IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 January 2024, accepted 7 March 2024, date of publication 18 March 2024, date of current version 22 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3376473

APPLIED RESEARCH

Efficient Object Detection and Recognition of Body Welding Studs Based on Improved YOLOv7

HONG HUANG[1], XIANGQIAN PENG[1], XIAOPING HU[2], AND WENCHU OU[1]
[1]School of Mechanical Engineering, Hunan University of Science and Technology, Xiangtan 411201, China
[2]Hunan Provincial Key Laboratory of Health Maintenance for Mechanical Equipment, Xiangtan 411201, China

Corresponding authors: Hong Huang (huanghong3722@126.com) and XiangQian Peng (redtailfox@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61572185.

ABSTRACT The welding stud is a widely used part in automobile manufacturing, and its welding quality plays a crucial role in component assembly efficiency and vehicle quality. In welded stud target inspection, the complex body environment and different lighting conditions will have a certain impact on the inspection accuracy, and most of the existing methods have limited efficiency. In this paper, in order to solve the problems of low accuracy and slow speed in the stud target inspection process, we propose an innovative welding stud target inspection method based on YOLOv7. First, the EfficientFormerV2 backbone network is adopted to utilize the new partial convolution, which can extract spatial features more efficiently, reduce redundant computation, and improve the detection speed. Secondly, the bounding box loss function is changed to NWD, which reduces the loss value, accelerates the convergence speed of the network model, and better improves the detection of studs. After the test, the improved YOLOv7 network model is better than the traditional network in both speed and accuracy of welded stud target detection. (1) The mAP0.5 increased from 94.6% to 95.2%, and the mAP0.5:0.95 increased from 63.7% to 65.4%. (2) The detection speed increased from 96.1 f/s to 147.1 f/s. The results of the study can provide technical support for the subsequent tasks of automatic detection and position estimation of body welding studs.

INDEX TERMS Welding studs, object detection, EfficientFormerV2, NWD, YOLOv7.

I. INTRODUCTION

In the process of automobile manufacturing, studs are the key parts to connect and fix the automobile body with other parts, and there can be hundreds of studs on each vehicle, and there are many different models [1], [2], [3]. The studs are mainly fixed on the metal surface of the car body by welding. If the welding position has a large deviation or the stud model welded on the surface is inconsistent with the design program, it may lead to subsequent assembly difficulties or even being unable to complete the assembly, which will have a greater impact on the reliability and stability of the whole car. Therefore, for accurate detection of the location and type of welded studs, timely detection of leakage welding, wrong welding, and other issues, for the enhancement of automotive raw for the body stud welding quality detection problems, in the past, due to technical reasons, the detection of the main relies on two ways: first, through the artificial visual confirmation of the work piece studs whether to be placed, the method exists in the high cost, low efficiency, heavy workload, large detection error, and other shortcomings [4], there will still be leakage welding and wrong welding. The second is to detect through the sensor; the detection accuracy of this method is higher, but the detection time is longer, and because of the limitation of the body size, the size of the stud cannot cover the different sizes of studs and cannot meet the needs of the modern industry for on-line, high-precision, high-efficiency measurement.

In automotive instrumentation inspection, welding process quality control, sheet metal contour tracking measurement, and other fields, machine vision inspection has become increasingly prevalent as computer and image processing

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

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

technologies continue to advance [5]. This is primarily due to the non-contact nature of the inspection, rapid speed, and high accuracy of machine vision inspection. In 2014, Wang et al. utilized the method of monocular vision to extract the feature information of the welded studs through the techniques of edge detection, feature extraction, etc., and finally through the computational model to realize the measurement of the positional offset and tilt angle offset of the welding stud [4]. The method can better meet the measurement needs of automotive and other manufacturing industrial fields, but the positional accuracy and angular accuracy are low. SONG et al. proposed an edge detection algorithm that utilizes morphology and wavelet transform to accurately identify bolt structures, which is verified to have strong noise immunity but can only handle images with simple backgrounds [6].

In recent years, people have begun to widely apply deep learning methods to machine vision target detection in different scenarios, including fruit detection [7], [8], [9], vehicle detection [10], ship detection [11], [12], defect detection [13], [14], [15], behavior detection [16], [17], etc., and have achieved good results. Compared with traditional target detection algorithms, deep convolutional networks can automatically learn multi-level feature models from training data and have strong generalization and feature extraction abilities. Deep learning-related algorithms are mainly classified into two categories: one is single-stage target recognition models, such as the YOLO (You Only Look Once) series and the SSD series; the other is two-stage recognition models, such as R-CNN and Faster R-CNN. Currently, some researchers have already applied these deep learning network structures to study target detection. In 2019, Lian proposed a detection method based on VGG and Faster RCNN models to study and experiment on the missing problem of stud nuts on top of automotive gold parts, which can get the location information of the picture where the stud nuts are located, but the number of samples in the method is small, the mAP value is only 30.4%, and there is no detection of the specification dimensions or information such as the vertical angle [18]. Since single-stage algorithms only need to scan the image once to produce detection results, they have higher efficiency compared to two-stage algorithms. Zhang et al. introduced an enhanced SSD network that utilizes multi-window, multi-scale fusion to rectify the drawbacks of the conventional SSD, namely its lack of sensitivity to small targets, and the mAP was improved to 43.2% [19]. Yang et al. used an improved YOLOv3-tiny network to recognize bolts, and the mAP increased from 81.3% to 83.9%, but the method is only applicable to the angle at which the camera shoots the object vertically [20]. In order to solve the problems caused by the missing bolt detection of steel structure canopies in passenger stations of high-speed railroads, Wang et al. used the YOLOv4 convolutional neural network algorithm to establish a bolt missing detection system, which first annotates the bolts of the steel structure canopies and the contact network collected in the field, and utilizes the data enhancement operations, such as CutMix and Mosaic, which are used to increase the diversity of the training data, and finally the accuracy of the system category recognition reaches more than 85%, but the method is more prone to overfitting phenomena [21]. Zhang et al. got around the problems with current detection methods by combining a lightweight YOLOv4 neural network with the photometric stereo method to find welded stud positions on targets [22]. A good answer to the issue is the proposed stud position detection system, which combines deep learning and photometric stereo for target detection in industrial production. Although the designed stud position detection system offers a foundation for integrating deep learning with photometric stereo for target detection in industrial production, the approach is unsuitable for mass production and has a high equipment cost. Wang et al. propose a bolt detection and positioning system based on a neural network and RGB-D camera that employs lightweight YOLOv5s-T to identify the bolts and screen the main bolts, realizing the fast guidance of the end of the 6-degree-of-freedom robotic arm to the fastening point of the bolts [23]. The mean average accuracy of the method reached 94%, but the FPS was only 5.83.

