IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 May 2023, accepted 13 June 2023, date of publication 10 July 2023, date of current version 19 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3293864

RESEARCH ARTICLE

An Improved YOLOv5 Algorithm for Wood Defect Detection Based on Attention

SIYU HAN, XIANGTAO JIANG, AND ZHENYU WU
College of Computer and Information Engineering, Central South University of Forestry and Technology, Changsha, Hunan 410000, China

Corresponding author: Xiangtao Jiang (xtjiang@csuft.edu.cn)

This work was supported in part by the Science and Technology Innovation Program of Hunan Province under Grant 2023JJ50058, and in part by the National Key Research and Development Program of China under Grant 2022YFD2200505 and Grant 2018YFB0703900.

ABSTRACT Wood defect detection is a research hotspot in the field of forestry at present. However, existing studies on wood defect detection mainly focus on detecting a single type of defect or common defects, such as knots, insect pests, and cracks, which cannot meet the processing needs of high-quality wood. Moreover, there are problems, such as low recognition rates of small-target defects and poor recognition integrity of dense defects. To address these issues, we construct a large-scale dataset containing multiple types of wood surface defects through data augmentation techniques. We also introduce the Coordinate Attention module, Transformer Encoder module, and Swin Transformer module in the YOLOv5 network structure. The backbone network CSP-Darknet53 is optimized, and BiFPN is introduced in the neck part to achieve multi-scale weighted bidirectional feature fusion. In addition, we implement three new heads: Shead, Mhead, and Lhead in the prediction part. Comparison experiments show that STC-YOLOv5 outperforms some object detection algorithms. Ablation experiments show that each module effectively improves the detection performance. Compared to YOLOv5, STC-YOLOv5 proposed in this paper improve the mAP by 3.1%. All types and scales of wood surface defects are detected better, with great potential for application in the forestry industry.

INDEX TERMS Object detection, transformer, wood defects, YOLOv5.

I. INTRODUCTION

Wood is a natural and renewable resource with a large accumulation in nature. It has the advantages of easy processing, good stability, and a large strength-to-weight ratio. In addition, wood also has unique material properties and excellent environmental characteristics. So it is widely used in life and production. However, there are many defects in wood due to physiology, pathology, and human reasons. Common wood defects include knots, decay, cracks, discoloration, deformation, insect pests, etc. The size, quantity, location, and type of wood defects can affect the quality and appearance of wood products. In addition, since different types of defects need to be treated differently during wood processing, it can also increase the difficulty of processing and reduce work efficiency. Therefore, it is of great significance to automatically classify and detect wood defects.

Initially, wood defect detection mainly relied on manual labor. However, this method is not only inefficient but also easily affected by subjective factors. The accuracy of artificial detecting wood defects is poor and can not meet the quality grading requirements of industrial production. Early studies show that wood defect detection causes 22% of waste materials due to artificial errors. The total output of wood products decreases from 63.5% to 47.4% [1]. With the development of science and technology, researchers began to use rays, ultrasonics, stress waves, and other methods to detect wood defects. However, these methods are difficult to popularize because of the high equipment cost and harsh environmental requirements.

In recent years, with the development of artificial intelligence, wood defect detection has gradually become intelligent. However, wood defect detection methods using deep

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

learning still have some problems. The existing data sets used for wood defect detection are small. Moreover, most of the existing wood defect detection studies are focused on a single type of defect or are limited to common defects such as knots, insect pests, and cracks, which cannot sufficiently cover the types of defects to be processed in industrial production. In addition, many complex, dense, and small-sized defects need to be treated in the actual wood processing. Therefore, although existing studies can achieve high recognition rates for detecting certain defects, they are not practical. In the study of this paper, a dataset collected during the actual production process is chosen [2]. This dataset covers a comprehensive range of wood defects, with a more detailed classification of knots. To better meet the needs of the wood industry production, we have adjusted the data set.

In addition to the limitations of the dataset, wood defect detection also has the problems of low detection accuracy of small targets and low recognition rate of dense targets.To solve these problems, we propose an improved model STC-YOLOv5 based on YOLOv5. The backbone part uses CSP-Darknet53. Several C3 modules in the backbone are removed to optimize the network structure. Based on this, the Transformer Encoder module is introduced to form a new module C3TF, which improves the feature extraction capability of the backbone. In the neck part, we use the weighted bidirectional feature fusion method of BiFPN. This refines the multi-scale fusion of feature maps and feature enhancement of the neck part. The prediction part introduces the Swin Transformer module. Three new detection heads, Shead, Mhead, and Lhead, are proposed to detect small, medium, and large targets, respectively. We add the attention mechanism to the backbone and neck of the new model. Attention can help the network to grasp global information better and coordinate information. In addition to the network structure, we use more practical strategies, including exponential moving average, group convolution, and post-processing. Compared with YOLOv5, our improved STC-YOLOv5 can detect wood defects better. Our contributions are as follows:

(1) In order to optimize the network structure, we improve the backbone of YOLOv5 by removing some C3 modules and adding some new modules called C3TF. The backbone of the new model can better extract wood defect features.

(2) In order to improve the multi-scale feature fusion, we consider that the different input resolutions have different effects on the output. Borrowing from BiFPN, we realize multi-scale weighted bidirectional feature fusion.

(3) In order to solve the problem of poor detection in regions with more dense defects, we introduce the Swin Transformer module. The formed new detection heads include Shead, Mhead, and Lead, which detect small, medium, and large targets, respectively. The proposed detection heads are able to detect complex and dense defects on the wood surface very well.

(4) In order to capture richer information, such as channel information and location information, we add the attention mechanism to the backbone, which enables the new model to identify the target region more accurately.

II. RELATED WORK

To fully utilize forest resources and improve economic efficiency, scholars have begun to use various techniques such as X-ray [3], microwave [4], ultrasonic [5], stress wave [6] and other methods to detect wood defects. However, such physical equipment-based defect detection methods are costly and susceptible to environmental factors. Most importantly, the detection results do not meet industrial quality standards. With the development of computer technology, researchers start to use feature-based computer vision algorithms for wood defect detection, such as histogram [7], support vector machine [8], [9], [10], gray level concurrence matrix [11], etc. Data augmentation is performed by grayscale transformation, spatial domain, and histogram processing of the original image. Then image segmentation and feature extraction are performed. Finally, the wood defects are identified by BP neural network and SVM. Yang and Yu [12] used ultrasound and principal component analysis to extract the defects features of artificial wood holes drilled into 120 elm samples. Qayyum et al. [13] used the PSO algorithm to classify and detect three types of defects: live knots, dead knots, and forked knots, with an accuracy of 78.26%. Kamal et al. [14] used grayscale co-occurrence matrix and texture energy measure to extract texture features as input to the neural network. The results showed that the overall classification accuracies of wood knot defects were 84.3% and 90.5%, respectively. Zhang et al. [15] used principal component analysis and compressive sensing to detect three defects in Quercus wood: live knots, dead knots, and cracks. The experimental results showed that PCA feature fusion could improve the detection speed. The detection accuracy of the SOM neural network was improved from 87% to 92% after compressive sensing. Chang et al. [16] used convex optimization with different weights as a smoothing preprocessing method and used the Otsu segmentation method to obtain the target defect region images to complete the classification and segmentation of four types of defects: pinholes, cracks, live knots, and dead knots. Compared to earlier detection methods, feature-based computer vision detection of wood defects is less costly, more efficient and more accurate.

