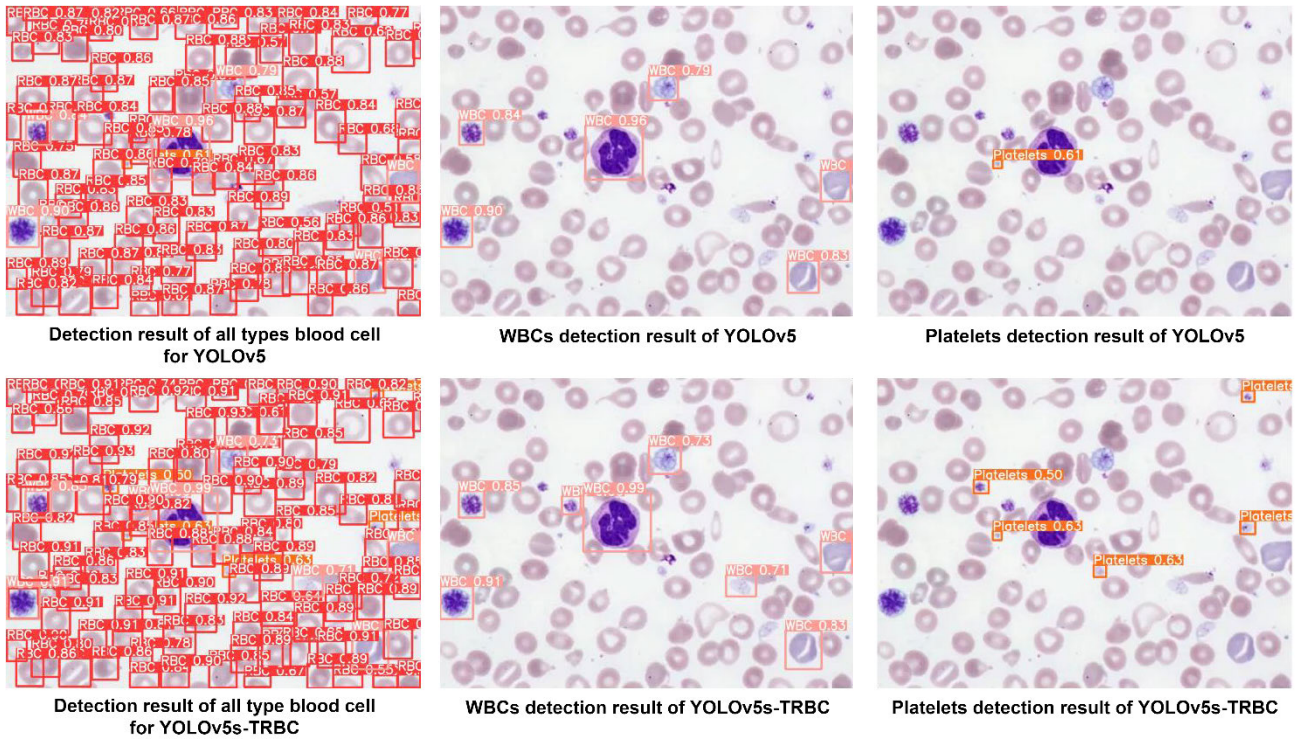


FIGURE 9. The detection results. The first lines are the detection result of YOLOv5; The second lines are the detection result of YOLOv5s-TRBC.