Despite the achievements in stud detection, there is still a lot of room for improvement in detection accuracy and speed for different bolt targets in industrial automated online inspection. Therefore, how to further improve balanced detection accuracy and speed is still a challenging issue. In addition, when the ambient light changes, the possible leakage of detection and the decrease in confidence level are also urgent problems to be solved. Based on the above discussion, this paper proposes a target detection algorithm for body stud recognition in an automated production shop environment to realize a two-way improvement in detection speed and accuracy for the body stud target detection task. This paper primarily presents the following contributions:

(1) Replace the backbone network of the YOLOv7 algorithm with EfficientFormerV2, so that the network reduces memory access and redundant computation and improves the computing speed of the detection process.
(2) Replace the loss function with NWD to solve the problem of slow convergence and low efficiency of the model during the training process, which effectively improves the performance of the network in accurately detecting small targets with welded studs.
(3) Propose an improved YOLOv7 algorithm applied to the task of detecting welded stud targets in automobile bodies, and conduct ablation experiments and comparative tests of the improved YOLOv7 algorithm on the stud dataset. Compared with the existing algorithms, the improved YOLOv7 algorithm in this paper shows significant improvements in detection accuracy and speed.

The rest of the paper is organized as follows: Section II introduces the real-time target detector, YOLOv7. Section III describes the improved YOLOv7 algorithm in detail. In Section IV, the experimental dataset and parameter settings are presented, and the results of the ablation and comparison

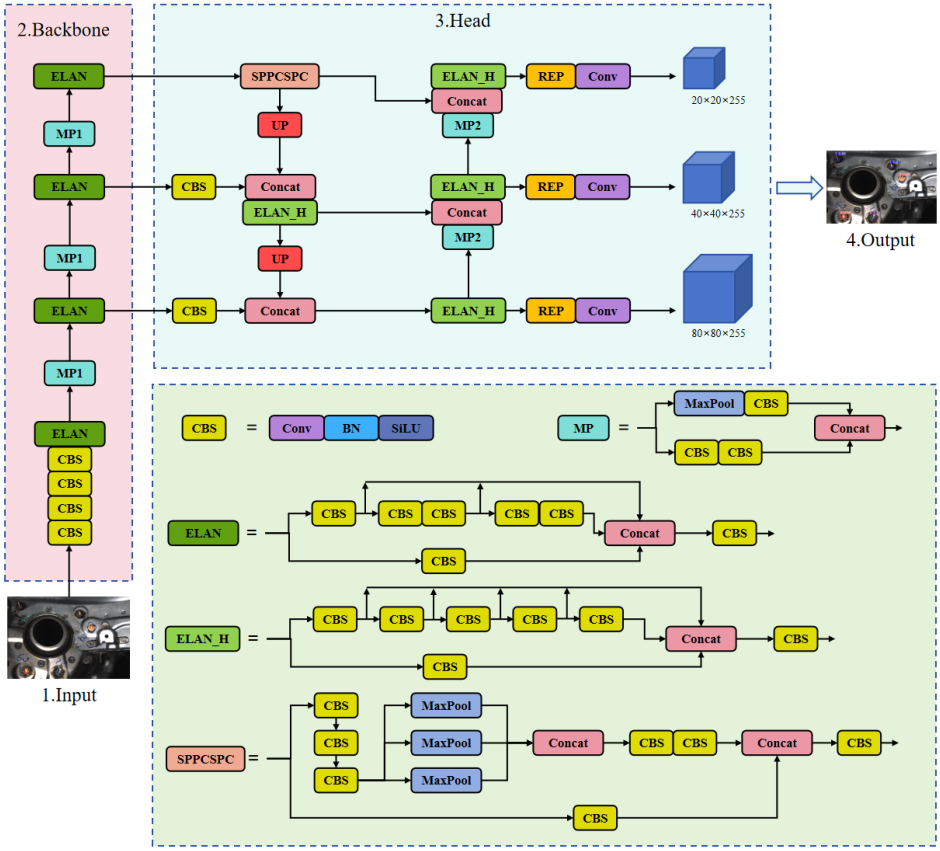


FIGURE 1. Structure of YOLOv7 model.

experiments are given. Section V summarizes the algorithm proposed in this paper and gives an outlook for future work.

II. RELATED WORK

The YOLO family of algorithms is a typical class of one-stage target detection algorithms released in 2015, and YOLOv7 is a more advanced version of the family with excellent speed and accuracy in the range of 5 FPS to 160 FPS [24]. In real-time image and video analysis, etc., YOLOv7 is able to quickly and accurately detect multiple objects and provide their position and category information, so in this paper, we choose the YOLOv7 model for stud detection [25]. The YOLOv7 network is mainly composed of the following parts, and the network structure of YOLOv7 is shown in Fig. 1.

(1) Input: in YOLOv7, following the Mosaic data enhancement method proposed by YOLOv4 [26], four pictures are randomly cropped and then spliced into one picture for training, enriching the dataset to improve the training efficiency while the training and inference costs remain unchanged. Following this, the training set is employed to calculate the optimal anchor points in an adaptive manner. The image is then resized to a standardized dimension prior to being transmitted to the backbone.

(2) Backbone: the main function of the backbone is to extract the feature information of the target. The YOLOv7 backbone network consists of the CBS module, the MP module, and the ELAN module. The convolutional layer, batch normalization layer, and activation function comprise the CBS module, which forms a standard convolutional block. The downsampling operation is carried out by the MP module, which comprises the maximal pooling layer and the CBS module and has upper and lower branches. The ELAN module is an efficient aggregation network that enhances the network's learning capability.