In recent years, deep learning technology has developed rapidly. More and more scholars have applied deep learning in the field of wood defect detection. The feature vectors of the original images are automatically extracted by convolution and pooling. Finally, wood defects are detected in the fully connected layer. Yang et al. [17] improved the SSD model with ResNet instead of the original VGG to detect live knots, dead knots, decay, mildew, cracks, and pinholes samples of more than 5000 solid wood panels with an average accuracy of 89.7% and an average detection time of 90 ms. Chen et al. [18] used deep learning to detect an edge, corner, joint defect of the wooden panels. The experimental results were 0.97, 0.90 and 0.92 for accuracy, recall and F1 score, respectively. Lim et al. [19] used a lightweight object detection model based on YOLOv4-Tiny to detect four types of wood defects: live knots, dead knots, ring cracks, and pinched bark. The accuracy was improved to 88.32% by modifying the loss function. Tu et al. [20] proposed an improved model GC-YOLOv3 with CIoU loss instead of IoU loss to detect defects in rubber wood and pine wood with the highest mAP of 86.00% and 92.29%, respectively. Compared with shallow neural network models such as BP neural network, RBF neural network, support vector machine, and extreme learning machine, deep neural network models can extract more complex and rich features with higher accuracy. The processing of wood and production of wood products requires intelligent wood defect detection methods. Using deep learning for wood defect detection largely compensates for the lack of manual detection and meets the requirements of industrial production.

Object detection is a research hotspot in the field of computer vision. The task is to locate and classify targets from images. Early object detection algorithms use sliding windows or selective searches to select regions of interest, followed by feature extraction. Finally, detection is performed with classifiers. However, these object detection algorithms, especially the sliding windows selection of the regions of interest, have redundant selection frames, high time complexity, low accuracy, poor real-time performance, and lousy robustness. With the development of artificial intelligence, more and more scholars use deep learning for object detection. By training a large amount of data, the accuracy and performance of object detection can be significantly improved. This approach is a good remedy for the shortcomings of earlier object detection algorithms.

At present, there are two most representative object detection algorithms. One is the regression-based one-stage detection algorithms, such as YOLO series, SSD [21], DETR [22], EfficientDet [23], etc. The other is the two stage detection algorithms based on candidate bounding boxes, such as Fast R-CNN [24], Faster R-CNN [25], Cascade R-CNN [26], Trident-Net [27], etc. Object detection algorithms can also be divided into anchor-based detection algorithms (e.g. ScaledYOLOv4 [28] and YOLOv5 [29]) and anchor-free detection algorithms (e.g. CenterNet [30], YOLOX [31], and RepPoints [32]). Some detection algorithms are used to process UAV images, such as RRNet [33] and PENet [34]. Among these object detection algorithms, the YOLO series algorithms, as classical one-stage detection algorithms, have a faster detection speed compared to the two-stage detection algorithms. YOLOv5 is lightweight with a small pre-trained model, nearly 90% smaller than YOLOv4. YOLOv5 is also easy to deploy. Therefore it is very suitable for industrial application scenarios that require online real time processing. In this paper, an improved YOLOv5 algorithm called STC-YOLOv5 is provided to address the problems in the field of wood defect detection, which can better detect wood surface defects.

III. MATERIALS AND METHODS

A. DATA SET AND PRE-PROCESSING

We use a large-scale image dataset of wood surface defects as the experimental data set. After selection, we retain 3606 images containing wood defects with seven types of defects, including live knots, dead knots, knots with cracks, knots missing, resin, marrow, and cracks. We convert the YOLO format labeled TXT files into VOC format labeled XML files and use the data labeling tool Labeling to correct the converted rectangular label boxes. Finally, we obtain a large-scale image dataset of wood surface defects in VOC format. The various types of wood defects are shown in Figure 1.

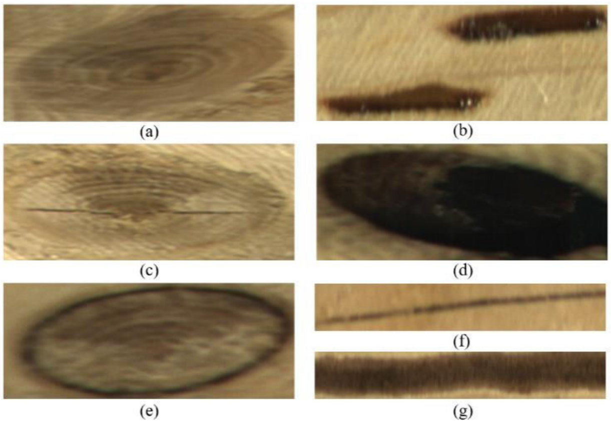


FIGURE 1. Typical samples of wood defects within the dataset: (a) Live knot, (b) Resin, (c) Knot with crack, (d) Knot missing, (e) Dead knot, (f) Crack, and (g) Marrow.

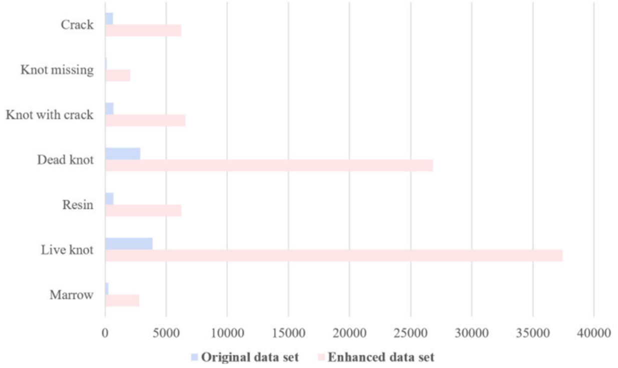


FIGURE 2. The number of labels of each category.

Data augmentation makes the model have better robustness and generalization ability. In addition to several methods such as cropping, mirroring, panning, rotating, masking, adding noise, and changing brightness, we also use Mosaic and Copy-Pasting for data augmentation. Mosaic data augmentation stitches the targets and the labeled boxes in four images to obtain a new image, which greatly enriches the background of targets. Copy-Pasting is a data augmentation method for small targets. Its main idea is to paste a small target to any position in the image to form a new annotation. And the pasted small targets can be scaled, folded, rotated, and other random transformations. This approach improves the contribution of small targets to the loss calculation during training by increasing the number of small targets in each image and anchor boxes that match them. The number of labels in each category of the data set is shown in Figure 2.

B. YOLOv5

YOLOv5 has several different serialized network structures, including YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. In these four network structures, model size and the number of parameters increase sequentially. YOLOv5x has the largest network depth and feature map width. In addition, YOLOv5 has other network structures for detecting larger resolution images, such as YOLOv5n6, YOLOv5s6, YOLOv5m6, YOLOv5l6, and YOLOv5×6. YOLOv5 (6.0) uses CSP-Darknet53 as the backbone, replacing the Focus module of the original network with a 6 × 6 sized convolutional layer. SPP is replaced with SPPF, and CSP is added to PAN to form CSP-PAN. The neck is a combined structure of FPN and CSP-PAN. The prediction part uses the same detection heads as YOLOv3 and YOLOv4, including a small, medium, and large target detection head. The overall structure of YOLOv5 is shown in Figure 3.