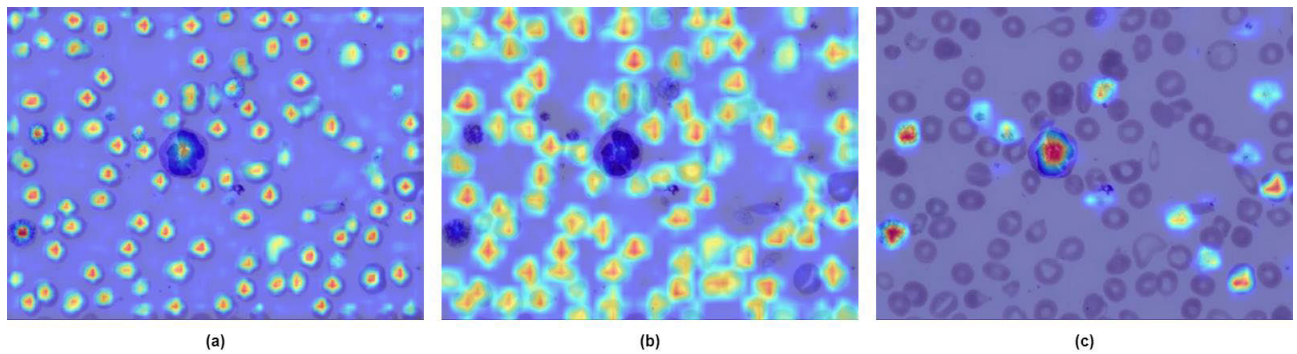


FIGURE 10. Visualization of the feature maps obtained by YOLOv5s-TRBC. (a) The heatmap of layer 18; (b) The heatmap of layer 22; (c) The heatmap of layer 26.

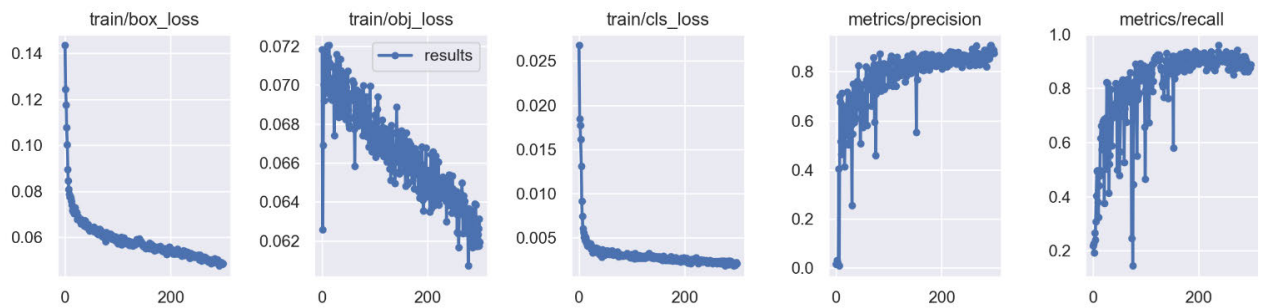


FIGURE 11. The loss, precision, recall performance of YOLOv5s-TRBC model.

TABLE 3. Comparison of different models on the BCCD dataset.

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	GFLOPs
Faster R-CNN [17]	74.00	94.33	90.42	-
CenterNet [20]	64.03	62.59	64.47	-
YOLOv3	87.30	91.00	93.00	154.6
YOLOv4	84.90	91.40	92.60	119.1
YOLOv5	87.90	87.70	92.20	107.7
YOLOv7	84.30	89.30	91.20	103.2
YOLOv8	84.30	91.00	92.20	164.8
EIoU-YOLOv5 [52]	-	91.70	92.20	-
DCBC DeepL [53]	79.90	80.30	82.40	-
FED [24]	-	-	89.86	-
Ref [54]	85.70	90.20	91.80	-
YOLOv5-CW [36]	84.00	85.70	89.90	-
AYOLOv5 [37]	86.20	91.50	93.30	-
YOLOv5s-TRBC	87.40	88.80	93.50	17.0

TABLE 4. Detection performance of multimodal fusion.

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	GFLOPs
YOLOv5s baseline	83.70	93.00	92.30	15.8
YOLOv5s+Transformer	88.10	85.80	92.40	15.6
YOLOv5s+Transformer+BiFPN	86.20	93.10	93.10	16.9
YOLOv5s+BiFPN	85.30	91.10	92.60	17.2
YOLOv5s+CBAM	87.20	90.20	91.90	15.8
YOLOv5s+BiFPN+CBAM	84.60	90.40	92.00	17.2
YOLOv5s+Transformer+CBAM	83.60	94.40	92.70	15.6
YOLOv5s+Transformer+BiFPN+CBAM	87.40	88.80	93.50	17.0

(b) As demonstrated in Table 4, experiments 3 and 4 indicate that using the BiFPN can effectively enhance the feature fusion at the neck network. The Precision and mAP@0.5 is increased by 2.5% and 0.3%, respectively. Through the analysis of the detection samples, the adoption of BiFPN has played a positive role in the model accuracy improvement.

