YOLOv7-DSE: An Efficient Safety Equipment Detection Network

Jiaxin Ren, Wenhua Cui, Ye Tao, Tianwei Shi.

Abstract—**Safety equipment detection is an important application of object detection, receiving widespread attention in fields such as smart construction sites and video surveillance. Significant progress has been made in object detection due to the rapid development of deep learning. Multi-scale targets and complex scenes increase the likelihood of false positives and missed detections, which can affect the accuracy of the detection. To address this issue, this study proposes YOLOv7-DSE. It is a small complex target scene detection network based on the improved YOLOv7. Also, we have created a private dataset of safety equipment. First, we enhanced the ELAN and MP backbone networks. Backbone is replaced by ordinary convolution by the depthwise separable convolution. We enabled the backbone network to extract deeper image features without increasing the amount of parameters and computation. Simultaneously, the model incorporates the EIOU loss function to improve its convergence speed and positioning effect. Secondly, we propose a new ELAN-SPD structure in the head network. Based on the ELAN structure, a space-to-depth convolutional layer is added to fully downsample the feature map, preserving all learnable features. Our network model can better detect objects with significant size differences faced with complex scenes. YOLOv7-DSE achieved the mAP of 82.38%, surpassing the original YOLOv7 with 2.64%. The YOLOv7-DSE model has a minor size compared to the baseline model. Our improvement has reduced the model parameters by 22.4%.**

Index Terms—**space-to-depth convolution, YOLOv7, Object detection, EIOU**

I. Introduction

THE rapid growth of the power industry has been accompanied by numerous high-risk incidents and frequent safety operation accidents. A primary cause of these safety accidents is the improper use or failure to wear electrical safety equipment by workers. This may result from prolonged exposure to high-voltage environments or the disregard for personal safety measures. It is an urgent issue to enhance the inspection of electrical operation safety equipment used by workers. Traditionally, safety equipment inspection involves personnel supervision. Workshop supervisors take turns to conduct visual observations. This method is inefficient and costly. Subjective results are not conducive to identifying safety hazards. Consequently, there is an urgent need within the traditional electricity industry for a more efficient and lightweight algorithm to inspect operation safety equipment.

With the development of computer vision technology, object detection has become an important area of research in computer vision. The developmental process can be broadly categorized into the subsequent phases: hand-crafted feature-based object detection, which began from 2001 to 2012. The earlier object detection techniques involved a combination of machine learning algorithms, like Haar features[1], HOG features[2], SVM[3], and Adaboost[4], along with hand-crafted features. Those methods were effective during that time but lacked universality and exhibited weak performance while dealing with intricate scenarios. Object detection methods based on deep learning arose in 2012-2016. It all started with the application of deep neural networks in the 2012 ImageNet[5] competition, as demonstrated by AlexNet[6]. Since then, deep learning has garnered significant interest from researchers in the field of computer science. In 2014, Ross Girshick's RCNN[7] made the target detection method based on deep learning surpass the hand-crafted feature method for the first time. In 2015, Fast R-CNN further improved its speed while maintaining its accuracy. In the same year, the YOLO[8] algorithm was proposed to further improve detection speed and accuracy. These algorithms all use convolutional neural networks (CNN) to extract features from images, which are then combined with classifiers or regressors for object detection. Object detection algorithms based on a one-stage method have been used since 2016. One-stage method directly regresses the position and category of objects from images, without generating candidate boxes first. While these methods have faster detection speeds, their accuracy is relatively lower. Examples of mature one-stage methods include SSD[9], YOLOv2[10], RetinaNet[11], and CornerNet[12]. Object detection algorithms based on two-stage methods have been used since 2018. Two-stage object detection methods generate candidate anchor boxes first and then classify and regress the positions of objects within those candidate anchor boxes. Compared with one-stage methods, two-stage methods usually have higher accuracy but slower detection speed. Examples of well-known two-stage object detection methods are Faster R-CNN [13], Mask R-CNN [14], and Cascade R-CNN [15].

Manuscript received August 31,2023; revised May 15, 2024.