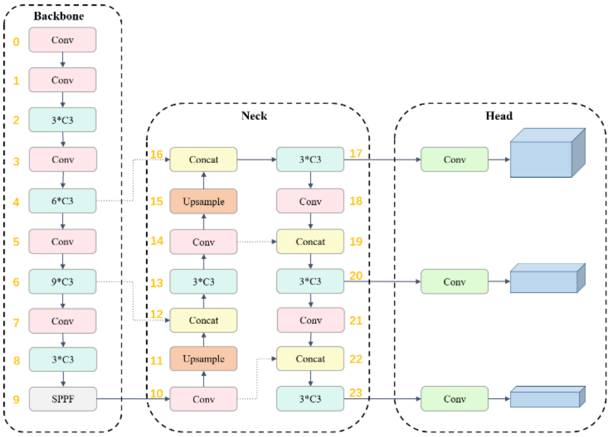


FIGURE 3. The overall structure of the YOLOv5 network.

The input side of YOLOv5 consists of three parts. The first part is the same as YOLOv4, using Mosaic data augmentation. Its main idea is to read four images simultaneously and combine them to form a new image through random cropping, random scaling, and random arrangement. Then calculate the target information of the new image to extend the training set. The second part is the adaptive anchor box calculation. The YOLO series sets initial anchor boxes for the dataset, compares the predicted boxes with the real boxes, calculates the loss, and continuously iterates to update the parameters. YOLOv3 and YOLOv4 set initial anchor boxes separately. But YOLOv5 embeds this function into the code, using adaptive anchor box calculation, that is, automatically selecting the most appropriate anchor box. The third part is adaptive image scaling. Since the datasets have different heights and widths, the input images are scaled to a uniform size for training convenience and then fed to the network. After scaling the image to 640 × 640, YOLOv5 adaptively adds the least black edges to the image, reducing redundant information and improving the inference speed.

C. PROPOSED METHOD

Our proposed STC-YOLOv5 is an improved attention based YOLOv5. First, perform adaptive image scaling, adaptive anchor box calculation, and data enhancement at the input side. Then use a neural network to extract and enhance wood defect features. Finally, complete the detection of wood defects. Compared with the baseline, STC-YOLOv5 can better detect wood defects and identify wood defect types, especially small target wood defects and dense wood defects. It helps to reasonably deal with wood defects and judge the grade of the boards by identifying the size, type, and quantity of defects. The overview of working pipeline using STC-YOLOv5 is shown in Figure 4.

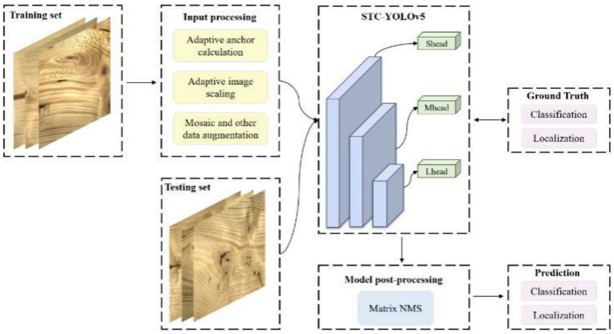


FIGURE 4. The overview of working pipeline using STC-YOLOv5.

YOLOv5 uses a combined structure of FPN and PAN as the neck. The top-down structure of FPN with lateral connections does not consider the problem of different input feature resolutions, that is, feature fusion is insufficient. Therefore, we use the weighted bidirectional feature fusion approach of BiFPN as the neck. This achieves a top-down and bottom-up differentiated bidirectional fusion of deep and shallow features, which enhances the transmission of feature information between different network layers.

STC-YOLOv5 network mainly consists of backbone part, neck part, and prediction part: (1) Backbone part is based on CSP-Darknet53 with some improvements, including Conv, C3, C3TF, CABlock, and SPPF; (2) Neck part uses BiFPN to achieve bidirectional fusion of multi-scale features; (3) Prediction part uses the Swin Transformer encoder module to form new prediction heads SHead, MHead and LHead, which can capture global information and rich contextual information. The overall structure of STC-YOLOv5 is shown in Figure 5.

1) BACKBONE OF STC-YOLOv5 MODEL

The backbone network is responsible for extracting wood defect features. The backbone network of YOLOv5 is

CSP-Darknet53, which mainly consists of C3 (CSP), Conv, and SPPF. The structure of its components is shown in Figure 6. To simplify the network structure while improving the feature extraction capability of the model, we remove some C3 modules and introduce the Transformer encoder module and coordinate attention module (CA module), which will be described in detail in the neck.

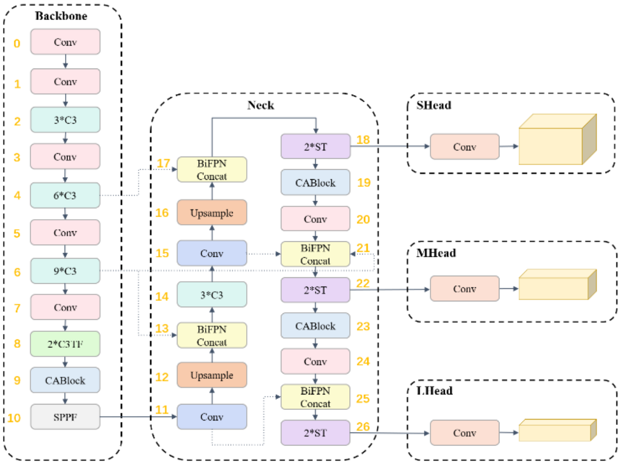


FIGURE 5. The overall structure of the STC-YOLOv5 network.

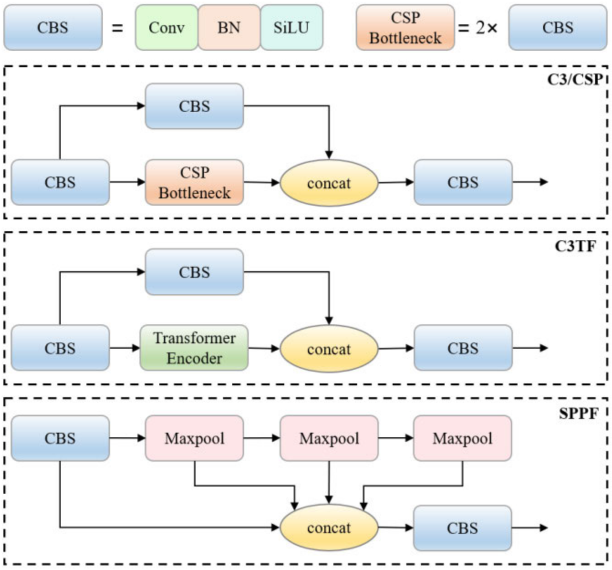


FIGURE 6. C3(CSP)module, C3TF and SPPF in CSP-Darknet53.