(c) As demonstrated in Table 4, embedding CBAM to the three outputs of the neck network alone will reduce the recall and mAP@0.5 of the model. However, when the CBAM module is embedded with the transformer and BiFPN, the precision is increased by 3.7% and mAP is increased by 1.2%. The results illustrate that CBAM modules can improve the accuracy of the model by highlighting information from feature maps enhanced by transformer and BiFPN while suppressing useless information.

VI. DISCUSSION

In recent years, artificial intelligence and DL technology have been used for diagnosis in medical imaging increasingly. Automated detection and computing of BCs system based on computer vision will solve the challenge of detection speed and accuracy in medical diagnostics. However, there are two significant problems that exist for detecting BCs. The first is the diversity in the physical form of different types of BCs. The second is the presence of cellular crossover, overlap of RBCs in blood smear images makes it difficult to detect and identify. To address these issues, in this study, Transformer module, BiFPN, and CBAM were added to YOLOv5s to enhance the accuracy of BCs detection and attained a 93.5% mAP in all classes.

As shown in Fig. 9, for the proposed model, the higher the contrast in size and color, the better detection and recognition. The diameter ratio of platelets and WBCs is approximately 1:10. Therefore, the mAP@0.5 for the WBCs and platelets reached 99.1% and 94.0%, respectively. However, the proposed model has difficulties in detecting edge and dense objects. An important challenge of blood cell detection is the denseness of RBCs. The detection mAP@0.5 of RBCs only reached 87.4% due to the density and overlap of RBCs. Finally, the insufficient dataset and low-quality images also affect the improvement of model performance.

VII. CONCLUSION

Automatically detecting and identifying BCs types by computer aids can improve the doctors' work efficiency as well as accuracy and is becoming increasingly important for medical diagnosis. This work presented a BCs detection and classification network called YOLOv5s-TRBC based on YOLOv5s. The proposed network uses a combination of DL techniques in computer vision, including Transformer, BiFPN, CBAM, and EIoU. Experimental outputs showcase that the method proposed gains better detection. Meanwhile, comparative experiment data verify the efficiency of the fusion method. However, the BC detection models proposed in this work still have some shortcomings. Therefore, further optimization of this model is considered. In the future, we will strive to improve the model by seeking more effective methods or advanced model, such as using the Dynamic Head [55] to obtain the optimization of the Detection Head network, applying light-weighted GhostNet [56] to improve the detection accuracy, among other methods [59], [60], [61].

The codes implementing the described model are available at <https://gitee.com/professor98911/YOLOv5s-TRBC>.

ACKNOWLEDGMENT

The authors would like to thank TopEdit (www.topeditsci.com) for the linguistic editing and proofreading during the preparation of this manuscript.

REFERENCES

- [1] R. Green and S. Wachsmann-Hogiu, "Development, history, and future of automated cell counters," *Clinics Lab. Med.*, vol. 35, no. 1, pp. 1–10, Mar. 2015, doi: 10.1016/j.cll.2014.11.003.

