IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 July 2023, accepted 27 July 2023, date of publication 1 August 2023, date of current version 8 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3300900

APPLIED RESEARCH

Segmentation Method for Whole Vehicle Wood Detection Based on Improved YOLACT Instance Segmentation Model

JISHI ZHENG[1], JUNJIE ZHENG[2], SHIWEN ZHANG[2], HONGHUI YU[2], LINGHUA KONG[3], AND DING ZHIGANG[3]

[1]Intelligent Transportation System Research Center, Fujian University of Technology, Fuzhou 350118, China
[2]School of Transportation, Fujian University of Technology, Fuzhou 350118, China
[3]School of Mechanical and Automotive Engineering, Fujian University of Technology, Fuzhou 350118, China

Corresponding author: Junjie Zheng (958542267@qq.com)

This work was supported in part by the School-Enterprise Cooperation Project of Fujian Jinsen Forestry Company Ltd., under Grant GY-H-20154, in part by the Forestry Technology Project of Fujian Province under Grant 2021FKJO6, and in part by the Nature Foundation of Fujian Science and Technology Department under Grant 2018JO1619.

ABSTRACT In order to overcome the problems of slow detection speed, low detection accuracy, dense wood stacks and easily obscured and overlooked, a segmentation method based on YOLACT_WOOD is proposed. YOLACT algorithm is proposed to explore the feasibility of a single-stage instance segmentation model for fast and accurate segmentation of whole-truck wood. In this study, based on the original YOLACT model, firstly, the ResNeXt network embedded with the CBAM attention mechanism module is used as the backbone network to imporve the feature extraction capability of the model; secondly, the image input size is increased to improve the detection ability of medium and small diameter class wood; then the CIoU bounding box regression loss function is used to improve the accuracy of bounding box regression; finally, DIoU is combined with Fast-NMS as a boundary box screening algorithm to improve the problem of false and missed detections. In this study, the YOLACT_WOOD algorithm is evaluated using five evaluation metrics: *mAP*, *FPS*, IoU_{mask}, wood true detection rate, and parametric size, and the wood segmentation mask map is fitted and counted using the OpenCV library. The experimental results show that the *mAP* of this study method is improved by 5.6% compared to the original network, the IoU_{mask} is improved by 2.6%, the *FPS* is improved by 14.7 frames/sec compared to the detection speed of the Mask R-CNN model, and the true detection rate of the logs in the test set reaches 96.61%, the false detection rate is 0.23%, and the parametric number of the model is not significantly improved. This result shows that the YOLACT_WOOD model not only ensures the segmentation speed but also improves the segmentation accuracy, solves the problem of false and omission, and the algorithm has strong robustness and generalisation ability.

INDEX TERMS Attention mechanism, CIoU loss function, DIoU, ResNeXt, wood detection segmentation, YOLACT.

I. INTRODUCTION

As China pledged to meet the targets for carbon neutrality by 2060, concerns were expressed by various countries and positive reactions were given by all parts of the world [1]; the Food and Agriculture Organization of the United Nations published reports mentioning that, with forests a major part of the global carbon cycle [2], trees have an important role in securing food, drinking water, renewable energy and the rural economy and that the rational utilization of forests has positive consequences for the development of the economy [3]. As the Chinese government encourages the manufacturing industry to speed up its transformation and upgrading, all aspects of forest production are gradually moving towards

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

FIGURE 1. Manual scale check.

information technology and intelligence. Several links in the production process, such as timber harvesting, transportation, and fluting management, need to determine the debarked diameter and timber length of control timber, and then refer to the national standard GB/T 4814-2013 ''Table of log timber volume'' to obtain the timber volume. In traditional logging, manual inspection tape is used to calculate log volume, and manual inspection tape requires a large amount of human labor and is influenced by subjective judgement, poor efficiency and prolonged cycle times, and also the labor cost is high.This labor-intensive process fatigues the inspector and affects measurement accuracy, which cannot be guaranteed to exceed 70-80% [4]. In Figure 1, the diameter is measured by manual inspection tape. As computer technology, especially computer vision technology, develops rapidly, the need to replace manual checks by computer image capturing and end face recognition technology becomes urgent.

II. RELATED WORK

In recent years, with the increasing maturity of computer technology, it has become possible to explore the use of computer vision methods instead of manual inspection. At this stage, the use of computer technology to achieve log end-face inspection and wood volume calculation is divided into three main directions:

A. IMAGE VISION ALGORITHMS

Firstly, traditional image vision algorithms, such as Galsgaard et al. [5] proposed a method for setting the weights of the graph using the information obtained by circular Hough transform (CHT) [6] combined with local circularity measure (LCM) and thus used to estimate the volume of the wood pile, which is sensitive to target distortion and noise, has high computational complexity and is unknown for the size of the a priori target; Kruglov [7] obtained log volumes by modeling logs in 3D space, using an image processing scheme combining clustering algorithm, Stoer-Wagner algorithm and watershed algorithm, which has a relatively impressive measurement result but is overly complex, not robust and has high requirements on the environment; Kruglov and Shishko [8] proposed an improved radially symmetric object detection method of pile volume measurement algorithm, the algorithm by combining Meanshift clustering, Delaunay triangular dissection, Boruvka minimum spanning tree algorithm, watershed and Boykov-Kolmogorov graph cut algorithm, the algorithm has a TPR value of 96.2% and the error is less than 9.2% compared with manual measurement, but the algorithm has a complex shape and texture of log end faces cannot be accurately segmented, sensitive to noise, blur or distortion in log images, and too complex; Kruglov [9] developed an automatic detection method for rounded targets so as to achieve log volume measurement, and the average detection probability of the method for targets was 95.7% after testing, but the method was not accurate enough for segmenting log boundaries, resulting in large errors; Budiman et al. [10] developed a portable handheld device consisting of a fixed-length iron rod, a camera, and a Raspberry Pi. The iron rod was placed on the end face of the log and a camera was used to capture the image, which was then processed on the Raspberry Pi by compression, grayscale conversion, contour analysis, and circumferential fitting to measure the wood diameter.This method has the advantage of portability and ease of use, achieving a measurement error of less than 3%.However, it should be noted that this method requires an LED light source, as operating in natural The robustness is limited when operating under lighting conditions; Guanghua et al. [11] designed an automated measurement system consisting of a CCD camera, a microcontroller and an image processing software for the upper computer using the principle of binocular vision, which is less efficient for single wood inspection; Keck and Schödel [12] used a log scanner as well as an edge projection system to generate a high-resolution grid of the log surface by streak projection. Using computational geometry and coordinate metrology techniques to compensate for subsurface scattering errors, a three-dimensional fitted model of the logs was obtained, and finally the log end-measure diameter was obtained.

B. MACHINE LEARNING

Another is machine learning methods, such as Samdangdech and Phiphobmongkol [13] proposed a method that combines a single-shot multibox detector (SSD) target detection model and a full convolutional network (FCN) semantic segmentation model to achieve on-board eucalyptus picture segmentation, which achieves 94.45% correct wood counting rate, but is not ideal for segmentation of obscured and split wood; Tang et al. [14] proposed a method to detect log endfaces in natural scenes using an SSD model, which uses annotated information of log endstock images to efficiently learn the unique features of log endface regions. By doing so, it mitigates the background interference during target recognition and enhances the learning ability of the model. The proposed method achieves an accuracy of 94.87% and a

recall of 91.34%. Compared with traditional techniques, this strategy greatly reduces the influence of ambient light on log recognition, solves the challenge of log end-face detection and recognition under complex background conditions, and paves the way for intelligent timber volume calculation of logs.

FIGURE 2. ZED binocular camera.

