IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 December 2023, accepted 20 December 2023, date of publication 25 December 2023,
date of current version 4 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3347451

RESEARCH ARTICLE

Lymphocyte Detection Method Based on Improved YOLOv5

PEIHE JIANG[1], YI LI[1], YING LIU[2], AND NING LU[2]
[1]School of Physics and Electronic Information, Yantai University, Yantai 264005, China
[2]Pathology Department, Yantaishan Hospital, Yantai 264003, China

Corresponding author: Ning Lu (ninglu314@163.com)

This work was supported in part by the Shandong Province Natural Science under Grant ZR2019LZH016.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Clinical Trial Ethics Committee of Yantaishan Hospital.

ABSTRACT To address the limitations of traditional burdensome and time-consuming manual diagnosis of Sjogren's syndrome, this study proposes and implements an improved version of YOLOv5s algorithm, named YOLOv5s-MSS. Using YOLOv5s-MSS, we are able to detect lymphocytic infiltrative lesions in pathological images and provide assistance for pathological diagnosis. Given the small size of lymphocytes and the difficulty in distinguishing them, we made four improvements to the YOLOv5s model. Firstly, we replace the original CIOU loss function with the Focal-SIOU loss function to accelerate model convergence and improve the detection accuracy. Additionally, we introduce the multi-head self-attention module into the backbone to enhance the model's ability to capture long range dependencies and overcome the challenges posed by complex background. Furthermore, we introduce the Shuffle Attention module into the neck, which enhances the model's ability to fuse features from both spatial and channel dimensions. Finally, we remove the 1/32 downsampling section in the neck and the corresponding large object detection head. This not only enhances accuracy but also reduces parameters and model complexity. Experimental results show that YOLOv5s-MSS achieves a mAP, Precision, and Recall of 93.2%, 87.2%, and 89%, representing increases of 2.9%, 2.6%, and 2.8% compared to the original YOLOv5s model. Additionally, YOLOv5s-MSS reduces the parameters by 28.2%. These results demonstrate the effectiveness and value of YOLOv5s-MSS for lymphocyte detection.

INDEX TERMS Attention mechanism, focal-SIOU, lymphocytes detection, multi-head self-attention, YOLOv5.

I. INTRODUCTION

Sjogren's syndrome (SS) is a chronic inflammatory autoimmune systemic disease characterized by lymphocyte proliferation and progressive damage to exocrine glands [1], [2], [3], [4]. In addition to mainly affecting salivary and lacrimal glands, it can also affect multiple organ systems such as the lungs, kidneys, skin, and blood. It frequently coexists with other systemic immune diseases, such as Rheumatoid arthritis (RA), Systemic lupus erythematosus (SLE). The cause of Sjogren's syndrome remains unknown, and it may involve genetic, viral infection, sex hormone levels, and other factors. It is noteworthy that Sjogren's syndrome is not rare, with an estimated 10 million patients worldwide. The prevalence of Sjogren's syndrome in China is approximately 0.3%-0.7%, and the incidence rate increases with age. The age of onset for Sjogren's syndrome is mostly between 40 and 50 years old, but it can also occur in children. However, many patients have limited awareness of Sjogren's syndrome and often delay seeking medical treatment. In the era of big data, AI has been widely utilized in medical imaging-based diagnostic assistance. With the rapid development of digital pathology technology, AI-assisted pathological diagnosis technology is gradually emerging. At present, in the diagnosis of lung

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

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

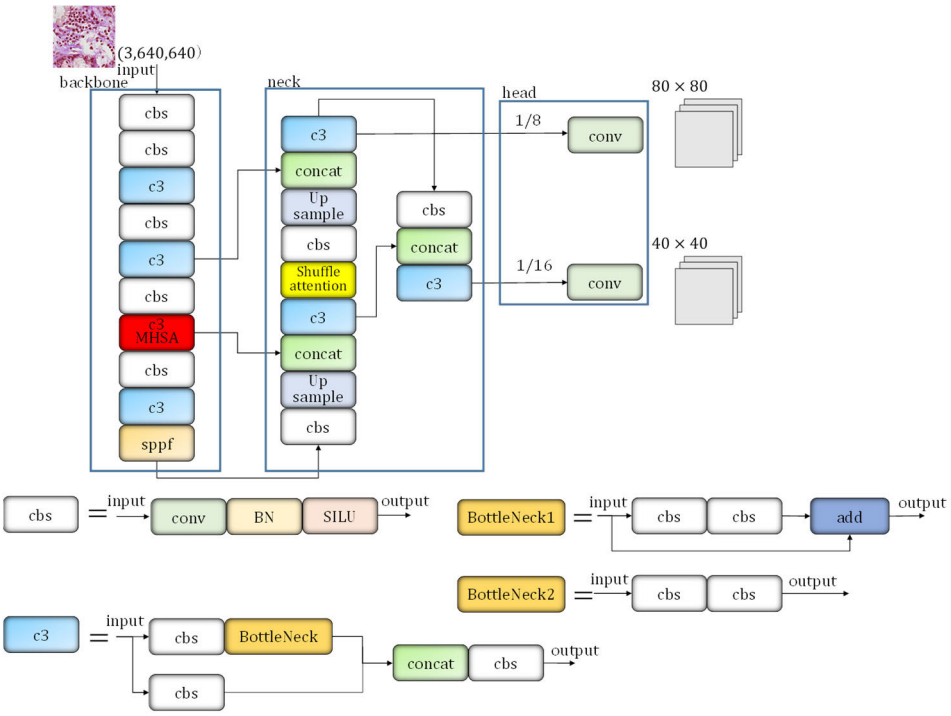

FIGURE 1. YOLOv5s-MSS framework. YOLOv5s-MSS consists of three primary parts: backbone, neck, and head. To enhance feature extraction from the input image, we introduce the C3MHSA module in the backbone to assist in processing and analysis. The Shuffle Attention module is introduced in the neck to enhance the diversity and robustness of features. Additionally, we remove the 1/32 downsampling section in the neck, along with its corresponding large object detection head. Furthermore, this figure provides a more detailed framework of the C3 and cbs modules. It should be noted that the C3 module in the backbone utilizes BottleNeck1, while the C3 module in the neck utilizes BottleNeck2.

cancer, breast cancer and other tumors, AI-assisted pathological diagnosis technology demonstrates not only efficiency, stability, and high repeatability but also a performance comparable to that of professional physicians. However, there is a lack of reports on the application of AI-assisted pathological diagnosis in Sjogren's syndrome.