- [2] D. Cruz, C. Jennifer, Valiente, L. C. Castor, C. M. T. Mendoza, B. A. Jay, L. S. C. Jane, and P. T. B. Brian, "Determination of blood components (WBCs, RBCs, and platelets) count in microscopic images using image processing and analysis," in *Proc. IEEE 9th Int. Conf. Humanoid, Nanotechnol., Inf. Technol., Commun. Control, Environ. Manage. (HNICEM)*, Dec. 2017, pp. 1–7, doi: [10.1109/HNICEM.2017.8269515](https://doi.org/10.1109/HNICEM.2017.8269515).
- [3] N. M. Deshpande, S. Gite, and R. Aluvalu, "A review of microscopic analysis of blood cells for disease detection with AI perspective," *PeerJ Comput. Sci.*, vol. 7, p. e460, Apr. 2021, doi: [10.7717/peerj-cs.460](https://doi.org/10.7717/peerj-cs.460).
- [4] M. M. Alam and M. T. Islam, "Machine learning approach of automatic identification and counting of blood cells," *Healthcare Technol. Lett.*, vol. 6, no. 4, pp. 103–108, Aug. 2019, doi: [10.1049/htl.2018.5098](https://doi.org/10.1049/htl.2018.5098).
- [5] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 8759–8768 doi: [10.1109/CVPR.2018.00913](https://doi.org/10.1109/CVPR.2018.00913).
- [6] M. Khodashenas, H. Ebrahimpour-komleh, and A. M. Nickfarjam, "White blood cell detection and counting based on genetic algorithm," in *Proc. Adv. Sci. Eng. Technol. Int. Conf. (ASET)*. Dubai, United Arab Emirates: IEEE, Mar. 2019, pp. 1–4 doi: [10.1109/ICASET.2019.8714455](https://doi.org/10.1109/ICASET.2019.8714455).
- [7] V. Acharya and K. Prakasha, "Computer aided technique to separate the red blood cells, categorize them and diagnose sickle cell anemia," *J. Eng. Sci. Technol. Rev.*, vol. 12, no. 2, pp. 67–80, Apr. 2019, doi: [10.25103/jestr.122.10](https://doi.org/10.25103/jestr.122.10).
- [8] S. Biswas and D. Ghoshal, "Blood cell detection using thresholding estimation based watershed transformation with Sobel filter in frequency domain," *Proc. Comput. Sci.*, vol. 89, pp. 651–657, Jan. 2016, doi: [10.1016/j.procs.2016.06.029](https://doi.org/10.1016/j.procs.2016.06.029).
- [9] S. Çelebi and M. Burcak Çötelı, "Red and white blood cell classification using artificial neural networks," *AIMS Bioeng.*, vol. 5, no. 3, pp. 179–191, 2018, doi: [10.3934/bioeng.2018.3.179](https://doi.org/10.3934/bioeng.2018.3.179).
- [10] P. Tiwari, J. Qian, Q. Li, B. Wang, D. Gupta, A. Khanna, J. J. P. C. Rodrigues, and V. H. C. de Albuquerque, "Detection of subtype blood cells using deep learning," *Cogn. Syst. Res.*, vol. 52, pp. 1036–1044, Dec. 2018, doi: [10.1016/j.cogsys.2018.08.022](https://doi.org/10.1016/j.cogsys.2018.08.022).
- [11] L. Jiao et al., "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128837–128868, 2019, doi: [10.1109/ACCESS.2019.2939201](https://doi.org/10.1109/ACCESS.2019.2939201).
- [12] H. Fu, G. Song, and Y. Wang, "Improved YOLOv4 marine target detection combined with CBAM," *Symmetry*, vol. 13, no. 4, p. 623, Apr. 2021, doi: [10.3390/sym13040623](https://doi.org/10.3390/sym13040623).
- [13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 779–788 doi: [10.1109/CVPR.2016.91](https://doi.org/10.1109/CVPR.2016.91).
- [14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Computer Vision—ECCV (Lecture Notes in Computer Science)*, vol. 9905. Cham, Switzerland: Springer, 2016, pp. 21–37, doi: [10.1007/978-3-319-46448-0_2](https://doi.org/10.1007/978-3-319-46448-0_2).
- [15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2017, *arXiv:1708.02002*.
- [16] R. Girshick, "Fast R-CNN," 2015, *arXiv:1504.08083*.
- [17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*.
- [18] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," 2016, *arXiv:1605.06409*.
- [19] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," 2018, *arXiv:1808.01244*.
- [20] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.
- [21] A. Rashwan, R. Agarwal, A. Kalra, and P. Poupart, "MatrixNets: A new scale and aspect ratio aware architecture for object detection," 2020, *arXiv:2001.03194*.
- [22] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," 2019, *arXiv:1904.01355*.
- [23] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point set representation for object detection," 2019, *arXiv:1904.11490*.