(3) Head: The Head part fuses the feature output from the Backbone and continues to extract features, generating a priori frames for classification prediction based on the strengthened features, which mainly include the SPPCSPC module, the MP module, the ELAN-H module, and the REP module. The SPPCSPS module introduces the CSPC structure on the basis of spatial pyramid pooling [27], so that it has one residual edge and is stacked with the feature layer after the maximum pooling process, reducing the amount of computation while increasing the accuracy. The REP module is divided into the training module and the inference module, which combines the different convolutional layers and batch normalization layers into one convolutional module when training the model and reparameterizes the parameters in the training module to the inference module when reasoning on the network, which accelerates the network reasoning under the condition of guaranteeing the performance of the model, reducing the model complexity but reducing the prediction

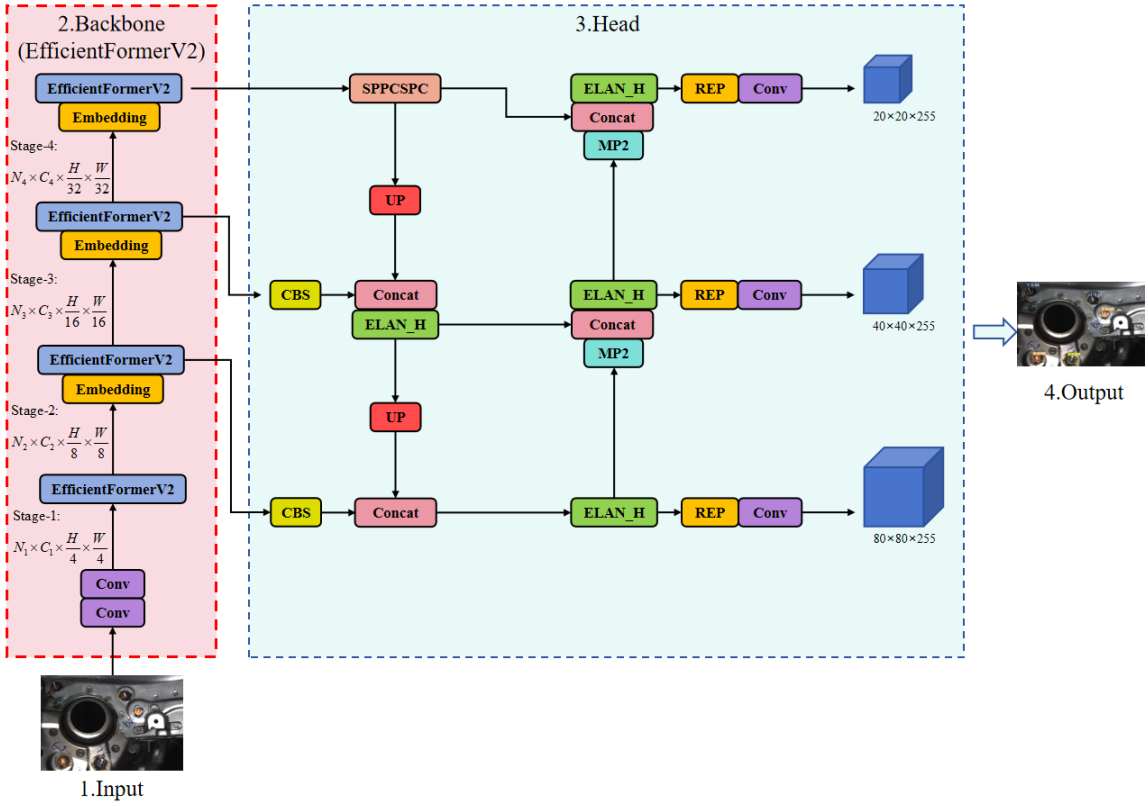


FIGURE 2. Improved YOLOv7 network structure.

performance. The complexity of the model is reduced, but the prediction performance is not degraded.

III. METHOD

This research presents an updated YOLOv7 detection network model that is optimized on the YOLOv7 backbone network and loss function in order to increase the speed and accuracy of body stud target recognition. Figure 2 depicts the fundamental architecture of the network. In this paper, the backbone network is EfficientFormerV2, the original network topology is replaced, and NWD (Normalized Gaussian Wsserstein Distance) is utilized in place of the loss function. The first image is fed into the backbone network for EfficientFormerV2 after data enhancement; the second and third blocks of EfficientFormerV2 are connected to the Head part's CBS module and Concat layer for input features, respectively; the final EfficientFormerV2 block of the backbone network is connected to the SPPCSPC feature fusion module to achieve an effective connection between the head part and the backbone network; since the training sample imbalance problem's impact on the bounding box regression process cannot be adequately taken into account by the CIoU loss used in YOLOv7, the NWD loss is proposed to lessen the negative effects of low-quality anchors on the bounding box regression process and increase the contribution of high-quality anchors.

A. EFFICIENTFORMERV2 BACKBONE NETWORK

In the original YOLOv7 backbone network, for the input image, four convolution operations are performed by the CBS module to obtain the underlying features, and then the MP and ELAN modules extract the fine-grained features. Such a structure will use a lot of repeated feature information, resulting in too many redundant operations, which increases the latency of the model. When inspecting welded studs, the high latency will make the time cost increase significantly. Therefore, without affecting the accuracy of the model, this paper proposes to replace the original YOLOv7 backbone network with a more efficient EfficientFormerV2 network structure. The network structure of EfficientFormerV2 is shown in Fig. 3.

EfficientFormerV2 [28] is an improvement on the previous version [29], which achieved higher performance with smaller models and faster reasoning. Two significant issues with the previously effective Vision Transformer (ViT) were its large model size and inability to be used on mobile devices. In order to generate efficient networks with low latency and small size based on the reference MobileNet, EfficientFormerV2 uses a fine-grained federated search method. It uses a depth-separable convolutional layer with the same kernel size in place of the Query mixer's average pooling layer, adhering to the standard ViT architecture and enhancing performance without adding latency. On the ImageNet

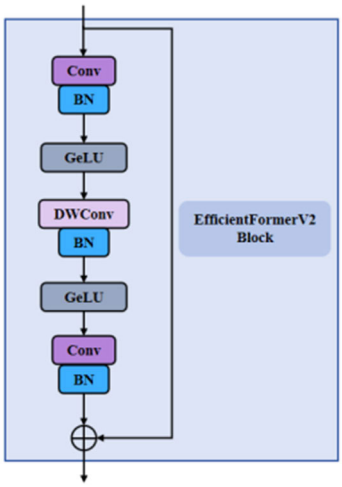


FIGURE 3. The EfficientFormerV2Block model structure diagram.

dataset validation set, the network outperforms MobileNetV2 by 4% while keeping the same latency and parameter counts.

Four stages of hierarchical design are used by EfficientFormerV2 to capture features at 1/4, 1/8, 1/16, and 1/32 of the input resolution, respectively. Like its predecessor, EfficientFormer, EfficientFormerV2 does not use inefficient non-overlapping patches; instead, it embeds the input picture from a small kernel convolutional stem. In the first two stages, just a uniform feed-forward network (FFN) is used to capture high-resolution local information; in the subsequent two stages, local FFNs and global MHSA blocks are used. In addition, building on previous versions, EfficientFormerV2 introduces fine-grained joint size and speed searches, resulting in extremely fast inference and smaller model sizes, outperforming previous techniques and becoming a powerful backbone for a variety of downstream tasks.