C. DEEP LEARNING

The last one is a deep learning method, such as Cai et al. [15] changed the convolutional layer in the decoding network of YOLOv4-tiny into a deep separable convolution to reduce the number of model parameters, and also implemented an attention mechanism to enhance the features of target logs by introducing compression and excitation networks to achieve log endface detection with a positive detection rate of up to 93.3%, while the weight of the model was reduced by 29.91% compared with the original model; Lin et al. [16] proposed the use of circular arc detection for endface detection and developed an equal-length log volume detection system based on the YOLOv3-tiny target detection model with Hough circle transformation, which can show good detection results for logs stacked in bundles with a true detection rate of 98.79%, but mainly for large-diameter wood; Lin et al. [17] Tiny to make the detection frame fit the log end face better, and improve the model recognition rate by combining soft thresholding with the SE module.

Whether a traditional or machine learning method is used for processing the log images, it needs to tune the parameters by different log images in order to find better outcomes. It is difficult, especially in the case of processing huge sets of complex images. One of the shortcomings of binocular shooting is that the camera calibration parameters vary from one environment to the other, requiring adjustment of the camera calibration parameters; Because most face of logs are close to oval shape rather than square shape, it is not practical to detect log end points by the use of Hoff circles; The log scanner is generally applied to a single, accurate measurement of larger (50 cm or heavier) logs with 3D modeling and cannot be employed in complex environments in which many logs. Although the log end face detection based on deep learning has greatly improved the detection accuracy and segmentation accuracy of the model, there are still inaccurate prediction frames in some areas and the occurrence of missed and false detections of wood. At the same time, in the actual production environment The lower model lacks rapid detection ability, and it is urgent to improve the comprehensive detection efficiency. At present, the research on the detection speed of log end face is relatively scarce. Therefore, a new algorithm is required to improve the detection efficiency of the network. It needs to improve the detection efficiency of the network as well as the detection accuracy of the model. To solve these problems, the paper has developed an experimental study on the whole lorry log, and has proposed a method of segmentation for the detection of whole lorry logs by YOLACT_WOOD. In this paper, a single stage instance segmentation algorithm for YOLACT [18] is presented. The algorithm is based on an enhanced single level instance model, which can effectively improve the decision speed of the model. By use of ResNeXt [19] network in the model, the accuracy of log end face identification can be increased without a significant increase of parameters; In the automatic feature extraction network, the attention-related module CBAM [20], with the aid of channel attention, can extract an important feature information from the primary image by extracting the important feature information from an image and spatial attention, thereby improving the capacity of extracting important feature information of the model; In order to increase the accuracy of the prediction box and reduce the false detection rate, the paper adopts the following steps: To increase the image input by 4 times for small and medium and small wood data, to use CIoU [21] to improve the prediction box accuracy and to introduce DIoU [22] in the Fast-NMS [18], and to reduce the false detection of prediction frame through the use of DIoU. The experimental results prove that this improved algorithm not only improves the detecting segmentation accuracy but also has a significant improvement in the processing speed, and meets the requirements of light-duty model with little parameter increase, which is suitable for the deployment of the model.

III. MATERIALS AND METHODS

A. DATA ACQUISITION

The experimental data used in this study was collected from a lumber yard of Fujian Jinlin Industrial Co., Ltd., using binocular cameras to collect images in different scenarios, as shown in Figure 2, to ensure the diversity of samples in the data set, so that the model can perform well It has strong robustness in complex scenes. A total of 500 wood images were collected in this study, and 150 clear images were finally retained by eliminating debris and blurred images. Some samples of the dataset are shown in Fig. 3. The ZED binocular camera is a depth-sensing camera developed by Stereolabs. It uses binocular vision technology to simulate the human eye to achieve three-dimensional perception and depth perception of the environment. ZED cameras combine the functions of visual imaging and depth sensing and can be used in various computer vision and depth perception applications. ZED camera has high-resolution color image and depth image output, which can be used in graphics processing, computer vision algorithm and 3D reconstruction, and has automatic stereo correction function, which can automatically calculate the internal parameters and external parameters of the

(a) Natural light angle

(b) Foreign Object Obstruction

(c) Pine wood

(d) Backlight angle

FIGURE 3. Sample dataset.

camera, so as to optimize the accuracy of the depth map sex and quality. When using the ZED binocular camera to take pictures of the whole vehicle's timber, it is necessary to ensure that the angle between the camera and the end surface of the timber does not exceed 15°. Usually around 10° has negligible effect on the results.

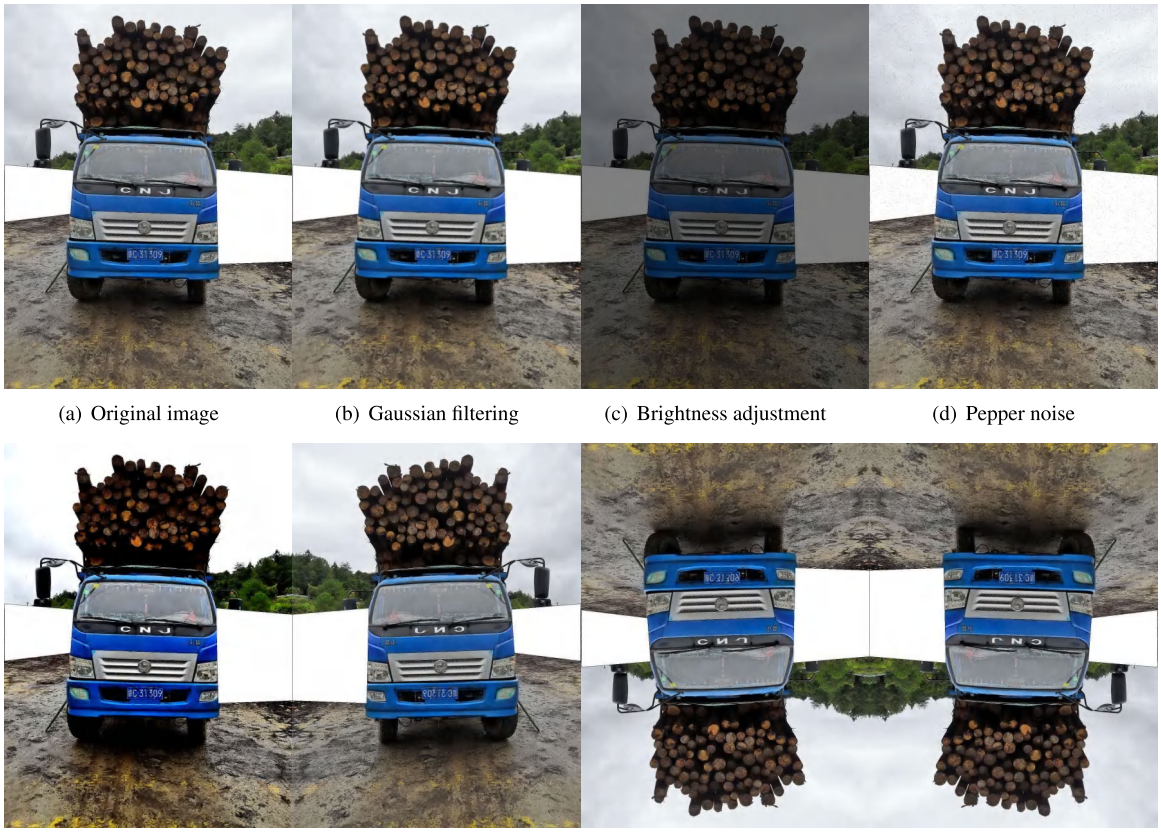
B. SAMPLE AMPLIFICATION

To enrich the experimental dataset, better extract wood endface features and improve the model generalization ability, the data enhancement technique was used to augment the wood endface dataset with samples, and the wood endface images were processed by Gaussian filtering, chromaticity, pretzel noise, contrast and flipping different angles, respectively. The augmented images are shown in Figure 4.