- [24] A. Shakarami, M. B. Menhaj, A. Mahdavi-Hormat, and H. Tarrah, "A fast and yet efficient YOLOv3 for blood cell detection," *Biomed. Signal Process. Control*, vol. 66, Apr. 2021, Art. no. 102495, doi: [10.1016/j.bspc.2021.102495](https://doi.org/10.1016/j.bspc.2021.102495).
- [25] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
- [26] Y.-M. Chen, J.-T. Tsai, and W.-H. Ho, "Automatic identifying and counting blood cells in smear images by using single shot detector and Taguchi method," *BMC Bioinf.*, vol. 22, no. S5, p. 635, Dec. 2022, doi: [10.1186/s12859-022-05074-2](https://doi.org/10.1186/s12859-022-05074-2).
- [27] S. McMahan, N. Sünderhauf, B. Uppcroft, and M. Milfordand, "How good are edge boxes, really?" in *Proc. Workshop Scene Understanding (SUNw), IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–2.
- [28] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2013, *arXiv:1311.2524*.
- [29] C. Di Ruberto, A. Loddo, and L. Putzu, "Detection of red and white blood cells from microscopic blood images using a region proposal approach," *Comput. Biol. Med.*, vol. 116, Jan. 2020, Art. no. 103530, doi: [10.1016/j.compbiomed.2019.103530](https://doi.org/10.1016/j.compbiomed.2019.103530).
- [30] S.-J. Lee, P.-Y. Chen, and J.-W. Lin, "Complete blood cell detection and counting based on deep neural networks," *Appl. Sci.*, vol. 12, no. 16, p. 8140, Aug. 2022, doi: [10.3390/app12168140](https://doi.org/10.3390/app12168140).
- [31] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," 2018, *arXiv:1807.06521*.
- [32] R. Liu, C. Ren, M. Fu, Z. Chu, and J. Guo, "Platelet detection based on improved YOLO_v3," *Cyborg Bionic Syst.*, vol. 2022, Jan. 2022, Art. no. 9780569, doi: [10.34133/2022/9780569](https://doi.org/10.34133/2022/9780569).
- [33] T. Xia, R. Jiang, Y. Q. Fu, and N. Jin, "Automated blood cell detection and counting via deep learning for microfluidic point-of-care medical devices," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 646, no. 1, Oct. 2019, Art. no. 012048, doi: [10.1088/1757-899X/646/1/012048](https://doi.org/10.1088/1757-899X/646/1/012048).
- [34] M. Habibzadeh, A. Krzyzak, and T. Fevens, "White blood cell differential counts using convolutional neural networks for low resolution images," in *Artificial Intelligence and Soft Computing (Lecture Notes in Computer Science)*, vol. 7895, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds. Berlin, Germany: Springer, 2013, pp. 263–274, doi: [10.1007/978-3-642-38610-7_25](https://doi.org/10.1007/978-3-642-38610-7_25).
- [35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.
- [36] M. Huang, B. Wang, J. Wan, and C. Zhou, "Improved blood cell detection method based on YOLOv5 algorithm," in *Proc. IEEE 6th Inf. Technol., Netw., Electron. Autom. Control Conf. (ITNEC)*. Chongqing, China: IEEE, Feb. 2023, pp. 992–996, doi: [10.1109/ITNEC56291.2023.10082206](https://doi.org/10.1109/ITNEC56291.2023.10082206).
- [37] W. Gu and K. Sun, "AYOLOv5: Improved YOLOv5 based on attention mechanism for blood cell detection," *Biomed. Signal Process. Control*, vol. 88, May 2023, Art. no. 105034, doi: [10.1016/j.bspc.2023.105034](https://doi.org/10.1016/j.bspc.2023.105034).
- [38] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," 2016, *arXiv:1612.08242*.
- [39] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
- [40] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
- [41] C.-Y. Wang, H.-Y. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, IEEE, Jun. 2020, pp. 1571–1580 doi: [10.1109/CVPRW50498.2020.00203](https://doi.org/10.1109/CVPRW50498.2020.00203).
- [42] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2016, *arXiv:1612.03144*.
- [43] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," 2019, *arXiv:1911.08287*.
- [44] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*. Hong Kong, IEEE, Aug. 2006, pp. 850–855 doi: [10.1109/ICPR.2006.479](https://doi.org/10.1109/ICPR.2006.479).
- [45] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [46] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," 2019, *arXiv:1911.09070*.
- [47] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An advanced object detection network," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 516–520, doi: [10.1145/2964284.2967274](https://doi.org/10.1145/2964284.2967274).