In their daily work, physicians need to examine each pathological section under different magnification lenses to diagnosis Sjogren's syndrome. This process is lengthy and time-consuming. Due to the subjective heterogeneity of physicians at different levels, misdiagnosis and missed diagnosis often occur. Accurate and efficient pathological diagnosis has become a significant challenge. Our goal is to utilize an object detection algorithm to detect lymphocytic infiltrative lesions in pathological images and assist in the diagnosis of Sjogren's syndrome.

II. RELATED WORK

Object detection is a fundamental task in computer vision that aims to identify and localize objects of interest within an image or video sequence. Significant advancements have been made in the development of object detection models over the years [5], leading to improved accuracy and efficiency.

The introduction of R-CNN [6] revolutionized the field of object detection. Building on R-CNN, Fast R-CNN [7] introduced a more efficient approach by sharing computation among region proposals. It used a region of interest pooling layer to extract fixed-size features from the entire image, which were subsequently processed by fully connected layers for classification and bounding box regression. Faster R-CNN [8] achieved further improvements in speed and accuracy through the introduction of Region Proposal Network (RPN) that shared convolutional features with the detection network. YOLO (You Only Look Once) [9], [10], [11], [12], [13], [14] is a series of algorithms in computer vision that are widely used for object detection tasks. YOLO algorithms belong to the one-stage object detection category, which means that they perform classification and bounding box regression simultaneously using a single network pass. This efficiency and speed make YOLO highly attractive for real-time applications.

Recent developments in object detection include architectures such as RetinaNet [15], which addressed the problem of class imbalance in the training data, leading to improved performance. Another noteworthy approach is the Transformer-based architecture, such as DETR (DEtection TRansformer) [16] and RT-DETR (Real-Time DEtection TRansformer) [17], which utilized self-attention mechanism to perform object detection by casting it as a set prediction problem.

YOLOv5 is built on the YOLO series of algorithms and adopts a more lightweight network structure. YOLOV5 is

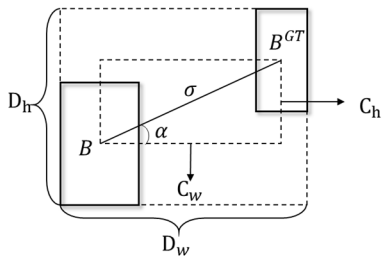

FIGURE 2. Illustration for calculating angle loss and distance loss.

well-suited for object detection tasks across diverse scenarios. Yi et al. [18] proposed the YOLO-S model for insulator and defect detection, incorporating a novel attention module, MaECA, to enhance target perception. This model replaces the original CIOU loss function with the SIOU loss function [19], effectively improving both the FPS and detection accuracy. Xu et al. [20] proposed a fire detection algorithm called Light-YOLOv5, which utilizes SepViT technology to improve the accuracy of smoke and fire detection while reducing the parameters. Additionally, this paper introduces a novel Light-BiFPN structure, which not only reduces computational costs and parameters but also enhances the fusion of multi-scale features and enriches semantic features. Zhu et al. [21] proposed an algorithm called TPH-YOLOv5, replacing the original prediction heads with Transformer Prediction Heads. This facilitates the detector accurately localize objects in high-dense scenes.

Based on the experience of the aforementioned researchers, this paper proposes the utilization of the YOLOv5s model for detecting lymphocytic infiltrative lesions in pathological images, with the aim of assisting in diagnosis. To overcome the challenges posed by the small size and difficulty in distinguishing lymphocytes, this paper introduces four improvements to the YOLOv5s model: (1) The original CIOU is replaced with the Focal-SIOU [22]. (2) The C3MHSA module is introduced in the backbone. (3) The Shuffle Attention module [23] is introduced in the neck. (4) The 1/32 downsampling section in neck is removed, along with its corresponding large object detection head. The improved YOLOv5s framework is presented in Fig. 1.

III. IMPROVEMENT OF YOLOV5

A. FOCAL-SIOU

The efficiency of object detection is highly dependent on the definition of the loss function. The conventional loss function typically focuses on several metrics related to bounding box regression, including the distance, overlapping area, and aspect ratio between the predicted and ground truth boxes, but does not take into account the direction mismatch between the ground truth and predicted boxes. This can cause the predicted box to wander around during the training process, leading to slower model training and less effective convergence, ultimately affecting the detection performance of the model. Focal-SIOU takes into account angle loss and addresses the aforementioned problem. The loss function of Focal-SIOU mainly consists of four parts: angle loss, distance loss, shape loss, and IOU loss. The calculation principle for angle loss and distance loss is illustrated in Fig. 2.