Transformer is a deep learning model based entirely on attention mechanisms, consisting of Encoders and Decoders, which are responsible for encoding and decoding, respectively. We only integrate the Encoder module. It consists of multiple identical layers stacked on top of each other, mainly including a multi-head self-attention block and a position-wise feed-forward network. Both sub-layers are stitched together using residual connections. The structure of Transformer Encoder is shown in Figure 7.

Existing research results [35] show that Transformer can extract contextual information well and can also capture global information effectively. To simplify the network model, some C3 modules of the original backbone are removed. A new module C3TF with an integrated Transformer Encoder is added. The backbone of the new model with C3TF can do feature extraction better than the baseline.

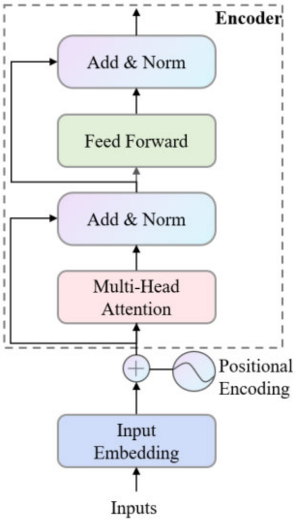


FIGURE 7. Structure diagram of Transformer Encoder.

2) NECK OF STC-YOLOv5 MODEL

To capture richer global information for better performance, coordinate attention mechanisms are introduced in the backbone and neck. The coordinate attention module (CA module) compensates for the disadvantage that other attention modules ignore location information [36]. The input features in vertical and horizontal directions are aggregated into two independent direction-aware feature maps using two one-dimensional global pooling operations, respectively. These two feature maps embedded with specific orientation information are encoded as two attention maps, each capturing the long-range dependence of the input feature maps along a spatial direction. Thus, the location information is stored in the generated attention maps. The working principle of the coordinate attention module is shown in Figure 8.

The CA module can be divided into two steps: coordinate information embedding and coordinate attention generation. Given the input X, two spatial pooling kernels $(H, 1)$ or (1, W) are used to encode each channel along the horizontal and vertical coordinates, respectively. The aggregated feature maps $z_c^h(h)$ and $z_c^w(w)$ are obtained by Equation (1) and Equation (2). They are concatenated and sent to a shared 1×1 convolutional transformation function F_1 to get f. Here r is the reduction rate for controlling block size. Then f is splitted along the spatial dimension into two separate tensors, $\mathrm{f}^h \in \mathrm{R}^{C/r \times H}$, $\mathrm{f}^w \in \mathrm{R}^{C/r \times W}$. Another two 1×1 convolutional transformations, F_h and F_w are used to separately transform f^h and f^w to tensors with the same number of channels to the input X. The obtained outputs g^h and g^w are expanded and used as attention weights, respectively. In summary, the equations for the coordinate attention module are as follows:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i) \tag{1}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w) \tag{2}$$

$$\mathrm{f} = \delta(F_1([z^h, z^w])) \tag{3}$$

$$\mathrm{g}^h = \sigma(F_h(\mathrm{f}^h)) \tag{4}$$

$$\mathrm{g}^w = \sigma(F_h(\mathrm{f}^w)) \tag{5}$$

$$y_c(i, j) = x_c(i, j) \times \acute{A}g_c^h(i) \times \acute{A}g_c^w(j) \tag{6}$$

where $[\cdot, \cdot]$ denotes the concatenation operation along the spatial dimension. δ is the non-linear activation function. $\mathrm{f} \in \mathrm{R}^{C/r \times (H+W)}$ is the intermediate feature map that encodes spatial information in the horizontal and vertical directions.

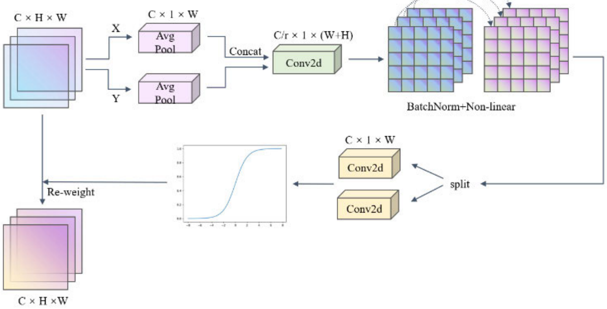


FIGURE 8. Schematic diagram of the coordinate attention module.

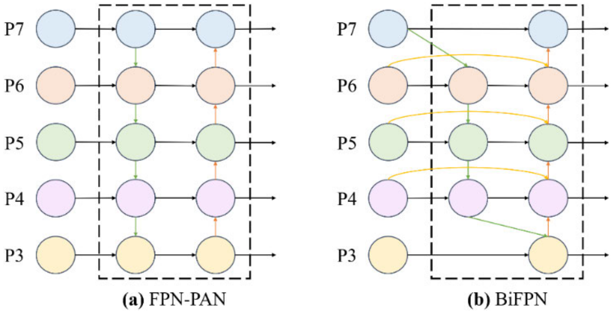


FIGURE 9. Schematic diagram of the neck network.

The neck utilizes multi-scale training. Instead of using the FPN-PAN structure, we borrow BiFPN proposed in EfficientDet, as shown in Figure 9. The neck part combines a weighted bidirectional feature fusion approach, as shown in Figure 5. The 13th, 17th, and 25th layers each realize weighted bidirectional feature fusion of two-layer feature maps with different resolutions. The 21st layer realizes weighted bidirectional feature fusion of three-layer feature maps with different resolutions. The neck is more powerful after adding the attention mechanism and weighted bidirectional feature fusion.

3) HEAD OF STC-YOLOv5 MODEL

The head of our proposed STC-YOLOv5 introduces the Swin Transformer module, which consists of LayerNorm, Multi-head Self-Attention module, MLP, and residual connections, similar to connecting two consecutive Transformers in series. Swin Transformer replaces the standard Multi-head Attention module (MSA) in Transformer with a shift window-based Multi-head Self-Attention module (W-MSA / SW-MSA), using W-MAS for the former and SW-MSA for the latter, with the other layers remaining unchanged, as shown in Figure 10. These two structures are used in pairs, so the number of Swin Transformer blocks is even.

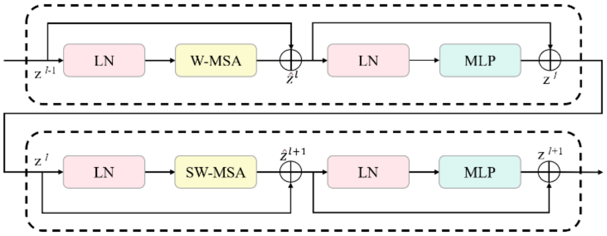


FIGURE 10. Structure diagram of Swin Transformer block.