C. DATASET PRODUCTION

YOLACT is a supervised learning model, which needs to label the log contours in the dataset images. The logs in the dataset are labeled using the polygon labeling tool using Labelme [23] software, and the labeling effect is shown in Figure 5. Since the label information of the annotation will not be displayed in the image but can only be seen in other views of labelme, the annotated json file is converted into a visual image and the result is shown in Figure 5(c). The log ends in the sample image will be covered with a red mask and the labelme will be displayed in the lower right corner. At this point, the json file corresponding to each image contains only the corresponding original image labeling information, and the script of labelme2coco is used to synthesize all the json files of the labeled images into one json file containing all the labeled image labeling information, which is converted into a COCO data set and input to the network for training. The wood labeling information can be provided to the model to learn wood contour features, and wood counting can be realized according to the wood contour mask map.

The 600 clear images obtained after 4-fold data augmentation of the above dataset were divided into training set, validation set and test set according to the ratio of 4:1:1, and the large, medium and small targets (small target pixel area less than 32*32, medium target pixel area between 32*32 and 96*96, large target pixel area greater than 96*96) were



(a) Original image (b) Gaussian filtering (c) Brightness adjustment (d) Pepper noise

(e) Contrast adjustment (f) X-axis flip (g) Y-axis flip (h) XY-axis flip

FIGURE 4. Data amplification.

distinguished according to the target size division method of COCO dataset.), and the statistics of the labeled dataset are shown in Table 1.

IV. IMPROVEMENT OF YOLACT NETWORK

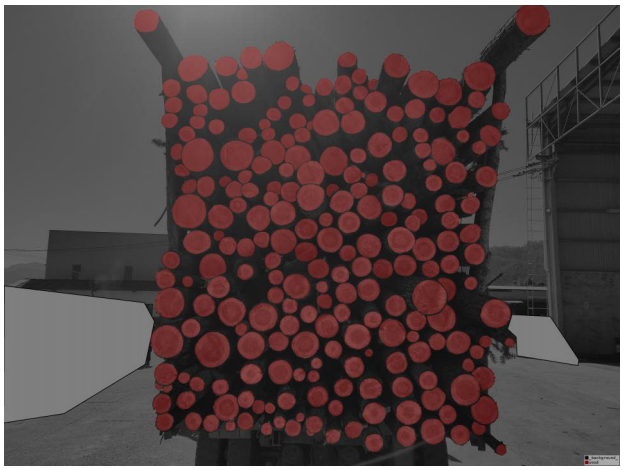
A. YOLACT ARCHITECTURE

The YOLACT model is a single-stage instance segmentation network that divides the instance segmentation task into two subtasks: prototype masks generation (prototype masks) and mask coefficients prediction (mask coefficients) for each instance.The structure of the YOLACT model is shown in Figure 6 and consists of five parts: backbone network (Backbone), mask template generation branch (Protonet [24]), prediction module (Prediction module), aggregation branch (assembly) and clipping module. The backbone network consists of ResNet and feature pyramid (FPN [25]), based on FPN to obtain feature images P5, P4, P3, and convolution operation on feature image P5 to obtain feature images P6, P7. Subsequently, the instance segmentation is divided into two parallel subtasks, one subtask inputs feature image P3 into Protonet to generate a series of mask templates (prototype masks), different mask templates have different sensitivity to different instances. Another subtask adds a mask coefficients prediction branch to the target detection branch, and generates mask coefficients (mask coefficients) representing instance masks in the mask templates while predicting the location and class of the target object bounding box. Finally, the mask coefficients are linearly combined with the mask template to obtain the instance mask, and then the image is cropped according to the predicted bounding box to achieve instance segmentation.

B. YOLACT-WOOD INSTANCE SEGMENTATION ALGORITHM

To address the problems of the YOLACT model in whole-vehicle wood detection and segmentation, this study makes the following improvements to the model to improve the detection and segmentation of log endfaces: 1) replace the feature extraction network and introduce the ResNeXt network as the backbone feature extraction network, compared with the ResNet network in the original YOLACT model, the ResNeXt network has 2) adding CBAM attention mechanism between ResNeXt feature extraction and FPN feature pyramid to enhance the specific target region of interest by channel attention to strengthen important features, suppress non-important features and spatial attention, while weakening irrelevant background regions to help the model locate and identify the region of interest more accurately The model also reduces the number of modules to retain more shallow features and enhance the feature extraction of small

(a) Wood Original Picture (b) End-face mask map



(c) Wood contour marking mask map

FIGURE 5. Data set annotation effect.

TABLE 1. Log end data set statistics.

Dataset information	Large Wood	Medium wood	Small wood	Total
Dataset	1413	47044	13697	621545
Training set	1128	31940	10008	43076
Validation set	145	7556	1841	9542
Test set	140	7548	1848	9536

diameter-level logs. 3) The CIoU loss function is used as the bounding box regression loss function to accurately measure the location of the prediction box, and the selection of the prediction box is optimized using DIoU-NMS, which can adjust the box while de-weighting to make the detection results more accurate. With DIoU, a distance metric, DIoU-NMS can better consider the relationship between frames to improve the detection accuracy of the model. the network structure of YOLACT-WOOD algorithm is shown in Figure 7.

1) ResNeXt MODULE

The initial backbone network of YOLACT is ResNet, which has high computational complexity and generates feature maps with low resolution and ignores the correlation of different spatial locations, which leads to low detection accuracy of the model. In order to improve the detection accuracy of the whole wood, this paper uses ResNeXt as the backbone network of the YOLACT model. the ResNeXt module can replace the original ResNet block with a parallel stack of blocks of the same topology without significantly increasing the parameter magnitude, thus improving the average recognition accuracy and improving the YOLACT model The ResNeXt module is used as the backbone feature extraction network, and the structural improvement of ResNeXt makes it have stronger feature expression capability.The structures of ResNet and ResNeXt are shown in Figure 8. From Fig. 8, it can be seen that ResNeXt is a single convolution of ResNet changed into a convolution with 32 branches, and the input

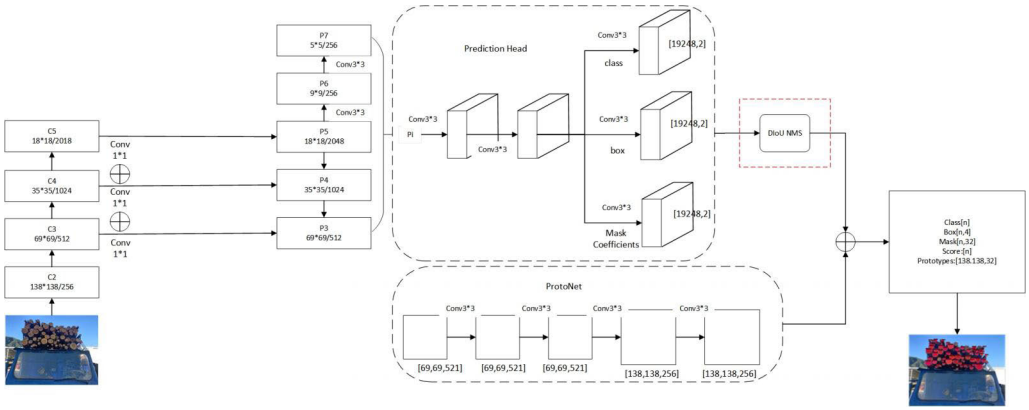


FIGURE 6. YOLACT network structure.