First, let α be the angle less than or equal to $\pi/4$ between the coordinate centers of the predicted box and the ground truth box. C_h and C_w represent the vertical and horizontal distance between the two coordinate centers. D_h and D_w represent the maximum horizontal and vertical distance between the predicted box and the ground truth box. The linear distance σ and angle α between the two coordinate centers can be calculated by the following formulas:

$$\sigma = \sqrt{C_h^2 + C_w^2} \tag{1}$$

$$\alpha = sin^{-1}\frac{C_h}{\sigma} \tag{2}$$

The calculation formula for the angle loss $\Lambda$is as follows:

$$\Lambda = sin(2\alpha) \tag{3}$$

The formulas for the distance loss Δ are as follows, where b^{gt} represents the ground truth box and b represents the predicted box. cx and cy stand for the horizontal and vertical coordinate of the center, respectively. ρ represents the factor for distance loss, while γ represents the factor for angle loss.

$$\Delta = \sum_{t=x,y}\left(1 - e^{-\gamma\rho_t}\right) \tag{4}$$

$$\rho_x = \left(\frac{b_{cx}^{gt} - b_{cx}}{D_w}\right)^2 \tag{5}$$

$$\rho_y = \left(\frac{b_{cy}^{gt} - b_{cy}}{D_h}\right)^2 \tag{6}$$

$$\gamma = 2 - \Lambda \tag{7}$$

It can infer from the above formulas that the distance loss incorporates angle loss. As α approaches 0, the contribution of the distance cost decreases significantly. On the other hand, the closer α is to $\pi/4$, the greater the contribution of the distance cost. It should be noted that the cost of distance will become conventional as α approaches 0.

The formulas for the shape loss Ω are as follows:

$$\Omega = \sum_{t=w,h}\left(1 - e^{-\omega_t}\right)^{\theta} \tag{8}$$

$$\omega_w = \frac{\left|w - w^{gt}\right|}{max\left(w, w^{gt}\right)} \tag{9}$$

$$\omega_h = \frac{\left|h - h^{gt}\right|}{max\left(h, h^{gt}\right)} \tag{10}$$

w and h represent the width and height of the box, respectively. ω represents the factor for shape loss. θ indicates the attention given to the shape loss and needs to be adjusted accordingly based on a specific dataset. In this paper, θ is set to 4.

After integrating the losses of the aforementioned indicators, the SIOU loss function formula is as follows:

$$L_{SIOU} = 1 - IOU + \frac{\Delta + \Omega}{2} \tag{11}$$

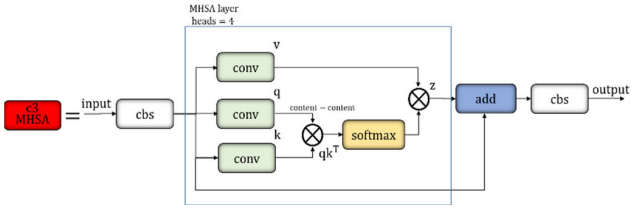

FIGURE 3. C3MHSA module. The q represents query vector, k represents key vector, and v represents value vector. ⊗ represents matrix multiplication. z represents the output of self-attention layer. Although we use 4 heads, we do not show them on the figure for simplicity.

Δ represents distance loss, Ω represents shape loss, and IOU represents the intersection over union ratio between the ground truth box and the predicted box.

When predicting the bounding box regression of the object, the process is affected by the problem of imbalanced training samples. In an image, there are fewer high-quality anchor boxes with small regression errors compared to low-quality anchor boxes with large errors. The poor quality anchor boxes can generate excessive gradients, which can affect the training process negatively. To address this problem, we integrated the Focal loss with SIOU to distinguish between high-quality and low-quality anchor boxes. The Focal-SIOU loss function formula is as follows, where γ represents the Focal factor and is set to 0.5.

$$L_{Focal-SIOU} = IOU^{\gamma} L_{SIOU} \tag{12}$$

B. C3MHSA MODULE

The multi-head self-attention module is a simple yet powerful mechanism that is well-suited for various machine vision tasks [24], [25], including image classification [26], [27], [28], [29], object detection, and visual tracking [30], [31], [32], [33]. In this paper, we replace the third C3 module of the original YOLOv5 with the C3MHSA module, as presented in Fig. 1. Convolution can effectively capture local information, but it lacks the ability to capture long range dependencies. In order to aggregate the locally captured filter responses globally, convolution-based architectures require the stacking of multiple layers. Therefore, utilizing self-attention to model global dependencies can be a more powerful and scalable solution, eliminating the need for as many layers. Self-attention implements pairwise entity interactions with a content-based addressing mechanism, thereby learning a rich hierarchy of associative features across long sequences of data.

The framework of C3MHSA is presented in Fig. 3. The input first generates the query vector q, key vector$k$, and value vector v through point convolution. The query vector q and key vector k are then multiplied to generate the corresponding content-content vector qk^T. This vector passes through a SoftMax layer and is multiplied with the value vector v to obtain the output z. We also attempt to incorporate position encoder within the MHSA layer. However, experimental results indicate that the introduction of the position encoder increases the parameters and negatively impacts the algorithm's precision on our dataset. Compared to the original C3 module, the C3MHSA module reduces the parameters while capturing more long range dependencies. This module successfully overcomes the challenges presented by complex background and significantly improves the accuracy of the model.

C. SHUFFLE ATTENTION MODULE

Attention mechanism has emerged as a crucial component in enhancing model detection performance. There are two widely used types of attention mechanisms in computer vision tasks: spatial attention and channel attention. Both types of attention mechanisms enhance the original features by aggregating the same feature from all positions using different aggregation strategies, transformations, and strengthening functions. Since each channel of a feature map is considered as a feature detector, channel attention focuses on 'what' is meaningful given an input image. In contrast to channel attention, spatial attention focuses on 'where' an informative part is located, which complements the channel attention. Although fusing them together may lead to better performance than their individual implementations, it will inevitably increase the computational overhead and complexity. To address this problem, we introduce the Shuffle Attention (SA) module in the neck of YOLOv5. The SA module can efficiently capture information from both the channel and spatial dimensions with fewer parameters and lower computational cost. The framework of SA module is presented in Fig. 4.