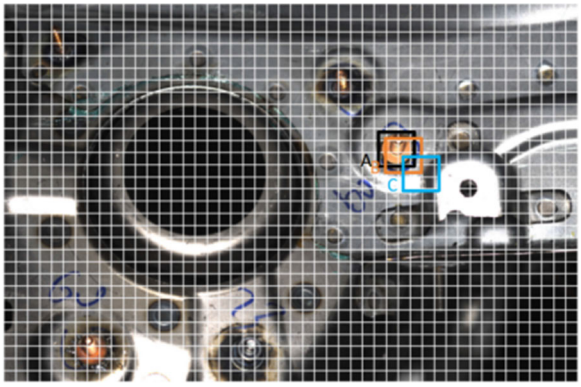
Replacing the backbone of YOLOv7 with EfficientFormerV2 effectively reduces memory access and latency. With this improvement, the inspection process is capable of being completed in less time by the model, resulting in higher frames per second (FPS).

B. NWD LOSS FUNCTION

The loss function of YOLOv7 consists of 3 parts: localization loss, confidence loss, and classification loss, and the computational representation is as follows:

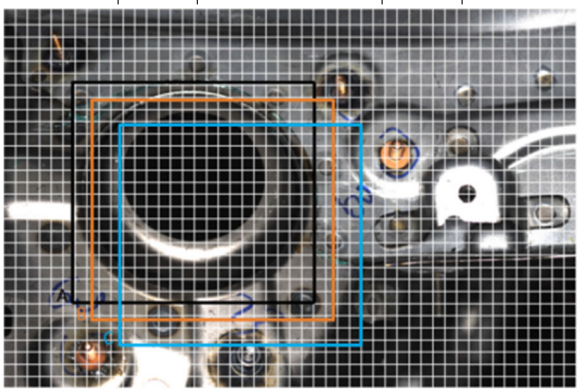
$$L_{loss} = \lambda_1 L_{obj} + \lambda_2 L_{cls} + \lambda_3 L_{box} \tag{1}$$

where L_{obj} denotes the confidence loss; L_{cls} denotes the classification loss; L_{box} denotes the localization loss; and $\lambda_1, \lambda_2, \lambda_3$ are the balance coefficients. Both confidence loss and classification loss are calculated using the binary cross entropy loss function; confidence loss is calculated for all samples, classification loss is calculated for positive samples only, and localization loss is calculated using CIoU (Complete-IoU) as the regression loss



(a) Stud target

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|} = 0.51 \rightarrow \text{IoU} = \frac{|A \cap C|}{|A \cup C|} = 0.09$$



(b) Other larger target

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|} = 0.87 \rightarrow \text{IoU} = \frac{|A \cap C|}{|A \cup C|} = 0.62$$

FIGURE 4. IoU analysis of stud targets vs. other larger targets.

function (L_{CIoU}), calculated as:

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

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{3}$$

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

where $p^2(b, b^{gt})$ represents the Euclidean distance between the two centroids and IoU indicates the consistency of the metric aspect ratio; The weight parameter is represented by α; the aspect ratio similarity between the real and predicted boxes is represented by v; the diagonal distance of the minimum enclosing box is represented by c; the width and height of the real box are represented by w^{gt} and h^{gt}, while the predicted box's width and height are represented by w and h. Bounding box regression is a key part of figuring out how well target localization worked, and the original YOLOv7 network's CIoU loss didn't take into account the issue of training sample imbalance. High-quality anchors exhibit a high IoU with the real box, meaning they closely align with the target. Conversely, a small IoU of real frames paired

with low-quality anchors suggests a poor match with the target. Because of the anchor base's limitations, only a small percentage of the many anchors generated are high-quality enough to match the target object, leading to a significant imbalance between positive and negative samples that interferes with the model's training process.

As demonstrated in Fig. 4(a), a slight positional deviation between the pre-tested frame and the real frame causes the IoU value to significantly decrease (from 0.51 to 0.09) when the body studs are actually detected because the stud target has a small area share in the entire image; for the normal-sized target, as demonstrated in Fig. 4(b), the positional deviation only causes a minor modification to the IoU (from 0.87 to 0.62). It is evident that the IoU is not suitable for body stud detection and is highly sensitive to the positional deviation of small targets. Instead of using the IoU, the Normalized Gaussian Wasserstein Distance (NWD) is used in this work to assess the similarity between the predicted frame and the real frame using a two-dimensional Gaussian distribution.

Since the actual stud target is not precisely rectangular, there will be some background pixels in the bounding box. The foreground pixel studs are mostly found in the center of the bounding box, whereas the background pixels are scattered around the borders. In order to improve the separation between the foreground and background pixels, a two-dimensional Gaussian distribution describes the bounding box. The stud pixels in the bounding box's center are given the greatest weight, and the weight value decreases as one moves toward the boundary. The two-dimensional Gaussian distribution $N(\mu, \sum)$ for the horizontal bounding box $R = (c_x, c_y, w, h)$ (where (c_x, c_y) represents the coordinates of the center point and w and h stand for the bounding box's width and height, respectively) is calculated as follows:

$$\mu = \begin{bmatrix} c_x \\ c_y \end{bmatrix} \tag{5}$$

$$\sum = \begin{bmatrix} \frac{w^2}{4} & 0 \\ 0 & \frac{h^2}{4} \end{bmatrix} \tag{6}$$