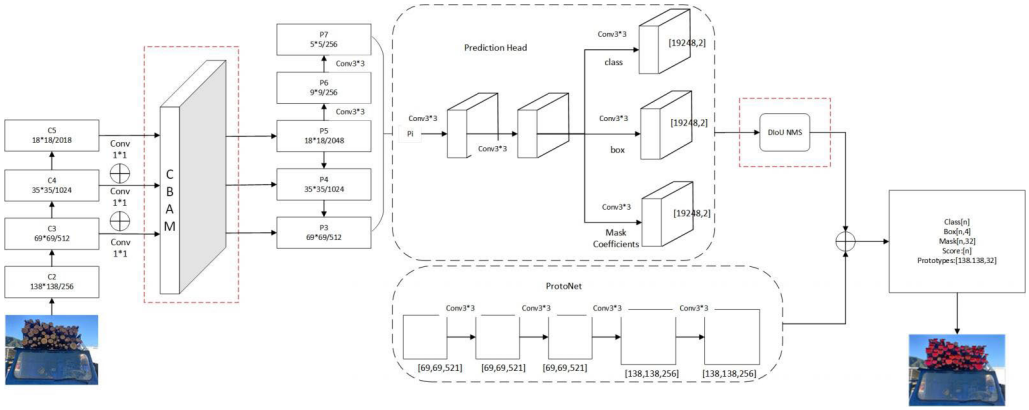


FIGURE 7. YOLACT-WOOD algorithm structure.

feature images are sent to each branch for convolution operation, and then the feature maps output from each branch are dimensionally stitched to get the final output.

In order to reduce the backbone model parameters and match the feature pyramid structure, the YOLACT-WOOD algorithm uses ResNeXt-50 as the backbone feature extraction network.The feature extraction module of the ResNeXt network is divided into four phases, each phase consists of 1×1 and 3×3 convolutional modules.The number of ResNeXt-50 corresponding to the four feature extraction layers is 3, 4, 6, 3.

2) CBAM ATTENTION MODULE

In the whole-vehicle wood inspection scale, the model needs to focus on the feature information of the wood end face. Therefore, to better extract the target features, the convolutional block attention module (CBAM) is added to the feature extraction network ResNeXt. CBAM is a lightweight dual attention mechanism proposed by Xie et al. [19] in 2018, which is a simple and effective for feedforward convolutional neural networks CBAM is a simple and effective attention module for feedforward convolutional neural networks, which differs from the attention mechanism of SE [26] in that the feature map will pass through both channel and spatial attention modules in turn to achieve dual conditioning, which can achieve better results in practical applications.The CBAM attention module consists of two parts: the channel attention module (CAM) and the spatial attention module (The structure of CBAM is shown in Figure 9.

The structure of the channel attention module is shown in Figure 10. In CAM, the input feature map F is extracted from the global features through two parallel branches of global maximum pooling and global average pooling; then the number of channels is compressed to $1/r$ by the multilayer perceptron (MLP) module respectively, and then expanded back to the original number of channels; then the outputs of the two branches are Then the outputs of the two branches are added element by element, and the weight coefficients M_c of CAM are obtained by a Sigmoid activation function; finally, the weight coefficients M_c are multiplied with the input feature map F to obtain the input features F' of the SAM module. computational equation (1) shows.

$$M_C(F) = \sigma\left\{W_1\left[W_0\left(F_{avg}^C\right)\right] + W_1\left[W_0\left(F_{max}^C\right)\right]\right\} \quad (1)$$

In equation (1), F is the input feature map, MLP is the multilayer perceptual layer, W_0 and W_1 are the single fully

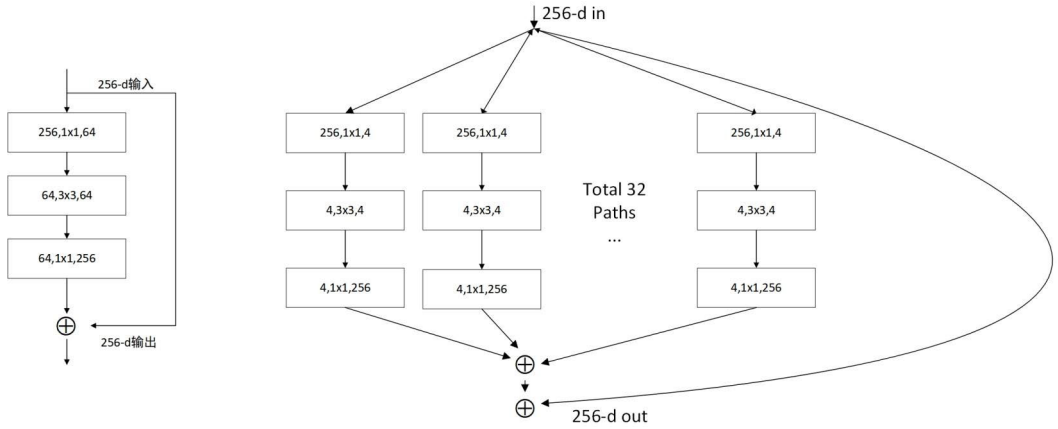


FIGURE 8. ResNet and ResNeXt structures.

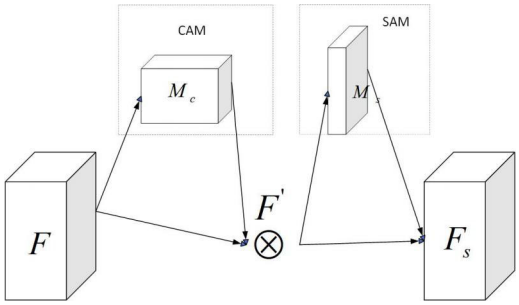


FIGURE 9. Attention mechanism CBAM overall network architecture.

connected layer in MLP, avg denotes the average pooling operation, max denotes the global maximum pooling operation, F^c_{avg} denotes the channel description feature after average pooling, and F^c_{max} denotes the channel description feature after maximum pooling.

The structure of the spatial attention module is shown in Figure 11. SAM first performs the maximum pooling operation and the average pooling operation on the channel for the input feature map F' of size $H \times W \times C$ to obtain two feature maps of $H \times W \times 1$ and splices these two feature maps together based on the channel. Then the convolution operation of convolution kernel 7×7 and Sigmoid activation operation are performed to obtain the weight coefficients M_s of the feature map, and finally the final features are obtained by multiplying and scaling the weight coefficients M_s and the feature map F'. The calculation formula is shown in (2):

$$M_s\left(F'\right) = \sigma\left[f^{7\times7}\left(F^s_{avg}; F^s_{max}\right)\right] \tag{2}$$

The $f^{7\times7}$ in equation (2) is the convolution operation with a convolution kernel size of 7×7.

3) CIoU LOSS FUNCTION AND DIoU-NMS

The loss function L_{loss} of the YOLACT model is defined as the sum of the classification loss L_{cls}, the bounding box regression loss L_{box}, and the segmentation loss L_{mask}. Among them, the bounding box regression loss L_{box} uses $SmoothL1$ as the bounding box regression loss function, and the loss is calculated for the length and width of the prediction box as well as the bias of the horizontal and vertical coordinates of the center point, where the calculation process is shown in Equation (3)(4)(5)(6)(7)(8)(9):

$$L_{smoothL1} = \sum_{i\in\{x,y,w,h\}} Smooth_{L1}\left(t^*_i - t_i\right) \tag{3}$$

$$SmoothL1\left(x\right) = \begin{cases} 0.5x^2, \, if \, |x| < 1 \\ |x| - 0.5, \, otherwise \end{cases} \tag{4}$$

$$u = \left(t_x, t_y, t_w, t_h\right) \tag{5}$$

$$t^*_x = \left(x^* - x_a\right) \tag{6}$$

$$t^*_y = \left(y^* - y_a\right) \tag{7}$$

$$t^*_w = \ln\left(w^*/w_a\right) \tag{8}$$

$$t^*_h = \ln\left(h^*/h_a\right) \tag{9}$$

where x_a,y_a,w_a,h_a denotes the centroid coordinates, length and width of the anchor; x^*,y^*,w^*,h^* denotes the centroid coordinates, length and width of the real frame; u denotes the anchor bias matrix of the network prediction.