Swin Transformer abandons the computation of attention from the global window in ViT and proposes to compute attention in the local window, which reduces the amount of computation. Assuming that each window contains M×M patches, the computation complexity of the global MSA module and the window based on h×w patches images can be calculated as follows:

$$\Omega(\mathrm{MSA}) = 4\mathrm{hwC}^2 + 2(\mathrm{hw})^2\mathrm{C} \tag{7}$$

$$\Omega(\mathrm{W-MSA}) = 4\mathrm{hwC}^2 + 2\mathrm{M}^2\mathrm{hwC} \tag{8}$$

where h and w are the height and width of the image, respectively. C is the number of channels. M is the size of the window.

The shift window in Swin Transformer restricts the self attention calculation to non-overlapping local windows while also allowing cross-window connections for window-to-window communication. In this way, the network model not only grasps rich contextual information but also improves detection efficiency to a great extent. The resulting new detection heads Shead, Mhead, and Lhead have good effects on the detection of complex and dense defects on wood surfaces.

D. TRAINING STRATEGIES

1) EMA

The exponential moving average uses exponential decay to calculate the moving average of the training parameters. Preserving the moving average of the parameters is beneficial for training the model. For each variable v, a shadow variable V is obtained, which is calculated as follows:

$$\mathrm{V} = \lambda\mathrm{V} + (1 - \lambda)\mathrm{v} \tag{9}$$

where λ is the decay rate.

2) GROUP CONVOLUTION

Group convolution is used only in the backbone, as shown in Figure 11(b). Group convolution groups the input feature maps and convolution kernels. Then convolution is performed within the group. When the number of groups is one, it is the standard convolution, as shown in Figure 11(a). Compared with the standard convolution, group convolution can reduce parameters and amount of operations.

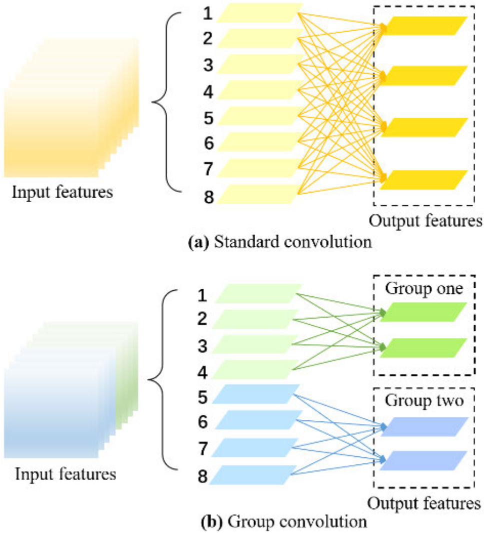


FIGURE 11. Schematic diagram of standard convolution and group convolution.

3) MATRIX NMS

The purpose of NMS is to remove redundant predicted boxes. Soft NMS is easier to realize than traditional NMS because it does not require additional training. Matrix NMS [37] is an improvement on Soft-NMS [38], which makes up for the drawback that Soft-NMS cannot be implemented in parallel. It is faster than the previous NMS because all operations can be implemented simultaneously. In the post processing part, we use Matrix NMS. The decay factor of predicted mask m_j is affected by two aspects. One is the penalty of each prediction m_i on m_j ($s_i > s_j$), where s_i and s_j are the confidence scores. The other is the probability that m_i is suppressed. The calculation formulas are as follows:

$$f(iou_{\cdot,i}) = f(iou_k, i) \tag{10}$$

The final decay factor is:

$$decay_j = \frac{f(iou_{i,j})}{f(iou_{\cdot,i})} \tag{11}$$

Updated score is computed by $s_i = s_j \cdot decay_j$. Consider two decremented functions, linear and Gaussian, with the following equations:

$$f(iou_{i,j}) = 1 - iou_{i,j} \tag{12}$$

$$f(iou_{i,j}) = exp(1 - \frac{iou_{i,j}^2}{\sigma}) \tag{13}$$

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. EXPERIMENTAL ENVIRONMENT AND DETAILS

The experiments in this paper are all implemented on Pytorch 1.10 deep learning framework. The specific parameters are shown in Table 1. We carry out multiple data augmentation methods described in III on the wood defect dataset. All experiments are performed on the large scale image dataset of wood surface defects. Different optimization methods are applied to the weights, bias layer, and BN layer of the model, respectively. The learning rate initially set to 0.01 is updated by the one-dimensional linear interpolation and cosine annealing algorithm. The batch size is set to 16. Finally, the detection of seven types of defects in wood is achieved.

TABLE 1. Software and hardware information.

Name	Parameter Information
CPU	R7-5800H
GPU	NVIDIA GeForce RTX 3060
RAM	16G
OPerating Platform	CUDA 11.1
Framework and Language	Pytorch 1.10&Python 3.6

B. EVALUATION INDICTOR

Several indicators are needed to evaluate the performance of the model after its training is completed. The results can be classified as true positive (TP), false positive (FP), false negative (FN), and true negative (TN) according to the similarities and differences between the predicted and true values. Intersection over Union (IoU) is the ratio of the intersection to union of the predicted boxes and the labeled boxes. AP can be calculated by the area under the P-R curve (with Precision as the vertical axis and Recall as the horizontal axis). mAP is the average AP of all categories. The formulates for Precision, Recall, AP, and mAP are as follows:

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

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

$$AP = \int_0^1 P(r)dr \tag{16}$$

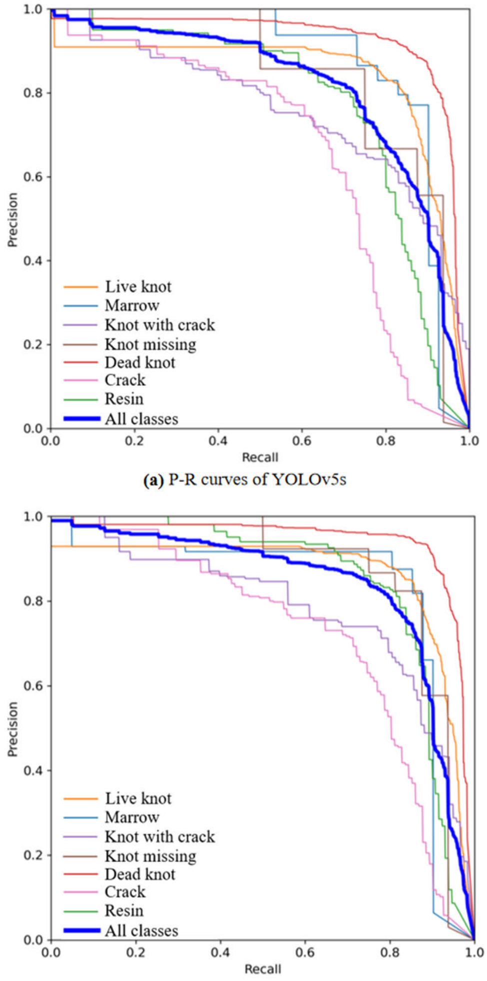
$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{17}$$

C. COMPARISON EXPERIMENTS

To acquire the best research baseline for wood defect detection, we do comparison experiments on several serialized network structures of YOLOv5, as shown in Table 2. By comparing and considering the wood industry production requirements, we select YOLOv5s with the highest mAP and the lowest number of parameters as the study baseline. Based on this, we make improvements to pursue the best performance.