The Wasserstein distance can still be used to determine the degree of similarity between the real and predicted frames, despite the fact that the values of KL dispersion (Kullback-Leibler divergence) and JS dispersion (Jensen-Shannon divergence) are meaningless and constant, respectively, due to the absence or negligible overlap area between the two. Hence, the Gaussian distribution distance between the predicted frame $R_a = (c_{xa}, c_{ya}, w_a, h_a)$ and the actual frame $R_b = (c_{xb}, c_{yb}, w_b, h_b)$ is computed utilizing the Wasserstein distance:

$$W_2^2(N_a,N_b) = \left\| \left(\left[c_{xa}, c_{ya}, \frac{w_a}{2}, \frac{h_a}{2} \right]^{\mathrm{T}}, \left[c_{xb}, c_{yb}, \frac{w_b}{2}, \frac{h_b}{2} \right]^{\mathrm{T}} \right) \right\|_2^2 \tag{7}$$

where $W_2^2(N_a, N_b)$ is the Gaussian distribution distance between the prediction frame and the real one; (c_{xa}, c_{ya}) is the coordinates of the center point of the prediction frame, and w_a and h_a are the width and height of the prediction frame, respectively; (c_{xb}, c_{yb}) is the coordinates of the center point of the real frame, and w_b and h_b are the width and height of the real frame, respectively. The Gaussian distribution distance is normalized in exponential form and expressed as follows, given that it is a distance metric and not a similarity metric:

$$N_{NWD}(N_a, N_b) = \exp\left(-\frac{\sqrt{W_2^2(N_a, N_b)}}{C} \right) \tag{8}$$

where C is a constant related to the dataset. In addition, CIoU computes the aspect ratio of the two bounding boxes by taking into account the distance between the centroids of the overlapping regions. However, the aspect ratio has an impact on the loss function in situations where the predicted box and the actual box do not overlap or overlap completely in the context of stud target detection. As the NWD is more suitable for the measurement of small target studs and is more sensitive to changes caused by positional deviation, it can also reflect the similarity between the predicted and actual frames. As a result, the NWD is utilized as the loss function (L_{NWD}) in this paper. The expression for NWD is:

$$L_{NWD} = 1 - N_{NWD}(N_p, N_g) \tag{9}$$

where N_p and N_g denote the Gaussian distribution of the predicted and true frames.

C. EVALUATION METRICS

The experiments in this paper use several metrics to measure the model's detection performance, including Precision (P), Recall (R), mean Average Precision (mAP), and Frames Per Second (FPS) for all classes. P is the proportion of samples predicted by the model to be positive categories that are actually positive, i.e., how many of the samples predicted to be positive are actually positive samples. TP (True Positives) refers to the number of samples correctly predicted by the model to be positive, i.e., the number of positive samples that are correctly identified. FP (False Positives) refers to the number of samples incorrectly predicted as positive by the model for samples of negative categories, i.e., the number of FN (False Negatives) refers to the number of samples that the model incorrectly predicts as positive classes, i.e., the number of positive samples that are incorrectly identified as negative samples. TN (True Negatives) refers to the number of samples that the model correctly predicts as negative classes, i.e., the number of negative samples that are correctly identified. FP and FN are related to the probability of wrong and missed tests. P is calculated as:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{10}$$

R is the proportion of all actual positive samples that are correctly identified, indicating the ability of the model to detect true positive samples. R is calculated as:

$$R = \frac{TP}{TP + FN} \times 100\% \tag{11}$$

Precision and Recall are often contradictory performance metrics, where the higher the value of one, the lower the other, so it is necessary to combine these two evaluation metrics in conjunction with Average Precision (*AP*) and assess the performance of the model. In calculating mAP, the Average Precision (*AP*) of each category needs to be calculated first, which represents the average value of detection accuracy for that category in the dataset. Then the *AP* values of different categories are averaged to obtain *mAP*. The calculation process of *AP* and *mAP* is shown in Eqs. (12) and (13), respectively:

$$AP = \int_0^1 P(R)dR \tag{12}$$

$$mAP = \frac{\sum_{i=0}^{n} AP(i)}{n} \tag{13}$$

where n denotes the number of categories to be detected in the dataset. In this experiment, $n = 5$, i.e., 5 different models of welded studs. The *FPS* frame rate is the number of images that can be predicted by the network per second. The larger the *FPS*, the faster the network's inference is. The higher the FPS of the model, the better it can meet real-time detection needs. The *FPS* calculation process is shown in Equation (14):

$$FPS = \frac{1000}{t} \tag{14}$$

where t denotes the time in milliseconds required for the model to infer the image. When measuring the accuracy of target recognition, we need to set a threshold for the intersection ratio between the prediction frame and the ground truth. Only when the intersection area of the two exceeds the specified threshold is it recognized as correctly identified. This ratio is measured by the change in IoU values at different thresholds, and the recognition accuracy for each category is presented as *mAP* values. In evaluating the model's detection performance for stud targets, we used *mAP0.5* as a metric for average accuracy, setting the threshold at 0.5, and also examined the *mAP0.5:0.95* values, which gradually increased at intervals of 0.05 in the range of 0.5 to 0.95.

IV. RESULTS AND DISCUSSIONS

A. DATASET

The dataset used in this study was taken at the site of a domestic automobile manufacturing workshop, and the object of the study is the body-in-white welded studs. Because in the actual inspection task, different parts of the body are made of different materials, and the light reflected from the metal surface has different energies, which can lead to the phenomenon of light and dark, and there are some reasons such as blocking the light from the body parts, it is necessary to complete the inspection task under different lighting conditions, so we design two shooting modes: dark-light and bright-light environments. After removing the duplicated and blurred images, the stud image data set consists of 1397 images, including 650 images in the dark light environment and 747 images in the bright light environment. Labelimg software was employed to manually label the stud dataset. The labeling box was selected to be the stud's tiniest outer rectangle, and the labeling information file generated after labeling was of xml type, storing the file name of the stud image, the position information of the four corners of the rectangular box in the labeling area, and the type of labeling information. The training set, test set, and validation set are divided in the ratio of 6:2:2 for model training, testing, and validation.

B. EXPERIMENTAL CONDITION

The experiments in this paper use the Windows 10 operating system and the Pytorch deep learning framework to train and test the models. The software environment is CUDA 11.3, CUDNN 8.2, and Python 3.8. The CPU used for training the dataset is 13th generation Intel(R) Core (TM) i5-13600K(F) 3.50 GHz 32 G, and the GPU is NVIDIA GeForce RTX 4070. In addition, the batch size in the optimizer, the learning rate, and the epoch number are set to 16, 0.001, and 300, respectively, and the image resolution size is set to 640 × 640.

C. EXPERIMENTAL RESULTS

To verify that the enhancements made to our model for body implant target detection have improved in terms of precision and acceleration, we conducted comparative experiments on different stud detection models. We used Precision, Recall, mAP0.5, mAP0.5:0.95, and FPS values as metrics to evaluate the performance of our models. We first compared the initial Yolov7 model with models incorporating the EfficientFormerv2 and NWD loss functions, respectively, and Table 1 displays the outcomes of the experiments.

The data presented in Table 1 after replacing EfficientFormerV2 with the backbone network of Yolov7, the PFS is increased from the initial 96.1 to 144.9, and the detection speed of the image is increased by 50.8%. However, the mAP0.5 is slightly decreased by 0.1%, and the effect of the accuracy is not obvious.