Since $SmoothL1$ lacks the calculation of the intersection ratio (IoU) and the minimum outer rectangle, it is not accurate enough to measure the position of the predicted frame. Therefore, the CIoU loss function is introduced to include the intersection ratio, minimum outer rectangle, geometric center distance and aspect ratio of the predicted frame and the real frame into the boundary frame regression loss calculation, which can accurately measure the boundary frame regression performance compared with the original loss function. The definition of the CIoU boundary frame regression loss function is shown in Equation (10)(11)(12)(13):

$$L_{CIoU} = 1 - IoU\left(b, b^{gt}\right) + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha v \tag{10}$$

$$IoU\left(b, b^{gt}\right) = \frac{\left|b \cap b^{gt}\right|}{\left|b \cup b^{gt}\right|} \tag{11}$$

$$\alpha = \frac{v}{\left(1 - P_{IoU} + v\right)} \tag{12}$$

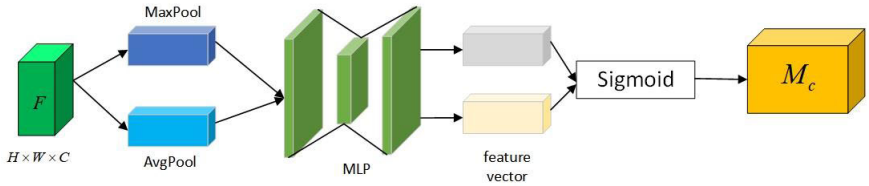


FIGURE 10. CAM channel attention module structure diagram.

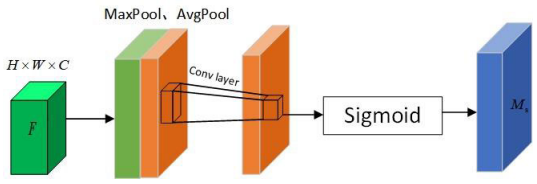


FIGURE 11. SAM spatial attention module structure diagram.

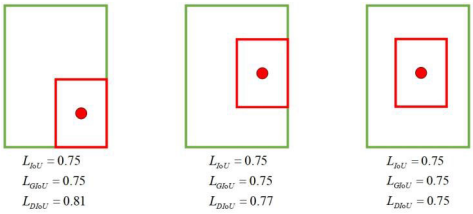


FIGURE 12. Comparison of three loss functions.

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{13}$$

where ρ denotes the distance between the prediction frame b and the geometric center of the target frame b^{gt}; c denotes the diagonal length of the smallest outer rectangle of the prediction frame and the target frame.

The YOLACT model structure uses fast non-maximum suppression(Fast-NMS) algorithm to eliminate redundant candidate frames and obtain the final prediction frame. although the Fast NMS algorithm will greatly reduce the redundancy of candidate frames, it is easy to cause the target candidate frames of different instances with high overlap rate to be mistakenly deleted, which leads to some neighboring similar objects to be easily regarded as one instance. Due to the very close distance between each log of the whole truckload of wood, the high overlap of target frames causes the wood miss detection problem. Therefore, in this paper, we introduce DIoU in Fast NMS calculation, and the definition of DIoU is shown in Equation (14)(15):

$$DIoU\left(B_i, B_j\right) = IoU\left(B_i, B_j\right) - R_{DIoU}\left(B_i, B_j\right) \tag{14}$$

$$R_{DIoU}\left(B_i, B_j\right) = \frac{\rho^2\left(B_i, B_j\right)}{c^2} \tag{15}$$

The ρ in equation (15) denotes the Euclidean distance between the centroids of candidate boxes B_i and B_j; c denotes the diagonal length of the smallest outer rectangle of candidate boxes B_i and B_j. Figure 12 compares the three loss functions of IoU, GIoU and DIoU. The figure gives the overlap relationship between the three groups of target bounding boxes and predicted bounding boxes, the overlap positions are different in the three cases but the IoU loss and GIoU loss are exactly the same, so these two losses cannot express the bounding box overlap relationship well, but the losses calculated by DIoU are not the same for the three cases, obviously DIoU is more reasonable. In this paper, by comparing the detection effects of Fast-NMS and DIoU-NMS in the test set images, it can be found that DIoU-NMS can improve the problem that the log end face is occluded and missed in the test set images, and the results are shown in Figure 13, DIoU-NMS can improve the situation when multiple targets are too close to each other causing smaller targets to be occluded and thus filtered out.

V. EXPERIMENTS AND RESULTS ANALYSIS

A. EXPERIMENTAL ENVIRONMENT AND PARAMETER SETTING

The experimental environment is a deep learning framework built with Pytorch 1.8.2 under Linux using NVIDIA GeForce RTX 3090 24GB (GPU) server, and GPU acceleration using CUDA11.1 toolkit. The MMDetection [27] toolbox was used to complete the training of the YOLACT model. The model is trained by first using pre-trained weights for migration learning to complete the initialization of the network parameters, and then the labeled whole-vehicle wood dataset is transformed into COCO dataset [28] format and fed into the network for training. According to the detection requirements of this experiment, other general training configuration parameters are: the number of target classes *num_classes* = 1, the detection class is ''wood''; set the learning rate to 0.001, the training Epoch size is 55, and the first 500 iterations are set to change linearly in the Warm up, which is used to Stabilize the parameter gradient at the early stage of training, and optimize the gradient transfer based on stochastic gradient descent (SGD); use GeLU as the activation function of the model. At Epoch equal to 20,42,49,52, the learning rate is multiplied by 0.1 times of the descent strategy.

B. EVALUATION INDICATORS

In order to evaluate the feasibility of the optimized YOLACT model for whole-truck wood detection and segmentation more comprehensively and objectively, four indexes,

(a) Fast-NMS (b) DIoU-NMS

FIGURE 13. Comparison of the effect of non-extreme value suppression function.

namely, mean accuracy (mAP [29]), mask intersection ratio (IoU_{mask}), wood recognition rate, and frames per second (FPS) transmission, are used to evaluate the log detection results.

1) mAP METRICS

In this study, the common evaluation metric mAP (Mean average precision) of target detection and instance segmentation model defined by COCO dataset is used. mAP is selected for the training accuracy of the model under the IoU threshold of 0.5 because the selection of IoU threshold affects the accuracy and recall size. mAP is calculated as shown in Equations (16) to (19):

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$AP = \int_0^1 P(r)\, dr \tag{18}$$

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{19}$$

where TP denotes the number of samples where the model predicted category matches the true labeled category; FP denotes the number of samples where the model predicted category does not match the true labeled category; and FN denotes the number of samples where the prediction is background but the true label is other categories. mAP_s, mAP_m, and mAP_l are used to denote the average mean accuracy of the target size at small, medium, and large levels, respectively.