This work was supported by Joint Fund Project of the National Natural Science Foundation of China (U1908218) , the Natural Science Foundation project of Liaoning Province (2021-KF-12-06) , and the Department of Education of LiaoningProvince (LJKFZ20220197).

Jiaxin Ren is a postgraduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (e-mail: 873951356@qq.com).

Wen Hua Cui is a professor at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (corresponding author to provide e-mail: taibeijack@126.com).

Ye Tao is an associate professor at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (e-mail: taibeijack@163.com).

Tianwei Shi is an associate professor at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (e-mail: tianweiabbcc@163.com).

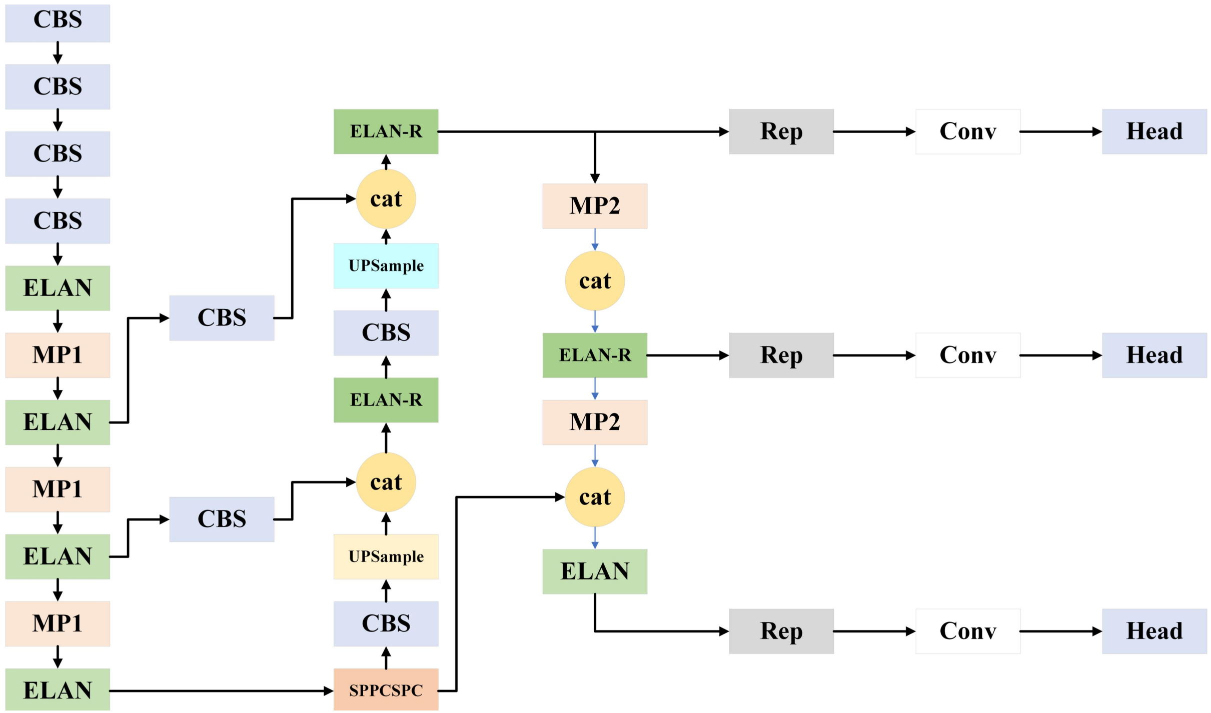


Fig. 1. YOLOv7 Network Structure

Deep learning-based object detection technology has seen extensive adoption in engineering construction in recent years. As a result, experts and scholars worldwide have shown a heightened interest in algorithms that detect safety equipment usage. Given that safety helmet detection algorithms are relatively mature. However, there is a scarcity of research on the detection of other safety equipment used during operations. This study aims to enhance the detection methods for worker safety equipment by leveraging safety helmet detection techniques. Therefore, it is necessary to investigate the current state of research on safety helmet detection algorithms.