In order to further improve the accuracy of the model, we replaced the loss function of the original Yolov7 model with NWD, which is more suitable for measuring small objects, and this replacement resulted in a 0.4% increase in mAP0.5. Therefore, we can conclude that replacing EfficientFormerV2 alone slightly reduces the accuracy of target detection but significantly improves the detection speed; replacing NWD alone improves the detection accuracy but is still limited in the detection speed.

In order to make progress in the performance of the model in both speed and accuracy, we use EfficientFormerV2 as the backbone and optimize the loss function part of the network using NWD. In order to verify the trend of the accuracy change of this paper's algorithm during the iteration process, the loss curve of 300 epochs of iteration of the model testing process as well as the mAP0.5 curve were compared with the

TABLE 1. Target detection model results.

Experiments	P(%)	R(%)	$mAP0.5$(%)	$mAP0.5:0.95$(%)	FPS
YOLOv7	90.8	94.6	94.6	63.7	96.1
YOLOv7+EfficientFormerV2	93.6	92.0	94.5	65.3	144.9
YOLOv7+NWD	92.5	93.5	95.0	64.7	97.1
Ours	**93.9**	93.7	**95.2**	**65.4**	**147.1**

TABLE 2. *mAP0.5* for five body stud.

Experiments	$mAP0.5$(%)			
	YOLOv7	YOLOv7+EfficientFormerV2	YOLOv7+NWD	**Ours**
1	95.9	95.7	96.7	**96.9**
2	94.4	94.5	94.5	**94.5**
3	92.3	91.6	92.6	**92.9**
4	97.9	97.7	98.5	**98.9**
5	92.5	92.9	92.7	**93**
average	94.6	94.5	95.0	**95.2**

TABLE 3. Experimental results of different attention modules with EfficientFormerV2 backbone.

Experiments	P(%)	R(%)	$mAP0.5$(%)	$mAP0.5:0.95$(%)	FPS
EfficientFormerV2+CIoU	93.6	92.0	94.5	65.3	144.9
EfficientFormerV2+EIoU	93.2	90.6	93.1	63.9	142.8
EfficientFormerV2+WIoU	88.5	93.8	94.2	62.8	142.8
EfficientFormerV2+SIoU	90.7	94.3	94.0	62.1	140.8
Ours	**93.9**	93.7	**95.2**	**65.4**	**147.1**

TABLE 4. Comparison of our model with traditional welding studs target detection model.

Experiments	P(%)	R(%)	$mAP0.5$(%)	$mAP0.5:0.95$(%)	FPS
FasterRCNN	83.6	95.9	89.7	57.2	12.0
YOLOv5s	90.6	91.7	93.9	60.4	98.7
YOLOv7	90.8	94.6	94.6	63.7	96.1
YOLOv8	90.8	93.6	94.0	61.9	243.9
Ours	**93.9**	93.7	**95.2**	**65.4**	147.1

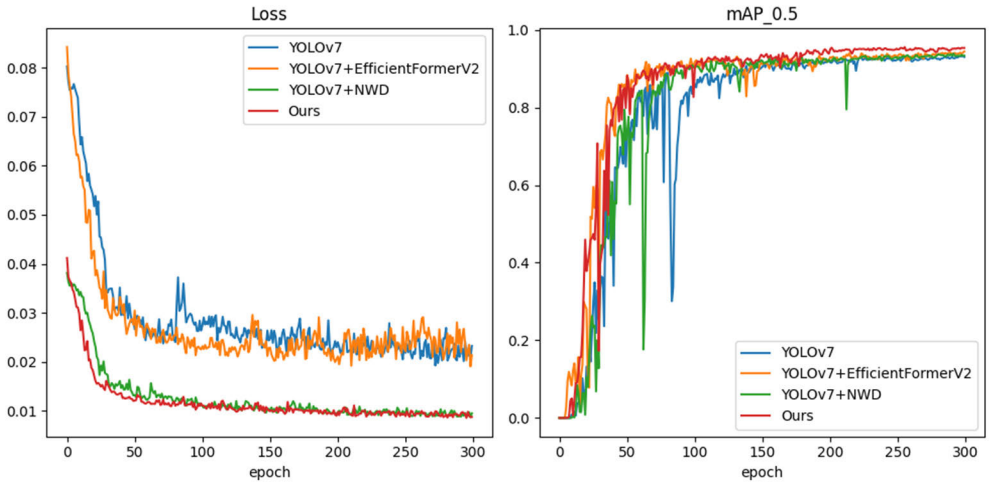


FIGURE 5. Performance of different models in training.

other 3 algorithms, as shown in Fig 5. The indicators gradually stabilized after 40 epochs, and the dataset converged faster, which illustrated that the parameters of the target detection model in this paper were set reasonably. Compared with the other three algorithms, the curve fluctuation during the training process of this paper's method is small, the

FIGURE 6. Experimental results of body stud target detection. (a) dark light condition; (b) bright light condition.

model fitting effect is better, and the method shows better performance in body stud target detection.

Our model outperforms the YOLOv7 and YOLOv7+ EfficientFormerV2 networks by 3.1% and 0.3%, respectively, in terms of detection accuracy, with a P-value of 93.7%. The R-value is 93.7%, greater than the 1.7% and 0.2%, respectively, of the YOLOv7+EfficientFormerV2 and YOLOv7+NWD networks. The percentages of mAP0.5 and mAP0.5:0.95 were 95.2% and 65.4%, respectively, higher than YOLOv7 by 0.6% and 1.7%. The model in this study delivers an FPS improvement of 144.9 in terms of detection speed, which is higher than the 96.1 FPS improvement of the

original YOLOv7 model. The experimental results show that the method in this paper outperforms YOLOv7 in terms of both accuracy and speed.

The corresponding mAP0.5 for five different body studs is shown in Table 2. The experimental results show that, compared with YOLOv7, our model improves the mAP0.5 values by 1%, 0.1%, 0.6%, 1%, and 0.5% for stud types 1–5, respectively. At a threshold of 0.5 for the IoU, our model has an superiority in target detection for these five types of studs, which signifies an enhanced target detection performance for these body studs within the specified IoU threshold.

To evaluate the effectiveness of various loss functions, in this paper, while keeping the backbone network of EfficientFormerV2 unchanged, the loss function CIOU of the original YOLOv7 is replaced with EIoU, GIoU, SIoU, and NWD, respectively, and Table 3 shows the performance of the above-mentioned five loss functions applied to YOLOv7+EfficientFormerV2.