2) IoU_{mask} METRICS

Although mAP has better characterization performance for deep learning models, mAP is more suitable for evaluating the classification confidence and cannot examine the actual segmentation effect of the mask. To address this problem, the IoU value is chosen to evaluate the quality of Mask. The calculation method of IoU_{mask} is shown in formula (20):

$$IoU_{mask} = \frac{area\,(P) \cap area\,(G)}{area\,(P) \cup area\,(G)} \tag{20}$$

The quality of Mask is quantitatively evaluated by calculating the intersection ratio between the wood Mask region (P) inferred by the model and the manually labeled wood profile region (G), and the IoU_{mask} value is used to further measure the accuracy of the model for wood profile segmentation.

3) WOOD RECOGNITION RATE

The higher the recognition rate of the model, the better the detection performance of the model. The model counts the timber segmentation masks output from the test set images, and counts the number of large, small, medium and large sized timbers, the number of mis-detected timbers, the number of missed timbers and the number of true detected timbers, and compares the actual number of timber statistics in the test set to calculate the detection rate and true detection rate of the model.

C. EXPERIMENTAL DESIGN ANALYSIS

The image size of the dataset used in this paper is 1600*1200, as the image input size of the original YOLACT model is 1333*800, based on the short side of the size will lead to a 55.6% reduction of the wood size in the original image after inputting into the model, if the image input size is increased by 50% to 2000*1200, that is, the target size in the original model is increased by 125%, which is very beneficial for feature extraction of small targets. Five groups of experiments are designed in this study, which are the comparison of different backbone network detection performance, the ablation experiment of different improvement methods on model performance, the comparison experiment of different

loss functions, the comparison experiment of mainstream instance segmentation models and the comparison of different improvement methods on log true detection performance. The performance of this research method is comprehensively analyzed by the above five sets of experiments.

1) COMPARISON EXPERIMENTS OF DIFFERENT BACKBONE NETWORKS

In order to further verify the effectiveness of the backbone network ResNeXt on whole-vehicle wood recognition, based on the YOLACT framework, ResNet50, ResNet101, ResNeXt50 and ResNeXt101 were used as the backbone networks of the model for experimental validation, and the experimental results are shown in Table 2.

As can be seen from Table 2, comparing ResNeXt-50 with ResNet-50 we can see that the mAP of the model is improved by about 2.5%, due to the structural improvement of ResNeXt relative to ResNet which increases the network width and has a stronger feature extraction ability, so on the basis of the same number of network layers, ResNeXt-50 and ResNeXt-101 are more accurate than ResNet-50 and ResNet-101 with higher average recognition accuracy. According to the above table, it can be seen that ResNeXt-101 has deeper network depth and stronger theoretical feature extraction ability compared to ResNeXt-50, but too deep network will bring problems such as excessive amount of parameters and gradient disappearance, which will affect the recognition accuracy of the network. Therefore, the ResNeXt-50 network is selected as the backbone network of the model for extracting the wood endface features of the whole vehicle.

2) IMPACT OF IMPROVED METHODS ON MODEL PERFORMANCE

To analyze the impact of all the improvement methods proposed in this study on the YOLACT model algorithm, the different improvement parts were analyzed by designing eight sets of comparison experiments, using the same training parameters for each scheme. The effects of the different methods on the performance of the model are shown in Table 3.

By comparing Experiment 1 and Experiment 2, we can find that replacing the backbone network with ResNeXt improves the mAP of the model by 2.2% and IoU_{mask} by 1.3%; comparing Experiment 1 and Experiment 5, we can find that adding the CBAM attention mechanism to the original backbone network ResNet improves the mAP of the model by 2.8% and IoU_{mask} by 2.8% and IoU_{mask} by 1.8%; comparing Experiment 1 and Experiment 6, we can find that the introduction of the new loss function and non-maximum suppression improves the mAP by 2% and IoU_{mask} by 1.8; comparing Experiment 2 and Experiment 3, we can find that the addition of the CBAM attention mechanism to the improved backbone network improves the mAP by 1.7 and IoU_{mask} by 0.3%; comparing Experiment 1 and Experiment 4, we can see that mAP improves by 5.6% and IoU_{mask} improves by 2.6% under the simultaneous improvement of three improvement points to the model, both of which have a large improvement, but the speed decreases by 5.9 FPS.Grad-CAM is used to generate the heat map of the backbone network, and the output of the backbone network is visualized and analyzed. The heat map generated before and after the improvement of the backbone network is shown in Figure 14. The area of interest of the original backbone network includes wood end faces and background areas, and the focus on small target wood is not enough. The improved backbone network enhances the ability to extract the deep information of the image, which improves the recognition ability of the backbone network for small target wood and greatly suppresses the interference of the background.

3) COMPARISON EXPERIMENTS OF DIFFERENT LOSS FUNCTIONS

In order to compare the superiority of CIoU loss function in recognizing whole wood end faces, three different loss functions: DIoU, GIoU [30], and CIoU are compared based on the YOLACT framework. experiments are conducted on three different bounding box regression loss functions, and the three schemes have identical network structures except for different loss functions, and use the same training parameters.

As can be seen from Table 4, compared with Smooth L1 loss function used by the original network, which can only calculate the loss based on the offset of the prediction box but cannot accurately describe the position relationship between the prediction box and the real box, the average recognition accuracy and mask segmentation quality of the model are improved after CIoU loss function is used. The CIoU loss function takes into account the overlapping areas between boundary boxes and provides a more accurate measure of distance.

4) COMPARISON EXPERIMENTS OF DIFFERENT NETWORK MODELS

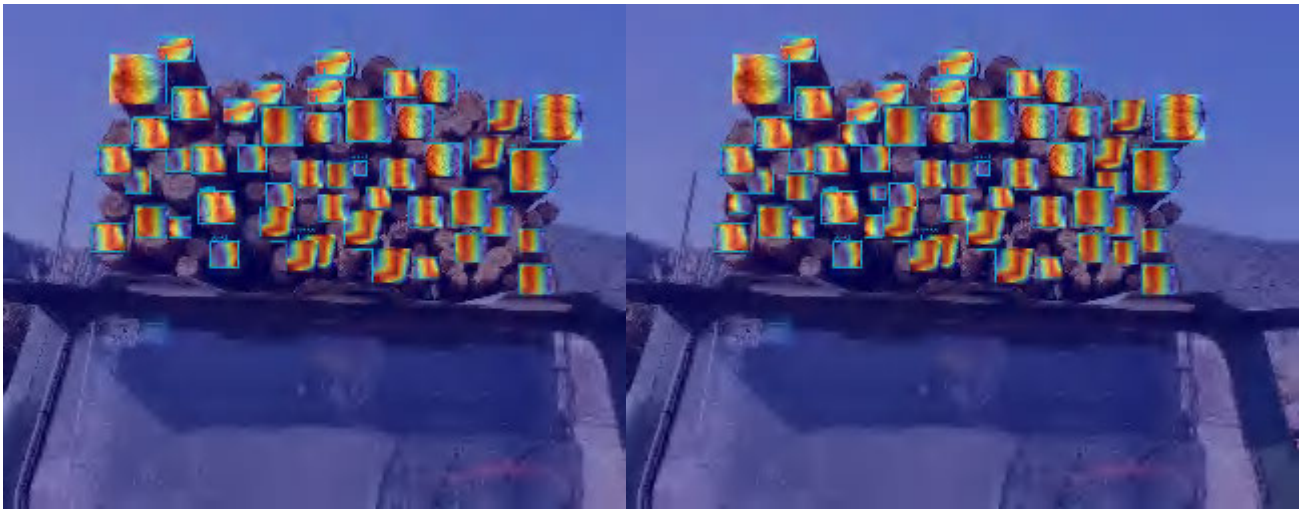
To further verify the instance segmentation effect of the algorithm in this paper, the algorithm in this paper is compared with the current advanced instance segmentation algorithms, and the Mask R-CNN [31], Cascade Mask R-CNN [32], YOLACT and YOLACT_WOOD models are trained respectively, and the trained models are used to verify the detection and segmentation effect of the whole vehicle wood. The training results are shown in Table 5, from which we can see that the YOLACT_WOOD algorithm model proposed in this paper is higher than Mask R-CNN, Cascade Mask R-CNN, and YOLACT models in mAP_50 all by 4.2%, 3.3%, and 5.6%, respectively. Although the detection segmentation accuracy of Mask R-CNN and Cascade Mask R-CNN models is higher, their parametric numbers are too large to limit the inference speed of the models. the main reason for the lower detection segmentation accuracy of YOLACT model is that YOLACT model is a single-stage instance segmentation model, while Mask R-CNN and Cascade Mask R-CNN models are both Therefore, the detection

TABLE 2. Comparison experiment of different backbone networks.