TABLE 2. The performance of different models of YOLOv5.

	mAP_{50}	mAP_{50-95}	Parameters/M	GFLOPS
YOLOv5s	81.1	49.3	7.1	88.5
YOLOv5m	78.8	47.4	20.9	48.0
YOLOv5l	80.4	51.0	46.1	107.9
YOLOv5x	78.5	49.1	86.2	204.1



(a) P-R curves of YOLOv5s

(b) P-R curves of STC-YOLOv5

FIGURE 12. P-R curves of YOLOv5 and STC-YOLOv5.

P-R curves obtained from experiments are shown in Figure 12. It can be seen intuitively that the detection of cracks is the worst and the detection of dead knots defects is the best. Both Table 3 and Figure 12 show that the detection accuracy of STC-YOLOv5 for each type of defects is higher than that of the YOLOv5s, improving the overall performance.

The heat map is shown in Figure 13. We can see that STC-YOLOv5 covers the target object regions better than YOLOv5. That is, adding attention can learn and use the information in the target object regions very well and aggregate features from them.

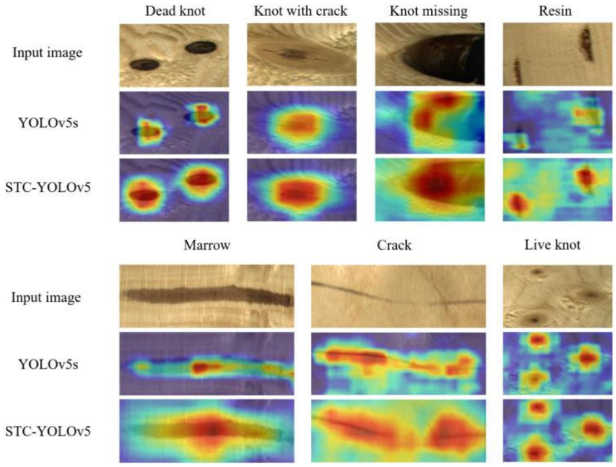


FIGURE 13. Heat map visualization results.

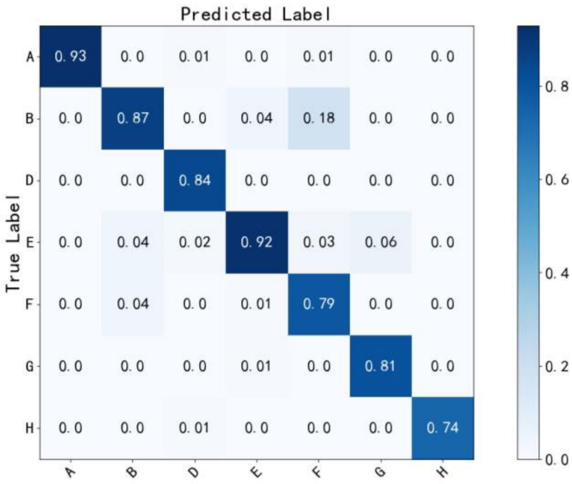


FIGURE 14. Confusion matrix.

D. ABLATION EXPERIMENTS

To evaluate effectiveness and feasibility of the proposed method, we perform ablation experiments to verify the performance of different components and their influence on the detection of different defects. For the accuracy of the ablation experiments, the operating environment and hyperparameters of the model are the same.

Table 3 shows the results of the ablation experiments for different modules. TF denotes Transformer encoder block, ST denotes Swin Transformer module, and CA denotes coordinate attention module. A, B, D, E, F, G, H denote marrow, live knot, resin, dead knot, knot with crack, knot missing, crack, respectively. The results show that the modifications of the model backbone, neck, and head have positive effects on the model. Overall the CA module has the greatest impact, with a 2.6% improvement in mAP. The Swin Transformer module, the Transformer encoder module, and the coordinate attention module perform best for detection of resin, knot

TABLE 3. Ablation studies for different modules.

Algorithms	mAP	A	B	D	E	F	G	H
YOLOv5s	81.1	81.9	85.3	84.8	90.1	74.7	78.6	67.8
+ST	82.5	90.6	84.3	**88.0**	93.7	71.8	78.0	71.3
+TF	83.6	82.8	86.3	85.0	94.0	78.0	**87.1**	72.0
+CA	83.7	**91.1**	86.3	86.4	92.9	78.8	81.1	69.0
STC-YOLOv5	**84.2**	87.9	**86.9**	85.9	**94.8**	**79.0**	82.2	**72.8**

TABLE 4. The performance of different models.

Methods	Backbone	Parameters/M	FPS	mAP(%)
YOLOv3	Darknet-53	61.5	54.3	82.5
YOLOv5	CSP-Darknet53	7.1	**88.5**	81.1
RetinaNet	ResNet-50 FPN	36.5	46.0	80.2
YOLOv7	CSP-Darknet53+E-ELAN+MPConv	37.2	22.7	77.6
SSD	VGGNet	34.3	55.1	79.3
YOLOv8	CSP-Darknet53	11.1	64.3	68.7
Ours	CSP-Darknet53	**6.0**	74.6	**84.2**