Let the input size is $c \times h \times w$, SA module first splits the input into G groups along the channel dimension, each group size is $c/G \times h \times w$. Then each group is further divided into two branches along the channel dimension. The two branches generate their own feature maps through the spatial attention mechanism and the channel attention mechanim, respectively. Following the extraction of relevant features, the two feature maps are concatenated, and the size changes back to $c/G \times h \times w$. After all G groups have extracted features, they aggregate again, generating an output of the same size as the input. Finally, the output is reordered through channel shuffle to ensure information flow between different groups and enhance the model's representational capability.

In the SA module, the channel attention mechanism performs average pooling to get a set of statistics related to the channels. After a linear transformation and sigmoid activation function, these sets of statistics are multiplied by their corresponding elements of the input to obtain the object feature information. The spatial attention mechanism normalizes the input to get spatially correlated statistics. After a linear transformation and sigmoid activation function, these sets of statistics are multiplied by their corresponding elements of the input to obtain the object position information.

We conducted a series of experiments to compare and analyze the detection performance of five different attention mechanisms. The results are presented in Table 1. Compared to SE [34], CBAM [35], CA [36], and ECA [37], the SA

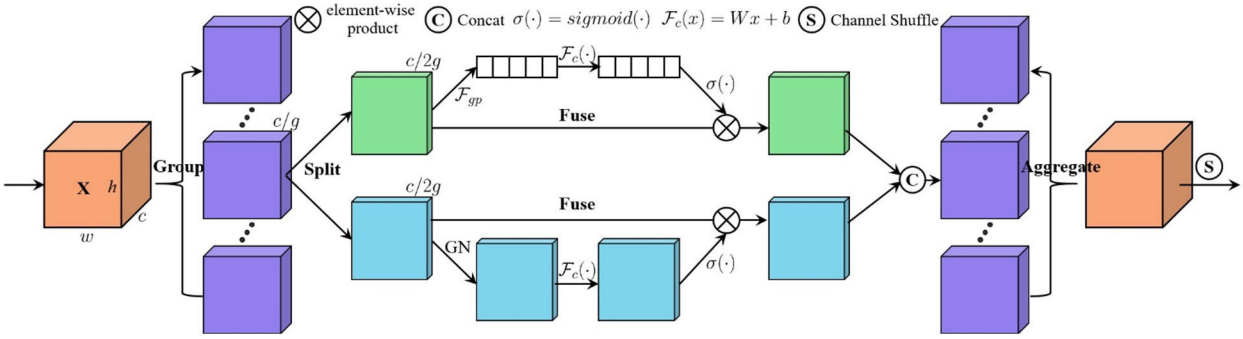

FIGURE 4. Shuffle attention module. The left orange block represents the input feature maps, which are grouped to generate g sets of feature maps of the same size, as represented by the purple blocks. Subsequently, each purple block is divided into two equally sized parts, which are separately subjected to the channel attention mechanism and spatial attention mechanism to facilitate feature extraction, as illustrated by the green and blue blocks. Once feature extraction is complete, the green and blue feature maps are concatenated. After all groups have completed this operation, they aggregate back to the original input size, as represented by the right orange block. Finally, the output is reordered through channel shuffle.

TABLE 1. Performance comparison of five attention mechanisms.

Module	P/%	R/%	mAP/%	Parameters	GFLOPs
YOLOv5s	84.3	86.4	90.4	7.02×10^6	15.8
+SE	85.2	88.8	90.8	7.06×10^6	16.1
+CBAM	84.9	88.7	90.9	7.06×10^6	16.1
+CA	85.5	86.6	91	7.06×10^6	16.1
+ECA	84.3	88.4	90	7.06×10^6	16.1
+SA	86.6	86.5	91.1	7.05×10^6	16.1

module has the highest precision and mAP with slightly lower recall, while having fewer parameters. Based on the experimental results, we introduce a SA module in the neck that enhances and fuses the features extracted by the backbone for subsequent detection, while balancing accuracy and parameter efficiency.

D. IMPROVING NECK STRUCTURE

The neck of the original YOLOv5 model connects three detection heads with sampling ratios of 1/8, 1/16, and 1/32, corresponding to small object detection, medium object detection, and large object detection. However, as the objects studied in this paper are all small and medium sized objects, large object detection head is not suitable. Experimental results demonstrate that the 1/32 detection head not only introduces additional parameters and computational complexity but also reduces the accuracy of the detector. Therefore, this paper has modified the neck of the original YOLOv5 by removing the 1/32 downsampling section and its corresponding detection head, as illustrated in Fig. 1.

IV. EXPERIMENT AND ANALYSIS

The YOLOv5 algorithm offers five different scales of models: N, S, M, L, and X. While the structure of these five scale models remains the same, each scale model possesses a different depth and width, resulting in varying sizes and complexities. In this paper, we examine and analyze the ability of the YOLOv5s model to detect lymphocytes in experiments. The platform used for model training in this experiment is Intel Core i9-13900 CPU and NVIDIA GTX4060 8G GPU. The software uses Windows system, Python 3.11, PyTorch 2.0.1, and Cuda11.8 deep learning framework.

The experiment involves 100 training epochs with a batch size of 5. The input image size is 640 × 640. The initial learning rate is set to 0.01, and SGD is used as the optimization algorithm. The weight decay is 0.005 and the momentum is 0.937.

A. LYMPHOCYTE DISCRIMINATION CRITERIA

In Whole-Slide Images (WSI) of labial gland biopsy, the detection of lymphocytic infiltrative lesions is a diagnostic criterion for Sjogren's syndrome. However, in addition to lymphocytes, other cell types such as epithelial cells and mucus cells may be present in the biopsy sample, which can interfere with the accurate identification of lymphocytes. To distinguish between different cell types at higher magnification accurately, a comprehensive analysis of cell morphological characteristics and staining effects is necessary. This analysis should also take into account background information to determine the cell category. The criteria for lymphocyte discrimination are as follows:

1. Appearance: Lymphocytes are typically medium-sized, single cells with a circular or ovoid shape. In WSI, they present as a nucleus with a small amount of cytoplasm. They usually lack protrusions or have smaller ones.