An analysis of the performance of the five loss functions shows that compared to CIoU, P, R, mAP0.5, and mAP0.5:0.95 using the NWD loss function model are 0.3%, 1.7%, 0.7%, 0.1% higher, and the FPS is 2.2 higher; compared to EIoU, P, R, and mAP0.5, mAP0.5 using the NWD loss function model are 0.7%, 3.1%, 2.1%, and 1.5% higher, and the FPS is 4.3 higher; and compared to WIoU, P, R, mAP0.5, and mAP0.5 using the NWD loss function model are 4.3 higher: 0.95 are 0.7%, 3.1%, 2.1%, and 1.5% higher, respectively, and FPS is 4.3 higher; compared to WIoU, P, mAP0.5, and mAP0.5:0.95 are 5.4%, 1.0%, and 2.6% higher, respectively, using the NWD loss function model, with P being 0.1% lower, and FPS being 4.3 higher; and compared to SIOU, P, R, mAP0.5, and mAP0.5:0.95 are 5.4%, 1.0%, and 2.6% higher, respectively, and P is 4.3 lower; and compared to SIOU, the P, mAP0.5, and mAP0.5:0.95 are 3.2%, 1.2%, and 3.3% higher, respectively, with a lower recall of 0.6% and a higher FPS of 6.3. From the above analysis, it is clear that the comprehensive advantages of model training using the NWD loss function are more obvious, with the highest detection precision and the fastest detection speed.

The comparison results of our proposed network with the traditional target recognition network for body stud detection are shown in Table 4. Our model outperforms the other listed algorithms in terms of P, mAP0.5, and mAP0.5:0.95, showing excellent model quality. As far as mAP0.5 is concerned, our method improves the performance over Faster-RCNN by 5.5%. In addition, it outperforms YOLOv5s, YOLOv7, and YOLOv8 of the YOLO family by 1.3%, 0.6%, and 1.2%, respectively. The results show that our method shows better performance than the other algorithms in the table in terms of mAP0.5. In terms of mAP0.5:0.95, it outperforms Faster-RCNN by 8.2% and outperforms YOLOv5s, YOLOv7, and YOLOv8 of the YOLO series by 5.0%, 1.7%, and 3.5%, respectively. In addition, the model proposed in this paper performs well in terms of detection speed, with an FPS value of 147.1 frames per second, which is a big improvement compared with the original YOLOv7 model; however, there is still a gap compared with the current state-of-the-art YOLOv8.

To confirm the method's capacity for generalization as shown in this paper, we conducted a comparison experiment using YOLOv7 and our model on some randomly selected images in the body stud dataset, and the outcomes of the experiment are illustrated in Fig. 6.

The outcomes of the experiment demonstrate that our model as well as the YOLOv7 model can accurately recognize five different types of studs under different light intensities. However, in darker environments, YOLOv7 has a misdetection phenomenon, as shown by the red wireframe in Fig. 6(a), identifying other parts of the bodywork as the No. 5 stud, which does not occur in our model, and it can be observed that our proposed improved method improves the confidence level of the stud identification more substantially, which effectively improves the performance of the target detection, identification, and classification of the bodywork studs.

V. CONCLUSION

To resolve the issues pertaining to the slow inference speed and the low target detection accuracy in the automatic body stud detection environment in the current vehicle manufacturing process, an improved YOLOv7 target detection algorithm is proposed. We replace the original YOLOv7 backbone with the EfficientFormerV2 model with a lightweight network structure to speed up the detection of features by the network. We replace the original loss function with NWD, which allows the network to get additional accurate feature details for smaller targets. Ablation experiments were conducted on the automobile body welded stud dataset. With an average detection accuracy of 95.2% on the test set, the enhanced YOLOv7 algorithm obtains a 0.6% improvement, according to the experimental results. The improved YOLOv7 model reaches 147.1 f/s in real-life scenarios, which meets the accuracy and real-time requirements of automated online detection. Compared with other mainstream algorithms, the improved YOLOv7 algorithm in this paper is more suitable to be applied to automated online detection tasks with effectiveness and superiority.

In future work, we will apply the improved YOLOv7 algorithm to the task of automatic body stud detection and position estimation to verify the performance of the algorithm in this task.

REFERENCES

[1] W. Nishikawa, "The principle and application field of stud welding," *Weld. Int.*, vol. 17, no. 9, pp. 699–705, 2003, doi: 10.2207/qjjws1943.71.575.
[2] E. P. Patrick, J. R. Auhl, and T. S. Sun, "Understanding the process mechanisms is key to reliable resistance spot welding aluminum auto body components," *SAE Trans.*, vol. 93, no. 2, pp. 435–448, Feb. 1984, doi: 10.4271/840291.
[3] S. Ramasamy, "Drawn arc aluminum stud welding for automotive applications," *JOM*, vol. 54, no. 8, pp. 44–46, Aug. 2002, doi: 10.1007/BF02711866.
[4] B. Wu, F. Zhang, and T. Xue, "Monocular-vision-based method for online measurement of pose parameters of weld stud," *Measurement*, vol. 61, pp. 263–269, Feb. 2015, doi: 10.1016/j.measurement.2014.10.041.

[5] M. L. Smith, L. N. Smith, and M. F. Hansen, "The quiet revolution in machine vision—A state-of-the-art survey paper, including historical review, perspectives, and future directions," *Comput. Ind.*, vol. 130, Sep. 2021, Art. no. 103472, doi: 10.1016/j.compind.2021.103472.

[6] Q. Song, X. Xiao, and H. Jiang, "The automatic recognition of large ball valve sealing bolt based on digital image," *Int. J. Signal Process., Image Process. Pattern Recognit.*, vol. 8, no. 6, pp. 311–320, Jun. 2015.

[7] B. Gu, C. Wen, X. Liu, Y. Hou, Y. Hu, and H. Su, "Improved YOLOv7-tiny complex environment citrus detection based on lightweighting," *Agronomy*, vol. 13, no. 11, p. 2667, Oct. 2023. [Online]. Available: https://www.mdpi.com/2073-4395/13/11/2667

[8] J. Hu, C. Fan, Z. Wang, J. Ruan, and S. Wu, "Fruit detection and counting in apple orchards based on improved YOLOv7 and multi-object tracking methods," *Sensors*, vol. 23, no. 13, p. 5903, Jun. 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/13/5903

[9] H. Yang, Y. Liu, S. Wang, H. Qu, N. Li, J. Wu, Y. Yan, H. Zhang, J. Wang, and J. Qiu, "Improved apple fruit target recognition method based on YOLOv7 model," *Agriculture*, vol. 13, no. 7, p. 1278, Jun. 2023, doi: 10.3390/agriculture13071278.