Backbone Network	mAP_50	mAP_s	mAP_m	mAP_l	FPS(img/s)	Number of participants
ResNet50	0.735	0.484	0.736	0.842	26.8	34.73M
ResNet101	0.742	0.451	0.644	0.717	21.7	53.72M
ResNeXt50	0.758	0.543	0.723	0.889	26.2	35.10M
ResNeXt101	0.753	0.579	0.772	0.911	20.6	59.96M

TABLE 3. Experimental comparison results of different improvement points.

No	ResNeXt	CBAM	CIoU and DIoU-NMS	mAP_50	mAP_s	mAP_m	mAP_l	FPS(img/s)	Parameter	IoU_{mask}
1	NO	NO	NO	0.735	0.484	0.746	0.892	26.8	34.73M	0.891
2	YES	NO	NO	0.758	0.543	0.723	0.909	26.2	35.10M	0.904
3	YES	YES	NO	0.775	0.583	0.791	0.911	21.6	37.26M	0.907
4	YES	YES	YES	**0.791**	**0.617**	**0.809**	**0.946**	**20.9**	**40.30M**	**0.917**
5	NO	YES	NO	0.763	0.557	0.754	0.916	26.0	35.24M	0.909
6	NO	NO	YES	0.755	0.492	0.759	0.895	26.5	35.11M	0.909
7	NO	YES	YES	0.777	0.598	0.810	0.933	21.4	37.93M	0.910
8	YES	NO	YES	0.769	0.579	0.820	0.912	22.9	39.75M	0.907



(a) Before backbone network improvement (b) After backbone network improvement

FIGURE 14. Backbone network heat map.

TABLE 4. Comparison experiments of different loss functions.

Loss function	mAP_50	IoU_{mask}
SmoothL1	0.775	0.907
DIoU	0.782	0.912
GIoU	0.769	0.889
CIoU	0.791	0.917

segmentation accuracy is not as good as that of the two-stage model. The segmentation accuracy of the YOLACT model is improved by optimizing the backbone network, introducing attention mechanism and loss function.

5) ANALYSIS OF COMPARATIVE EXPERIMENTAL RESULTS OF WOOD TESTING PERFORMANCE

Since this paper targets the whole wood end face for inspection segmentation, an important metric for evaluating the model is the detection of the number of woods in the test set. The contour finding and contour counting of the wood segmentation masks of the YOLACT model are completed using the OpenCV library. Figure 15 shows the counting results for the comparison experiments of different network models in Table 6. In the top left corner of it, the counting results of the wood in the picture are printed, divided into three sizes of wood, small, medium and large. And different shades of color are used to distinguish between small and large woods.

The models used in the five groups of experiments, Experiment 1, Experiment 2, Experiment 4, Experiment 5 and Experiment 6, were selected according to the different improvement points in Table 4 for wood detection performance, and the wood detection rate and wood true detection rate of the models were obtained as shown in Tables 6 and 7.

TABLE 5. Comparative experiment of different network models.

Model	mAP_50	mAP_s	mAP_m	mAP_l	FPS	Parameter	IoU_{mask}
Mask R-CNN	0.749	0.495	0.754	0.901	6.2	123.75M	0.898
Cascade Mask R-CNN	0.758	0.522	0.765	0.924	11.1	76.80M	0.905
YOLACT	0.735	0.484	0.746	0.892	26.8	34.73M	0.891
YOLACT_WOOD	**0.791**	**0.617**	**0.809**	**0.946**	**20.9**	**40.30M**	**0.917**

(a) YOLACT (b) Mask R-CNN

(c) Cascade Mask R-CNN (d) YOLACT_WOOD

FIGURE 15. Results of wood testing counts for different models.

From Table 6, it can be seen that the optimized model experiment 4 has improved the detection performance of each size of wood compared with the initial model experiment 1, especially the detection performance of small wood has improved more obviously, and the wood detection rate has increased from 89.835% to 96.836%, which is an increase of about 7%.

From Table 7, we can see that the original YOLACT model (Experiment 1) has the lowest wood true detection rate of 89.54% among the five groups of experiments, while the improved YOALCT_WOOD model (Experiment 4) has the highest wood true detection rate of 96.61% among the five groups of experiments. The highest false detection rate of 1.76% was achieved for the model (Experiment 6) after

TABLE 6. Performance statistics of different improvement points for each size of wood inspection.

Number of wood	No	All wood	Small wood	Medium wood	Large wood	Detection rate(%)
Actual number		9536	1848	7548	140	
number of detections	1	8567	1625	6891	51	89.835
	2	8922	1759	7062	101	93.559
	4	9234	1799	7321	114	**96.836**
	5	9038	1771	7148	119	94.781
	6	8726	1693	7003	30	91.491

TABLE 7. Statistical table of different improvement points on wood true inspection performance.

No	Number of logs	Number of detections	Number of false detections	Number of true inspections	Number of missed tests	True inspection rate(%)	False detection rate(%)
1	9536	8567	28	8539	997	89.54	0.35
2	9536	8922	44	8878	658	93.10	0.49
4	9536	9234	21	9213	323	**96.61**	**0.23**
5	9536	9038	76	8962	574	93.98	0.84
6	9536	8726	154	8572	964	89.89	1.76

improving only the loss function and non-maximum suppression, while the lowest false detection rate of 0.23% was achieved for the YOLACT_WOOD model (Experiment 4). Therefore, the YOLACT_WOOD model can effectively improve the detection capability for each size of wood.

VI. CONCLUSION

To address the problems of slow detection speed, low detection accuracy, dense wood stacking and easy to be obscured and missed, we propose a segmentation method YOLACT_WOOD based on YOLACT algorithm for end-face detection of whole-vehicle wood. Images are data enhanced to improve the detection capability of the model for small and medium diameter wood; finally, CIoU is selected as the boundary frame regression loss function to improve the problem of inaccurate boundary frame prediction, as well as combining DIoU with Fast-NMS to solve the problem of false detection and missed detection. The improved algorithm was tested under NVIDIA GeForce RTX 3090 test conditions using the test set. mAP improved by 5.6% compared to the initial algorithm, reaching 79.1%, and FPS reached 20.9 frames/second, compared to the detection speed of the Mask R-CNN model, which improved by 14.7 frames/second, and IoU_{mask} improved by 2.6% to 0.917, and the wood true detection rate reaches 96.61% and the false detection rate is 0.23%. The experiments show that the trained model has the ability to check the ruler quickly and also has good detection effect for all sizes of wood, and the model has strong robustness and generalization ability to meet the field demand of industrial production.

REFERENCES

[1] S. Mallapaty, "How China could be carbon neutral by mid-century," *Nature*, vol. 586, no. 7830, pp. 482–483, Oct. 2020, doi: 10.1038/d41586-020-02927-9.