- [48] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," 2019, *arXiv:1902.09630*.
- [49] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and efficient IOU loss for accurate bounding box regression," 2021, *arXiv:2101.08158*.
- [50] *BCCD Dataset*. Accessed: Dec. 15, 2022. [Online]. Available: https://github.com/Shenggan/BCCD_Dataset
- [51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
- [52] Z. Zhang, Z. Deng, Z. Wu, and G. Lai, "An improved EIoU-YOLOv5 algorithm for blood cell detection and counting," in *Proc. 5th Int. Conf. Pattern Recognit. Artif. Intell. (PRAI)*. Chengdu, China: IEEE, Aug. 2022, pp. 989–993, doi: [10.1109/PRAI55851.2022.9904093](https://doi.org/10.1109/PRAI55851.2022.9904093).
- [53] Md. A. Rahaman, M. M. Ali, Md. N. Hossen, M. Nayer, K. Ahmed, and F. M. Bui, "DCBC_DeepL: Detection and counting of blood cells employing deep learning and YOLOv5 model," in *Artificial Intelligence and Data Science (Communications in Computer and Information Science)*, vol. 1673, A. Kumar, I. Fister, P. K. Gupta, J. Debayle, Z. J. Zhang, and M. Usman, Eds. Cham, Switzerland: Springer, 2022, pp. 203–214, doi: [10.1007/978-3-031-21385-4_18](https://doi.org/10.1007/978-3-031-21385-4_18).
- [54] S. Shinde, J. Oak, K. Shrawagi, and P. Mukherji, "Analysis of WBC, RBC, platelets using deep learning," in *Proc. IEEE Pune Sect. Int. Conf. (PuneCon)*. Pune, India: IEEE, Dec. 2021, pp. 1–6, doi: [10.1109/PuneCon52575.2021.9686524](https://doi.org/10.1109/PuneCon52575.2021.9686524).
- [55] X. Dai, Y. Chen, B. Xiao, D. Chen, M. Liu, L. Yuan, and L. Zhang, "Dynamic head: Unifying object detection heads with attentions," 2021, *arXiv:2106.08322*.
- [56] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," 2019, *arXiv:1911.11907*.
- [57] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*. Montreal, BC, Canada: IEEE, Oct. 2021, pp. 2778–2788 doi: [10.1109/ICCVW54120.2021.00312](https://doi.org/10.1109/ICCVW54120.2021.00312).
- [58] J. Liang, X. Chen, C. Liang, T. Long, X. Tang, Z. Shi, M. Zhou, J. Zhao, Y. Lan, and Y. Long, "A detection approach for late-autumn shoots of litchi based on unmanned aerial vehicle (UAV) remote sensing," *Comput. Electron. Agricult.*, vol. 204, Jan. 2023, Art. no. 107535, doi: [10.1016/j.compag.2022.107535](https://doi.org/10.1016/j.compag.2022.107535).
- [59] L. Li, X. Yao, X. Wang, D. Hong, G. Cheng, and J. Han, "Robust few-shot aerial image object detection via unbiased proposals filtration," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5617011, doi: [10.1109/TGRS.2023.3300071](https://doi.org/10.1109/TGRS.2023.3300071).
- [60] X. Qian, B. Wu, G. Cheng, X. Yao, W. Wang, and J. Han, "Building a bridge of bounding box regression between oriented and horizontal object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5605209, doi: [10.1109/TGRS.2023.3256373](https://doi.org/10.1109/TGRS.2023.3256373).
- [61] X. Qian, Y. Huo, G. Cheng, C. Gao, X. Yao, and W. Wang, "Mining high-quality pseudoinstance soft labels for weakly supervised object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5607615, doi: [10.1109/TGRS.2023.3266838](https://doi.org/10.1109/TGRS.2023.3266838).



YINGGANG HE was born in Fujian, Zhangzhou, China, in 1981. He received the B.Sc. degree from the School of Information Engineering, Jimei University, Xiamen, China, in 2003, and the M.Sc. degree from the School of Informatics, Xiamen University, Xiamen, in 2010. He is currently a Lecturer with the College of Chengyi, Jimei University. His research interests include computational intelligence, machine learning, and computer vision.

...