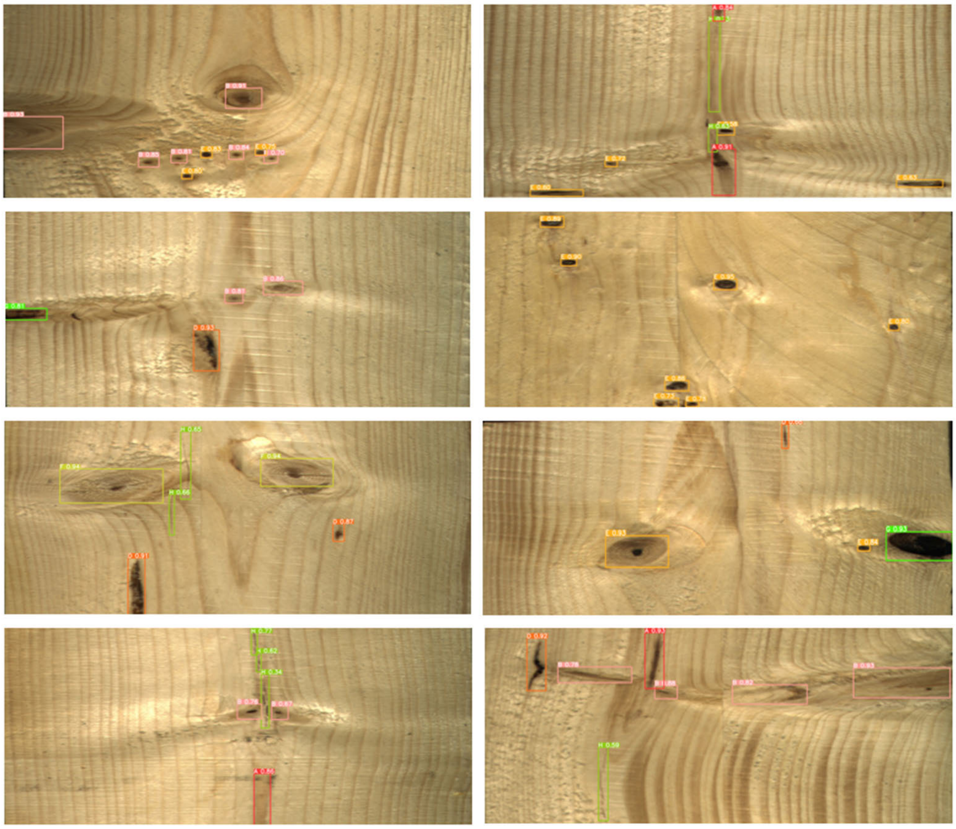
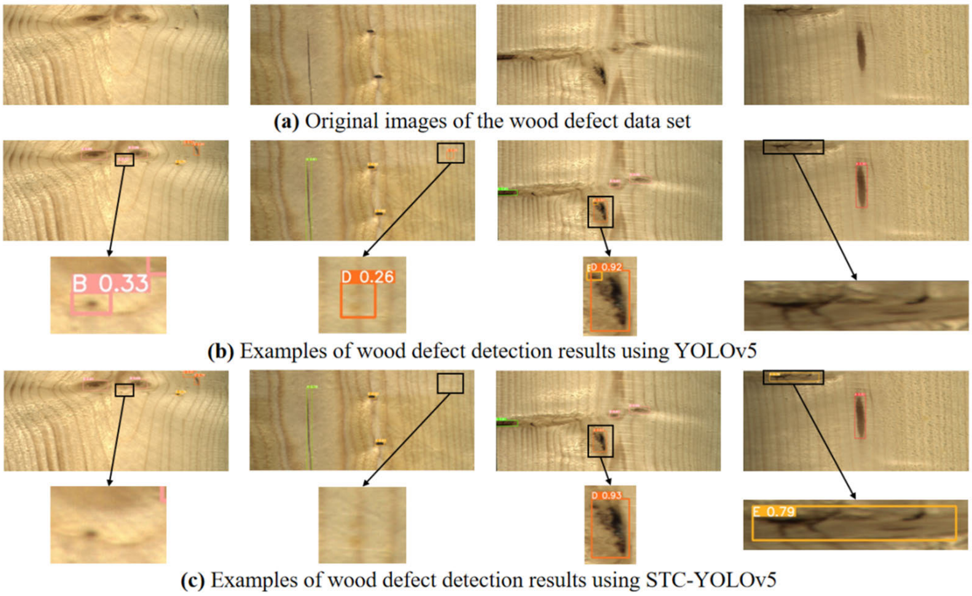


FIGURE 15. Some visualization results from STC-YOLOv5 on the large-scale image dataset of wood surface defects.

missing, and marrow, respectively. However, their combined improvement is more suitable for multi-class wood defect detection, increasing the mAP of the model to an optimal value of 84.2%.

E. DISCUSSION

As shown in Figure 14, a confusion matrix is used to provide accuracy rates reflecting the correct category. The highest prediction accuracy is A(marrow) and the worst is H(crack).



(a) Original images of the wood defect data set

(b) Examples of wood defect detection results using YOLOv5

(c) Examples of wood defect detection results using STC-YOLOv5

FIGURE 16. Examples of wood defect detection results using YOLOv5 and STC-YOLOv5.

The visualization results of our proposed method on the test set are shown in Figure 15. The results in Figure 15 show that our proposed method effectively detects all types of wood defects and is able to accurately identify and locate small-target and dense-target wood defects.

To demonstrate the performance of the proposed method, we compare the proposed method with several mainstream object detection algorithms in the same experimental setting, as shown in Table 4. Our STC-YOLOv5 has a great advantage with the least parameters and the highest mAP. Although the recent YOLOv7 and YOLOv8 algorithms perform well on public datasets, they are not suitable for the dataset of wood surface defects.

Figure 16 shows the different detection results of the wood surface defects data set using YOLOv5 and STC-YOLOv5. It can be seen that YOLOv5 has the problem of wrong detection and missed detection due to insufficient feature learning and confusion between some defects and wood grain. All experiments demonstrate the effectiveness of the improvements to YOLOv5 in this paper.

Although the overall detection accuracy is improved, it does not perform well for cracks and knots with cracks. This is due to the fact that some fine cracks on the wood surface are very shallow and even resemble the marks at the splices of the boards. Therefore our proposed method needs further optimization subsequently. In the future, we will consider using model compression to lighten YOLO series networks, reduce memory overhead and model file size to balance performance and inference speed. These efforts benefit intelligent wood industrial production and are worthy directions for research.

V. CONCLUSION

To develop smart forestry, a wood defect detection algorithm based on improved YOLOv5 is proposed. Our proposed algorithm adds Transformer Encoder block, coordinate attention module, Swin Transformer, and BiFPN to YOLOv5 to enhance the feature extraction capability and multi-scale feature fusion capability of the model. Experimental results show that STC-YOLOv5 can improve the detection accuracy of multiple categories of wood defects, which has certain advantages compared with other algorithms. And it effectively solves the problems of low detection rate of small-sized defects and incomplete recognition of complex and dense defects on wood surface. Therefore, our proposed algorithm provides a good reference for wood defect detection in forestry.

ACKNOWLEDGMENT

The authors would like to thank the editors and the reviewers for their valuable suggestions.

REFERENCES

[1] U. Buehlmann and R. E. Thomas, "Impact of human error on lumber yield in rough mills," *Robot. Comput.-Integr. Manuf.*, vol. 18, nos. 3–4, pp. 197–203, Jun/Aug. 2002.

[2] P. Kodytek, A. Bodzas, and P. Bilik, "A large-scale image dataset of wood surface defects for automated vision-based quality control processes," *F1000Research*, vol. 10, p. 581, Jun. 2021.

[3] H. Mu, D. Qi, M. Zhang, and P. Zhang, "Study of wood defects detection based on image processing," in *Proc. 7th Int. Conf. Fuzzy Syst. Knowl. Discovery*, Yantai, China, 2010, pp. 607–611, doi: 10.1109/FSKD.2010.5569454.

[4] M. Pastorino, A. Salvade, R. Monleone, T. Bartesaghi, G. Bozza, and A. Randazzo, "Detection of defects in wood slabs by using a microwave imaging technique," in *Proc. IEEE Instrum. Meas. Technol. Conf. (IMTC)*, Warsaw, Poland, May 2007, pp. 1–6, doi: 10.1109/IMTC.2007.379332.

[5] W. Lin and J. Wu, "Non-destructive testing of wood defects for Korean pine in northeast China based on ultrasonic technology," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, KunMing, China, Aug. 2013, pp. 1–4, doi: 10.1109/ICSPCC.2013.6664057.

[6] G. Li, X. Weng, X. Du, X. Wang, and H. Feng, "Stress wave velocity patterns in the longitudinal-radial plane of trees for defect diagnosis," *Comput. Electron. Agricult.*, vol. 124, pp. 23–28, Jun. 2016.

[7] W. Song, T. Chen, Z. Gu, W. Gai, W. Huang, and B. Wang, ''Wood materials defects detection using image block percentile color histogram and eigenvector texture feature,'' in *Proc. 1st Int. Conf. Inf. Sci., Machinery, Mater. Energy*, 2015, pp. 779–783.