2. Nuclear characteristics: The nucleus of lymphocytes is typically circular or ovoid with well-defined boundaries. One or more deeply stained nucleoli can be observed within the nucleus, appearing as one or more distinct dots or small masses.

3. Staining properties: The nuclear staining properties of lymphocytes are typically intense, using commonly employed staining agents such as eosin (H&E) staining or Kalman staining. The cytoplasm usually appears lighter in color.

B. EXPERIMENTAL DATASET

The experimental dataset used in this paper is derived from WSI of labial gland biopsy specimens obtained from

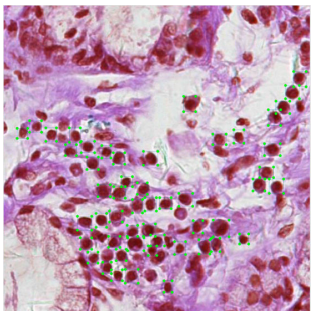

FIGURE 5. Illustration of manually annotated lymphocytes. Green boxes indicate the manually annotated lymphocytes.

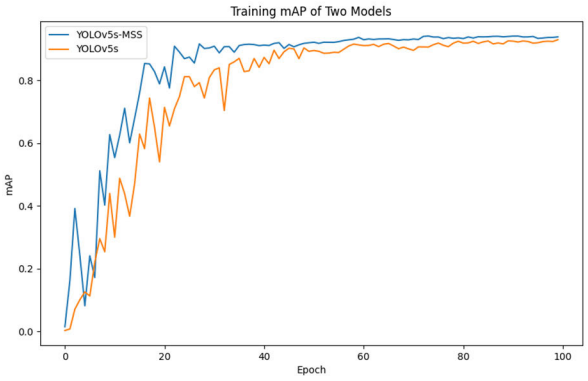

FIGURE 6. Training mAP curves of two models. Orange line represents the original YOLOv5s model, and the blue line represents the YOLOv5s-MSS model.

Yantai Shan Hospital. The use of this dataset has been approved by the hospital's ethics review committee. Due to the significant size of the WSI, it is not feasible to detect lymphocytes directly. Therefore, we segment the WSI into block images of size 640 × 640 at the highest resolution and create a dataset containing 300 images. Based on the criteria for lymphocyte discrimination, the lymphocytes within these images are manually annotated.

It should be noted that due to the average number of lymphocytes contained in a single image being 30, and some images even exceeding 70, the manual labeling process was extremely cumbersome and time-consuming. Our sole detection object is lymphocyte, and there are a significant number of lymphocytes with distinctive and consistent features present in a single image. We chose to manually annotate 300 images and opt for a lighter model for training, in order to achieve a balance between model training and annotation complexity. The dataset is divided into a train set, a validation set, and a test set with a ratio of 8:1:1. An example of an annotated image is presented in Fig. 5, where the green boxes indicate the manually annotated lymphocytes.

C. EVALUATION INDICATORS

We utilized a set of standard metrics to evaluate the performance of the improved YOLOv5 in lymphocyte detection tasks. The primary metrics considered in this paper are Recall (R), Precision (P), and mean Average Precision (mAP). Since the sole object to be detected in this study is lymphocyte, these metrics can be represented as follows:

$$mAP = \int_0^1 P(R)\, dR \tag{13}$$

$$P = \frac{TP}{TP + FP} \tag{14}$$

$$R = \frac{TP}{TP + FN} \tag{15}$$

Among these metrics, TP (true positive) refers to instances that are correctly predicted as positive, TN (true negative) refers to instances that are correctly predicted as negative, FP (false positive) refers instances that are incorrectly predicted as positive, and FN (false negative) refers instances that are incorrectly predicted as negative.

D. ABLATION EXPERIMENT

The YOLOv5s-MSS proposed in this paper has introduced several improvements to the original algorithm's loss function and network structure. To evaluate the effectiveness of different modules and their combinations, we conducted ablation experiments on our dataset. The experimental results are presented in Table 2.

As shown in Table 2, after replacing the original C3 module with the C3MHSA module in the backbone, the model's precision improved by 2.1%, while the recall remained relatively stable. The mAP, parameters, and GFLOPs slightly reduced. After introducing the Shuffle Attention module in the neck, the model's precision increased by 2.3%, while the recall remained relatively stable. The mAP increased by 0.7%, and the parameters and GFLOPs slightly increased. After modifying the neck structure, the model's precision improved by 0.4%, the recall increased by 4.3%, the mAP rose by 2.1%, and the parameters and GFLOPs significantly reduced. After replacing the original CIOU with Focal-SIOU, the model precision improved by 2.1%, the recall increased by 2.4% and the mAP rose by 0.2%. The experimental results demonstrate that the implementation of improvement measures in this paper resulted in variable enhancements in the detection performance of the original YOLOv5s algorithm.

After merging the improvement measures, the final precision of the model reached 87.2%, the recall reached 89%, and the mAP achieved 93.2%. In comparison to the original YOLOv5s, the precision improved by 2.9%, recall increased by 2.6%, mAP rose by 2.8%, and the parameters decreased by 28.2%, while GFLOPs decreased by 13.3%. The experimental results demonstrate that the improvement measures employed in this paper have a significant positive effect on enhancing lymphocyte detection performance.

The training mAP curves of the two models are presented in Fig. 6, where the orange line represents the original YOLOv5s model, and the blue line represents the YOLOv5s-MSS model. As is evident from the figure, the training mAP of YOLOv5s-MSS consistently surpasses that of YOLOv5s, and its convergence speed is also faster.

TABLE 2. Results of ablation experiment.