[2] J. G. Canadell and M. R. Raupach, "Managing forests for climate change mitigation," *Science*, vol. 320, no. 5882, pp. 1456–1457, Jun. 2008, doi: 10.1126/science.1155458.

[3] B. Vira, C. Wildburger, and S. Mansourian, "Forests, trees and landscapes for food security and nutrition: A global assessment report," IUFRO World Ser., Tech. Rep., 2015, vol. 33, doi: 10.11647/obp.0085.01.

[4] S.-W. Hwang, T. Lee, H. Kim, H. Chung, J. G. Choi, and H. Yeo, "Classification of wood knots using artificial neural networks with texture and local feature-based image descriptors," *Holzforschung*, vol. 76, no. 1, pp. 1–13, Jan. 2022, doi: 10.1515/hf-2021-0051.

[5] B. Galsgaard, D. H. Lundtoft, I. Nikolov, K. Nasrollahi, and T. B. Moeslund, "Circular Hough transform and local circularity measure for weight estimation of a graph-cut based wood stack measurement," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 686–693, doi: 10.1109/WACV.2015.97.

[6] D. H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern Recognit.*, vol. 13, no. 2, pp. 111–122, 1981, doi: 10.1016/0031-3203(81)90009-1.

[7] A. V. Kruglov, "The algorithm of the roundwood volume measurement via photogrammetry," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2016, pp. 1–5, doi: 10.1109/DICTA.2016.7797088.

[8] A. Kruglov and E. Shishko, "Log pile measurement through 3D modeling," in *Proc. 40th Int. Conf. Telecommun. Signal Process. (TSP)*, Jul. 2017, pp. 263–266, doi: 10.1109/TSP.2017.8075983.

[9] A. V. Kruglov, "Development of the rounded objects automatic detection method for the log deck volume measurement," in *Proc. 1st Int. Workshop Pattern Recognit.*, vol. 10011, 2016, pp. 13–18, doi: 10.1117/12.2242172.

[10] F. Budiman, R. Mardiyanto, Tasripan, and R. Rachmat, "A handy and accurate device to measure smallest diameter of log to reduce measurement errors," in *Proc. Int. Seminar Intell. Technol. Appl. (ISITIA)*, Jul. 2016, pp. 423–428, doi: 10.1109/ISITIA.2016.7828697.

[11] G. Chen, Q. Zhang, M. Chen, J. Li, and H. Yin, "Rapid detection algorithms for log diameter classes based on binocular vision," *J. Beijing Jiaotong Univ.*, vol. 42, no. 2, pp. 22–30, 2018, doi: 10.11860/j.issn.1673-0291.2018.02.004.

[12] C. Keck and R. Schödel, "Reference measurement of roundwood by fringe projection," *Forest Products J.*, vol. 71, no. 4, pp. 352–361, Oct. 2021, doi: 10.13073/FPJ-D-21-00024.

[13] N. Samdangdech and S. Phiphobmongkol, "Log-end cut-area detection in images taken from rear end of eucalyptus timber trucks," in *Proc. 15th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, Jul. 2018, pp. 1–6, doi: 10.1109/JCSSE.2018.8457388.

[14] H. Tang, K. Wang, J. Gu, X. Li, and W. Jian, "Application of SSD framework model in detection of logs end," *J. Phys., Conf. Ser.*, vol. 1486, no. 7, Apr. 2020, Art. no. 072051, doi: 10.1088/1742-6596/1486/7/072051.

[15] R. Cai, P. Lin, and Y. Lin, "A detection approach for bundled log ends based on an improved YOLOv4-Tiny network," *Video Eng.*, vol. 45, no. 9, pp. 92–99, 2021, doi: 10.16280/j.videoe.2021.09.028.

[16] Y. Lin, H. Zhao, Z. Yang, and M. Lin, "An equal length log volume inspection system using deep-learning and Hough transformation," *J. Forestry Eng.*, vol. 6, no. 1, pp. 136–142, 2021, doi: 10.13360/j.issn.2096-1359.202003022.

[17] Y. Lin, R. Cai, P. Lin, and S. Cheng, "A detection approach for bundled log ends using K-median clustering and improved YOLOv4-tiny network," *Comput. Electron. Agricult.*, vol. 194, Mar. 2022, Art. no. 106700, doi: 10.1016/j.compag.2022.106700.
[18] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9156–9165, doi: 10.1109/ICCV.2019.00925.
[19] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995, doi: 10.1109/CVPR.2017.634.
[20] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19, doi: 10.48550/arXiv.1807.06521.
[21] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12993–13000, doi: 10.1609/aaai.v34i07.6999.
[22] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666, doi: 10.1109/CVPR.2019.00075.
[23] A. Torralba, B. C. Russell, and J. Yuen, "LabelMe: Online image annotation and applications," *Proc. IEEE*, vol. 98, no. 8, pp. 1467–1484, Aug. 2010, doi: 10.1109/JPROC.2010.2050290.
[24] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, doi: 10.48550/arXiv.1703.05175.
[25] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.
[26] 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, doi: 10.1109/CVPR.2018.00745.
[27] K. Chen et al., "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*, doi: 10.48550/arXiv.1906.07155.
[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland: Springer, Sep. 2014, pp. 740–755, doi: 10.48550/arXiv.1405.0312.
[29] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010, doi: 10.1007/S11263-009-0275-4.
[30] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666, doi: 10.1109/CVPR.2019.00075.
[31] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.322.
[32] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162, doi: 10.1109/CVPR.2018.00644.

JISHI ZHENG received the Ph.D. degree in engineering from Central South University, in 2015. From January to July 2019, he was a Visiting Scholar with the Robotics Laboratory, Department of Computer and Electronic Engineering, University of Essex, U.K. He is currently a person in charge of the Internet of Things, the Director of the Department of Traffic Information and Control, and a part-time Executive Director of the Fujian Aeronautical Society. His main research interests include the application of artificial intelligence in the industry and the research on the flight control algorithm of drones. He has presided over and participated in more than ten provincial and municipal scientific research projects.

JUNJIE ZHENG was born in Xianyou, Fujian, in 1999. He received the bachelor's degree majoring in software engineering from the School of Computer Science and Mathematics, Fujian University of Technology, where he is currently pursuing the master's degree in electronic information with the School of Transportation. His research interests include machine vision, deep learning, and image processing.

SHIWEN ZHANG is currently pursuing the master's degree in electrical engineering with the Fujian University of Technology. His research interests include the application of deep learning in the field of forestry inspection.

HONGHUI YU is currently pursuing the master's degree in electrical engineering with the Fujian University of Technology. His research interests include the application of object detection in forestry small object detection.

LINGHUA KONG received the bachelor's degree in physics from Nankai University, in 1983, the master's degree in physics from the Institute of High Energy Physics, Chinese Academy of Sciences, in 1988, and the Ph.D. degree in mechanical engineering from the Department of Mechanical Engineering, McGill University, Canada, in 2004. He received a Postdoctoral Fellowship with the Georgia Institute of Technology, in 2005. He is currently a Professor with the School of Mechanical and Automotive Engineering, Fujian University of Technology. He designed and developed a variety of new products and equipment, and obtained 15 patents; published 20 influential articles included in SCI/EI as the first author. His main research areas are multispectral and plasma fields.

DING ZHIGANG received the master's degree in bioengineering from Jilin University, in 2007. He has been a Senior Engineer with the Fujian University of Technology, since 2008. He is currently the Head of the Vehicle Engineering Laboratory. He has participated in more than ten national 863 programs and provincial and municipal scientific research projects. Currently, he is mainly engaged in the research and development of new energy vehicle power system assembly, product visual inspection, and intelligent equipment.