[8] I. Y.-H. Gu, H. Andersson, and R. Vicen, ''Wood defect classification based on image analysis and support vector machines,'' *Wood Sci. Technol.*, vol. 44, no. 5, pp. 693–704, Apr. 2010.

[9] Q. Ye, H. Zhao, Z. Li, X. Yang, S. Gao, T. Yin, and N. Ye, ''L1-norm distance minimization-based fast robust twin support vector *k*-plane clustering,'' *IEEE Trans. neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 4494–4503. Sep. 2018, doi: 10.1109/TNNLS.2017.2749428.

[10] H. Yan, Q. Ye, T. Zhang, D.-J. Yu, X. Yuan, Y. Xu, and L. Fu, ''Least squares twin bounded support vector machines based on L1-norm distance metric for classification,'' *Pattern Recognit.*, vol. 74, pp. 434–447, Feb. 2018.

[11] X. YongHua and W. Jin-Cong, ''Study on the identification of the wood surface defects based on texture features,'' *Optik Int. J. Light Electron Opt.*, vol. 126, no. 19, pp. 2231–2235, Oct. 2015.

[12] H. Yang and L. Yu, ''Feature extraction of wood-hole defects using wavelet-based ultrasonic testing,'' *J. Forestry Res.*, vol. 28, no. 2, pp. 395–402, Mar. 2017.

[13] R. Qayyum, K. Kamal, T. Zafar, and S. Mathavan, ''Wood defects classification using GLCM based features and PSO trained neural network,'' in *Proc. 22nd Int. Conf. Automat. Comput. (ICAC)*, Colchester, U.K., 2016, pp. 273–277, doi: 10.1109/IConAC.2016.7604931.

[14] K. Kamal, R. Qayyum, S. Mathavan, and T. Zafar, ''Wood defects classification using laws texture energy measures and supervised learning approach,'' *Adv. Eng. Informat.*, vol. 34, pp. 125–135, Oct. 2017.

[15] Y. Zhang, C. Xu, C. Li, H. Yu, and J. Cao, ''Wood defect detection method with PCA feature fusion and compressed sensing,'' *J. Forestry Res.*, vol. 26, no. 3, pp. 745–751, Apr. 2015.

[16] Z. Chang, J. Cao, and Y. Zhang, ''A novel image segmentation approach for wood plate surface defect classification through convex optimization,'' *J. Forestry Res.*, vol. 29, no. 6, pp. 1789–1795, Jan. 2018.

[17] Y. Yang, H. Wang, D. Jiang, and Z. Hu, ''Surface detection of solid wood defects based on SSD improved with ResNet,'' *Forests*, vol. 12, no. 10, p. 1419, Oct. 2021.

[18] L.-C. Chen, M. S. Pardeshi, W.-T. Lo, R.-K. Sheu, K.-C. Pai, C.-Y. Chen, P.-Y. Tsai, and Y.-T. Tsai, ''Edge-glued wooden panel defect detection using deep learning,'' *Wood Sci. Technol.*, vol. 56, no. 2, pp. 477–507, Jan. 2022.

[19] W.-H. Lim, M. B. Bonab, and K. H. Chua, ''An optimized lightweight model for real-time wood defects detection based on YOLOv4-tiny,'' in *Proc. IEEE Int. Conf. Autom. Control Intell. Syst. (ICACIS)*, Shah Alam, Malaysia, 2022, pp. 186–191, doi: 10.1109/I2CACIS54679.2022.9815274.

[20] Y. Tu, Z. Ling, S. Guo, and H. Wen, ''An accurate and real-time surface defects detection method for sawn lumber,'' *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021, doi: 10.1109/TIM.2020.3024431.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg, ''SSD: Single shot MultiBox detector,'' in *Proc. 14th Eur. Conf., Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 21–37.

[22] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, ''Deformable DETR: Deformable transformers for end-to-end object detection,'' 2020, *arXiv:2010.04159*.

[23] M. Tan, R. Pang, and Q. V. Le, ''EfficientDet: Scalable and efficient object detection,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, May 2020, pp. 10778–10787, doi: 10.1109/CVPR42600.2020.01079.

[24] R. Girshick, ''Fast R-CNN,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Apr. 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169.

[25] 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, doi: 10.1109/TPAMI.2016.2577031.

[26] Z. Cai and N. Vasconcelos, ''Cascade R-CNN: Delving into high quality object detection,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Sep. 2018, pp. 6154–6162, doi: 10.1109/CVPR.2018.00644.

[27] Y. Li, Y. Chen, N. Wang, and Z.-X. Zhang, ''Scale-aware trident networks for object detection,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 6053–6062, doi: 10.1109/ICCV.2019.00615.

[28] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, ''Scaled-YOLOv4: Scaling cross stage partial network,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 13024–13033, doi: 10.1109/CVPR46437.2021.01283.

[29] G. Jocher, K. Nishimura, T. Mineeva, and R. Vilariño. (2020). *YOLOv5*. Accessed: Jan. 10, 2021. [Online]. Available: https://github.com/ultralytics/yolov5.Jan.2021.

[30] X. Zhou, D. Wang, and P. Krähenbühl, ''Objects as points,'' 2019, *arXiv:1904.07850*.

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

[32] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, ''RepPoints: Point set representation for object detection,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 9656–9665, doi: 10.1109/ICCV.2019.00975.

[33] C. Chen, Y. Zhang, Q. Lv, S. Wei, X. Wang, X. Sun, and J. Dong, ''RRNet: A hybrid detector for object detection in drone-captured images,''in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Seoul, Korea (South), Oct. 2019, pp. 100–108, doi: 10.1109/ICCVW.2019.00018.

[34] Z. Tang, X. Liu, and B. Yang, ''PENet: Object detection using points estimation in high definition aerial images,'' in *Proc. 19th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Miami, FL, USA, Dec. 2020, pp. 392–398, doi: 10.1109/ICMLA51294.2020.00069.

[35] 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, Oct. 2021, pp. 2778–2788, doi: 10.1109/ICCVW54120.2021.00312.

[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)*, Nashville, TN, USA, Jun. 2021, pp. 13708–13717, doi: 10.1109/CVPR46437.2021.01350.

[37] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, ''SOLOv2: Dynamic and fast instance segmentation,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Dec. 2020, pp. 17721–17732.

[38] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, ''Soft-NMS improving object detection with one line of code,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 5562–5570, doi: 10.1109/ICCV.2017.593.

SIYU HAN was born in Hebei, China, in 1999. She is currently pursuing the master's degree in software engineering with the Central South University of Forestry and Technology, Changsha, China. Her current research interest includes object detection.

XIANGTAO JIANG received the Ph.D. degree from Central South University, in 2019. He is currently an Associate Professor with the School of Computer and Information Engineering, Central South University of Forestry and Technology. His current research interests include artificial intelligence, big data analysis, and distributed storage. He is a Professional Member of CCF.

ZHENYU WU was born in Hunan, China, in 1995. He is currently pursuing the master's degree in software engineering with the Central South University of Forestry and Technology, Changsha, China. His current research interest includes object detection.