C3MHSA	SA	New-Neck	Focal-SIOU	P/%	R/%	mAP/%	parameters	GFLOPs
×	×	×	×	84.3	86.4	90.4	7.02×10^6	15.8
√	×	×	×	86.4	86.5	90.1	6.79×10^6	14.9
×	√	×	×	86.6	86.5	91.1	7.05×10^6	16.1
×	×	√	×	84.7	90.7	92.5	5.23×10^6	14.3
×	×	×	√	86.1	88.8	90.6	7.02×10^6	15.8
√	√	×	×	87	86.4	90.5	6.87×10^6	15.0
×	√	√	×	85.4	88.8	92.5	5.27×10^6	14.4
√	×	√	×	86.5	87.8	92.4	5.01×10^6	13.6
√	√	√	×	86.8	88.6	92.6	5.04×10^6	13.7
√	√	√	√	87.2	89	93.2	5.04×10^6	13.7

TABLE 3. CNN based algorithm comparison results.

Algorithm	P/%	R/%	mAP/%	Parameters	GFLOPs
YOLOv3-spp	81.2	87.3	89.9	4.12×10^6	12.0
YOLOv6n	77.9	83.8	88.3	4.23×10^6	11.8
YOLOv7-tiny	74.8	82.9	85.2	6.01×10^6	13.0
YOLOv7	78.3	80.6	85.5	9.32×10^6	26.7
RetinaNet	63.9	82.3	69.6	19.8×10^6	61.5
YOLOv8n	85.5	85.1	91.3	3.01×10^6	8.1
YOLOv8s	84.4	88.4	91.3	11.13×10^6	28.4
YOLOv5s-MSS	87.2	89	93.2	5.04×10^6	13.7

TABLE 4. Transformer based algorithm comparison results.

Algorithm	Epochs	TrainingTime/minute	mAP/%	Parameters	GFLOPs
DETR	300	170	70.1	41×10^6	86
RT-DETR	100	33	88.3	20×10^6	60
YOLOv5s-MSS	100	6	93.2	5.04×10^6	13.7

The detection performance of YOLOv5s-MSS is presented in Fig. 7, where the boxes represent the lymphocytes detected by the model. As is evident from the figure, YOLOv5s-MSS generates accurate and non-overlapping prediction boxes. In contrast, the prediction boxes produced by YOLOv5s overlap significantly, leading to some detection errors. The detection results demonstrate that YOLOv5s-MSS adheres strictly to the criteria for lymphocyte discrimination. It effectively extracts background information while efficiently recognizing interfering cells with similar color and shape, such as epithelial cells, ensuring accurate detection.

E. MODEL COMPARISON EXPERIMENT

To evaluate the performance of YOLOv5s-MSS in lymphocyte detection, we first compared it with other state-of-the-art and similar size CNN-based object detection models, including YOLOv3, YOLOv6, YOLOv7, RetinaNet, and YOLOv8. As presented in Table 3, YOLOv5s-MSS achieves a balance between model complexity and detection accuracy, and possesses certain advantages compared to other algorithms. As presented in Fig. 7, YOLOv7 incorrectly detects many other types of cells, indicating a lack of full understanding of lymphocyte characteristics and a tendency to be influenced by background interference during detection. Although the detection accuracy of YOLOv8s is relatively high, there are still some cases of overlapping bounding boxes in local regions with dense lymphocytes.

Then we compared our YOLOv5s-MSS with state-of-the-art Transformer-based object detection models, including DETR and RT-DETR, to further evaluate its performance. Table 4 demonstrates that YOLOv5s-MSS boasts a noteworthy accuracy advantage over the other two network models. Compared to DETR, YOLOv5s-MSS has an mAP advantage of 23.1%. When compared to RT-DETR, the mAP advantage is 4.9%. Additionally, YOLOv5s-MSS has fewer network parameters and lower GFLOPs. It should be noted that training DETR and RT-DETR is more challenging than training YOLOv5s-MSS. DETR requires 300 epochs and a total of 170 minutes to achieve a moderate level of mAP on our dataset. The training time is approximately 28 times that of YOLOv5s-MSS. RT-DETR, despite being relatively easy to converge, still has a training time that is approximately 5 times that of YOLOv5s-MSS. Therefore, in terms of training complexity, YOLOv5s-MSS, which is based on the CNN architecture, is simpler to train than DETR and RT-DETR, which are based on the Transformer architecture. As presented in Fig. 7, DETR's detection has a significant amount of overlapping boxes and erroneous detections, indicating a lack of sufficient feature learning of lymphocytes and a tendency to be influenced by background interference. Although