[10] Y. Zhang, Y. Sun, Z. Wang, and Y. Jiang, "YOLOv7-RAR for urban vehicle detection," *Sensors*, vol. 23, no. 4, p. 1801, Feb. 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/4/1801

[11] Y. Liu and X. Wang, "SAR ship detection based on improved YOLOv7-tiny," presented at the *Proc. IEEE 8th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2022, pp. 2166–2170.

[12] W. Wu, X. Li, Z. Hu, and X. Liu, "Ship detection and recognition based on improved YOLOv7," *Comput., Mater. Continua*, vol. 76, no. 1, pp. 489–498, 2023, doi: 10.32604/cmc.2023.039929.

[13] B. Chen and Z. Dang, "Fast PCB defect detection method based on FasterNet backbone network and CBAM attention mechanism integrated with feature fusion module in improved YOLOv7," *IEEE Access*, vol. 11, pp. 95092–95103, 2023, doi: 10.1109/access.2023.3311260.

[14] G. Wei, F. Wan, W. Zhou, C. Xu, Z. Ye, W. Liu, G. Lei, and L. Xu, "BFD-YOLO: A YOLOv7-based detection method for building Façade defects," *Electronics*, vol. 12, no. 17, p. 3612, Aug. 2023, doi: 10.3390/electronics12173612.

[15] Y. Yang and H. Kang, "An enhanced detection method of PCB defect based on improved YOLOv7," *Electronics*, vol. 12, no. 9, p. 2120, May 2023, doi: 10.3390/electronics12092120.

[16] B. Peirui, W. Rui, L. Qingyi, H. Chao, D. Hongxuan, X. Mengyu, and F. Yingxia, "DS-YOLOv5: A real-time detection and recognition model for helmet wearing," *Chin. J. Eng.*, vol. 45, no. 12, pp. 2108–2117, Dec. 2023, doi: 10.13374/j.issn2095-9389.2022.11.11.006.

[17] S. Liu, Y. Wang, Q. Yu, H. Liu, and Z. Peng, "CEAM-YOLOv7: Improved YOLOv7 based on channel expansion and attention mechanism for driver distraction behavior detection," *IEEE Access*, vol. 10, pp. 129116–129124, 2022, doi: 10.1109/access.2022.3228331.

[18] L. Liguang, "Research on missing detection of bolts and nuts in sheet metal based on computer vision," M.S. thesis, Harbin Inst. Technol., Harbin, China, 2019.

[19] J. Zhang, Z. Su, and Z. Xing, "An improved SSD and its application in train bolt detection," in *Proc. 4th Int. Conf. Electr. Inf. Technol. Rail Transp. (EITRT)*. Singapore: Springer, 2019, pp. 97–104.

[20] J. Yang, L. Xin, H. Huang, and Q. He, "An improved algorithm for the detection of fastening targets based on machine vision," *Comput. Model. Eng. Sci.*, vol. 128, no. 2, pp. 779–802, 2021, doi: 10.32604/cmes.2021.014993.

[21] W. Yuchen, F. Hailong, and L. Boqi, "Missing bolt detection system for steel structure canopy of high-speed railway passenger station based on YOLO algorithm," *Railway Comput. Appl.*, vol. 31, no. 6, pp. 1–5, Jun. 2022, doi: 10.3969/j.issn.1005-8451.2022.06.01.

[22] X. Zhang and G. Wang, "Stud pose detection based on photometric stereo and lightweight YOLOv4," *J. Artif. Intell. Technol.*, vol. 2, no. 1, pp. 32–37, Dec. 2021.

[23] X. Wang, M. Yang, S. Zheng, and Y. Mei, "Bolt detection and positioning system based on YOLOv5s-T and RGB-D camera," *Trans. Beijing Inst. Technol.*, vol. 42, no. 11, pp. 1159–1166, Nov. 2022, doi: 10.15918/j.tbit1001-0645.2021.339.

[24] S. Li, S. Wang, and P. Wang, "A small object detection algorithm for traffic signs based on improved YOLOv7," *Sensors*, vol. 23, no. 16, p. 7145, Aug. 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/16/7145

[25] X. Shi, D. Huang, W. Li, and X. Wang, "Application of remote sensing image processing for classification and recognition," in *Proc. IEEE 15th Int. Conf. Adv. INFOCOMM Technol. (ICAIT)*, Oct. 2023, pp. 1–13.

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

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. ECCV*. Cham, Switzerland: Springer, 2014, pp. 346–361.

[28] Y. Li, J. Hu, Y. Wen, G. Evangelidis, K. Salahi, Y. Wang, S. Tulyakov, and J. Ren, "Rethinking vision transformers for MobileNet size and speed," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 16889–16900.

[29] Y. Li, "EfficientFormer: Vision transformers at MobileNet speed," in *Proc. Adv. Neural Inf. Process.*, vol. 35, 2022, pp. 12934–12949.

HONG HUANG received the B.Eng. degree in vehicle engineering and instrument from Hunan University of Science and Technology, Xiangtan, China, in 2018, where he is currently pursuing the Ph.D. degree. His research interests include deep learning, machine vision, and image processing.

XIANGQIAN PENG received the B.S. degree in vehicle engineering from Wuhan University of Technology, in 2002, and the Ph.D. degree in mechatronics engineering from Huazhong University of Science and Technology, in 2008. He is currently a Senior Lecturer with Hunan University of Science and Technology. His research interests include visual guidance, machine vision, and image processing.

XIAOPING HU received the B.S. degree in automobile application engineering from Changsha Institute of Transportation, in 1987, and the M.S. degree in mechatronics engineering and the Ph.D. degree in mechanical manufacturing and automation from Wuhan University of Technology, in 2001 and 2010, respectively. He is currently a Secondary Professor and a Ph.D. Supervisor with Hunan University of Science and Technology. His research interests include robot control and robot vision and mechatronics technology.

WENCHU OU received the B.S. degree in measurement and control technology and instrumentation and the M.S. degree in precision instrumentation and mechanics from Sichuan University, in 2005 and 2008, respectively, and the Ph.D. degree in instrument science and technology from Shanghai Jiao Tong University, in 2016. He is currently a Senior Lecturer with Hunan University of Science and Technology. His research interests include ultrasonic actuation and intelligent detection.