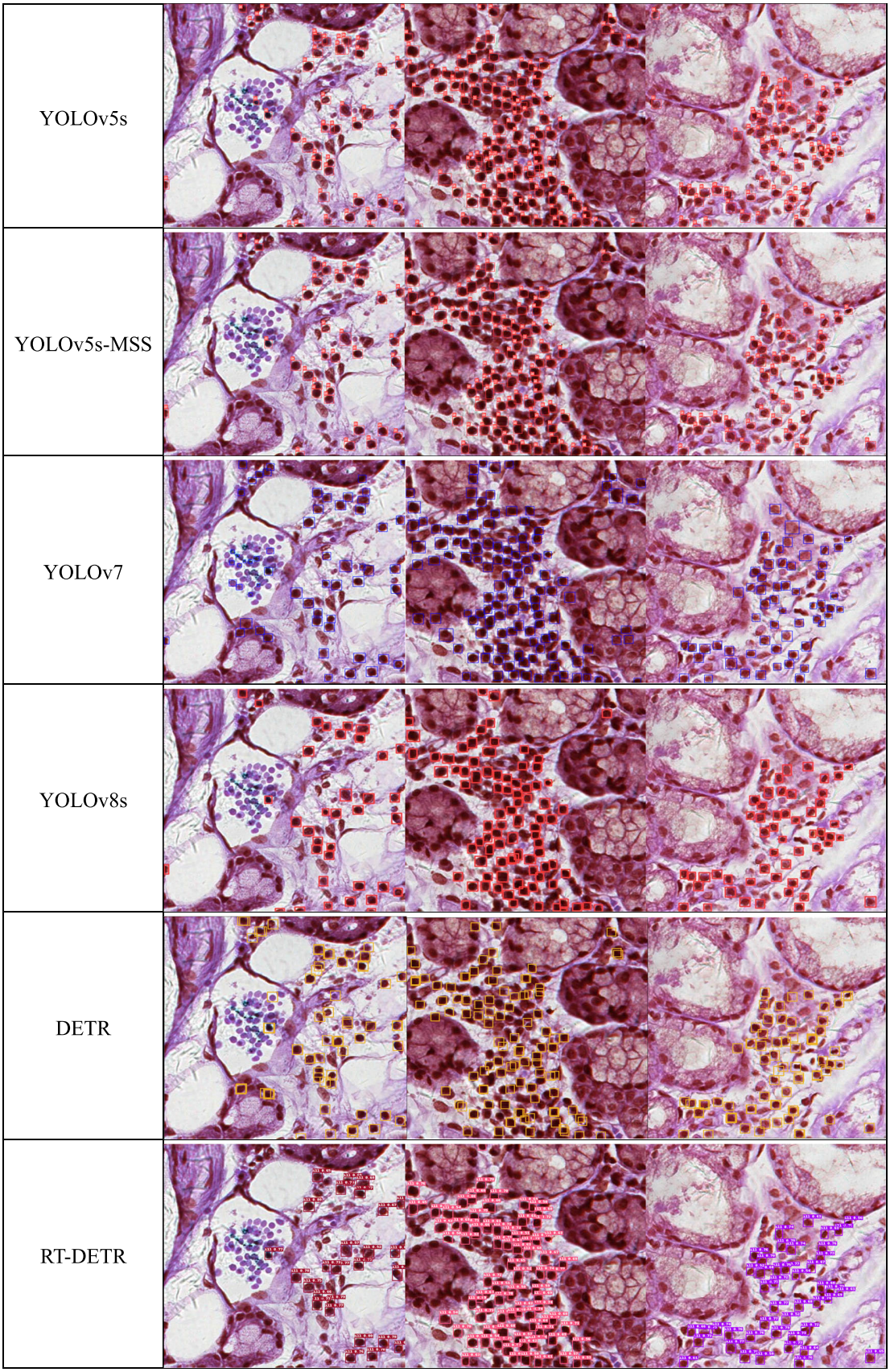

FIGURE 7. Comparison of different algorithm detection effects.

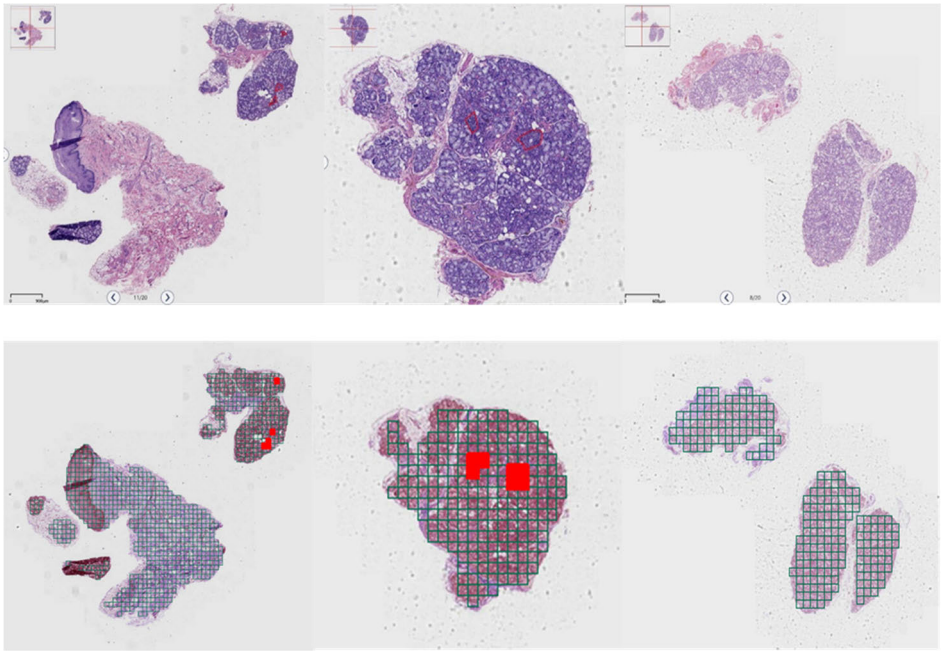

FIGURE 8. YOLOv5s-MSS assists pathological diagnosis. Top three images represent the lesions marked by the physician, and the bottom three images represent the lesions identified by YOLOv5s-MSS.

RT-DETR's detection accuracy is relatively higher, there are still some cases of missed detections in the edge area.

F. AUXILIARY PATHOLOGICAL DIAGNOSIS

We use the YOLOv5s-MSS algorithm to perform block detection on labial gland biopsy WSI and count the number of lymphocytes in each block. Blocks with high lymphocyte counts are labeled in red as suspicious lesions for further physician evaluation in diagnosing Sjogren's syndrome. As presented in Fig. 8, the top three images represent the lesions marked by the physician, while the bottom three images illustrate the lesions identified by YOLOv5s-MSS. It is evident that the lymphocyte detection model, based on improved YOLOv5s, effectively identifies and labels suspicious lesion areas, which roughly correspond to those annotated by the physician. In lesion-free areas, there are no instances of erroneous labeling, indicating that the model boasts high accuracy in lesion discernment and can assist in pathological diagnosis to a certain extent.

V. CONCLUSION

This paper presents the YOLOv5s-MSS algorithm, which is suitable for assisting in the diagnosis of Sjogren's syndrome. Four improvement measures have been proposed for lymphocyte detection. Firstly, we replace the original CIOU loss function with the Focal-SIOU loss function to accelerate the model training and improve the detection accuracy. Additionally, we introduce a multi-head self-attention module into the backbone to enhance the model's ability to capture long range dependencies and overcome the challenges posed by complex background. Furthermore, we introduce the Shuffle Attention module into the neck, which enhances the model's ability to fuse features from both spatial and channel dimensions. Finally, we remove the 1/32 downsampling section in the neck and the corresponding large object detection head, which not only improves accuracy but also reduces the parameters and model complexity. The experimental results demonstrate that YOLOv5s-MSS achieves mAP, Precision, and Recall of 93.2%, 87.2%, and 89%, respectively, representing increases of 2.9%, 2.6%, and 2.8% over the original YOLOv5s model. Additionally, it reduces the parameters by 28.2%, thus proving the efficacy of the proposed improvement measures. We also explore the application of YOLOv5s-MSS in assisting pathological diagnosis, and the results indicate that it is effective and valuable for pathological diagnosis.

REFERENCES

[1] F. B. Vivino, "Sjogren's syndrome: Clinical aspects," *Clin. Immunol.*, vol. 182, pp. 48–54, Sep. 2017.
[2] C. P. Mavragani and H. M. Moutsopoulos, "The geoepidemiology of Sjögren's syndrome," *Autoimmun. Rev.*, vol. 9, no. 5, pp. A305–A310, Mar. 2010.
[3] R. I. Fox, F. V. Howell, R. C. Bone, and P. E. Michelson, "Primary sjogren syndrome: Clinical and immunopathologic features," *Seminars Arthritis Rheumatism*, vol. 14, no. 2, pp. 77–105, Nov. 1984.
[4] H. M. Moutsopoulos, T. M. Chused, and D. L. Mann, "Sjögren's syndrome (Sicca syndrome): Current issues," *Ann. Intern. Med.*, vol. 92, pp. 212–226, Feb. 1980.
[5] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proc. IEEE*, vol. 111, no. 3, pp. 257–276, Mar. 2023.
[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[7] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[9] 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)*, Jun. 2016, pp. 779–788.

[10] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.

[11] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[12] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[13] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.

[14] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Nov. 2020, pp. 213–229.

[17] W. Lv, Y. Zhao, S. Xu, J. Wei, G. Wang, C. Cui, Y. Du, Q. Dang, and Y. Liu, "DETRs beat YOLOs on real-time object detection," 2023, *arXiv:2304.08069*.

[18] W. Yi, S. Ma, and R. Li, "Insulator and defect detection model based on improved YOLO-S," *IEEE Access*, vol. 11, pp. 93215–93226, 2023.

[19] Z. Gevorgyan, "SIoU loss: More powerful learning for bounding box regression," 2022, *arXiv:2205.12740*.

[20] H. Xu, B. Li, and F. Zhong, "Light-YOLOv5: A lightweight algorithm for improved YOLOv5 in complex fire scenarios," *Appl. Sci.*, vol. 12, no. 23, Dec. 2022, Art. no. 12312.

[21] 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)*, Oct. 2021, pp. 2778–2788.

[22] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and efficient IOU loss for accurate bounding box regression," *Neurocomputing*, vol. 506, pp. 146–157, Sep. 2022.

[23] Q.-L. Zhang and Y.-B. Yang, "SA-Net: Shuffle attention for deep convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 2235–2239.

[24] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[25] A. Vaswani, N. Shazeer, N. Parmar, and J. Uszkoreit, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2017, vol. 30.

[26] 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*.

[27] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16514–16524.

[28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[29] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin transformer v2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11999–12009.

[30] F. Gu, J. Lu, C. Cai, Q. Zhu, and Z. Ju, "EANTrack: An efficient attention network for visual tracking," *IEEE Trans. Autom. Sci. Eng.*, to be published. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10271268, doi: 10.1109/TASE.2023.3319676.

[31] D. Yuan, X. Shu, Q. Liu, and Z. He, "Aligned spatial–temporal memory network for thermal infrared target tracking," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 70, no. 3, pp. 1224–1228, Mar. 2023.

[32] F. Gu, J. Lu, and C. Cai, "RPformer: A robust parallel transformer for visual tracking in complex scenes," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.

[33] F. Gu, J. Lu, C. Cai, Q. Zhu, and Z. Ju, "Repformer: A robust shared-encoder dual-pipeline transformer for visual tracking," *Neural Comput. Appl.*, vol. 35, no. 28, pp. 20581–20603, Jul. 2023.

[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[35] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[36] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13708–13717.

[37] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

PEIHE JIANG received the B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology, China, in 2011, 2013, and 2018, respectively. He is currently an Associate Professor and a Master Tutor with the School of Physics and Electronic Information, Yantai University, Yantai, China. His research interests include image processing, embedded software and hardware design, signal processing technology, robotics, and autonomous navigation technology.

YI LI received the B.S. degree from Qingdao Agricultural University, Qingdao, China, in 2021. He is currently pursuing the M.S. degree in electronic science and technology with Yantai University, China. His research interests include image processing and deep learning.

YING LIU received the M.S. degree from Weifang Medical University. She is currently with the Pathology Department, Yantaishan Hospital, as an Associate Chief Physician. She is also a member of the Lymphoma Study Group and the Urinary Male Genital Diseases Study Group, Pathology Branch of the Shandong Medical Association.

NING LU received the M.S. degree from the Shandong University School of Medicine. She is currently with the Pathology Department, Yantaishan Hospital, as an Attending Physician. She is also a member of the Youth Committee of Tumor Pathology Experts, Shandong Clinical Oncology Pathology Society. She is also a member of the Molecular Pathology Group, Cancer Pathology Branch of the Shandong Anti-Cancer Association.