Many researchers have attached the importance of the detection of safety helmets. Wen et al. [16] proposed a circle/arc detection algorithm based on the improved Hough transform in 2003 and applied it to detect safety helmets in ATM monitoring systems. Rubaiya et al. [17] proposed a human detection algorithm in 2016, which combines the frequency domain information of the image with the Histogram of Orientation for the detection of construction workers. Li et al. [18] proposed the ViBe background modeling algorithm in 2018. Pedestrians are located accurately and quickly by using a real-time human classification framework. The framework relies on the segmentation results of moving objects. Finally, head positioning, color space transformation, color feature recognition, and other methods are used to detect wearing helmets according to the pedestrian detection results. Wu et al. [19] enhanced the YOLOv3 network model in 2018. It effectively overcomes challenges such as varied helmet colors, partial obstruction, multiple objects, and low image resolutions.

After analyzing the research, specific areas need improvement in the algorithmic investigation of deep learning-based detection of safety equipment. (1) Building comprehensive datasets is a priority. Current detection datasets for safety operation equipment are scarce, with limited coverage and scenes. A larger and more diverse dataset is necessary to enhance algorithm robustness. Therefore, it is necessary to build a larger and diversified dataset to improve the robustness of the algorithm. (2) Additionally, multi-category safety equipment testing should be considered. Workers in electrical environments require various equipment, including insulating gloves and safety operating bars. (3) Finally, the optimization of the algorithm cannot be overlooked. The safety equipment detection algorithm faces challenges. Identifying small targets and lightweight detecting models necessitate further optimization. Our paper proposes a safety operation equipment detection algorithm. Compared with the baseline network, our network has better performance on detection accuracy and model size. The main contributions of this paper are in four aspects. (1) We propose a detection method for workers' electrical safety equipment based on an improved YOLOv7 network model. The depthwise separable convolutions are used to replace the convolutions in the MP1 and ELAN of the backbone. And SPD is used to replace the convolution of ELAN on the head. (2) We have built a set of private data sets to make the network model achieve better performance. (3) We introduce the EIOU calculation in the loss function. It uses the minimum difference between the width and height of the target anchor box. It improves the localization effect of the model's object detection box in complex scenes. (4) We have done relevant comparative experiments and ablation experiments to verify the performance improvement of the network model.

The remainder of this paper is organized as follows: Section I describes the current national and international related work on safety equipment detection. Section II introduces the baseline YOLOv7 network. Section III describes our proposed DSConv, SPD, and EIOU in detail. Section IV describes our experimental procedure and discusses the experimental results in detail. Conclusions and introspection are presented in Section V.

The opening section explores the context of network models. The evolution of computer vision and advances in object detection techniques are involved in Section I. Furthermore, the section reviews the historical progress of safety equipment and summarizes the limitations of contemporary research. Subsequently, the section outlines the contributions and related works of this paper. Finally, the section discusses the organization of future sections.

II. The Model Structure Of YOLOv7

Wang et al. [20] proposed YOLOv7. It is one of the well-known object detectors that has both accuracy and speed. The network structure is depicted in Figure 1. It comprises three main components: the image input component, the backbone feature extraction component, and the head feature fusion component.

A. Input

The input part of the YOLOv7 network model uses the following four tricks. Mosaic data enhancement uses four images for random scaling, clipping, and arrangement. It mainly solves the problem of small object detection. It enhances the dataset and amplifies the number of minor targets after random fragmentation and recombination. In turn, it enhances the robustness of the network model. Adaptive computing of anchor boxes utilizes a combination of genetic algorithms and K-means clustering to calculate the optimal frame. The predicted values are then used to calculate the best possible recall rate of the frame, achieving the purpose of random data augmentation. The self-adaptive anchor frame changes from the conventional approach. Instead of scaling the source image to a standardized size before feeding into the backbone network, it adjusts the anchor frame dynamically. It dynamically adjusts the minimum amount of black borders required based on the actual usage scenario. This eliminates the redundancy of information and improves the processing speed. YOLOv7's input reuses YOLOv5's preprocessing mode as a whole. We think that this preprocessing mode is mature in the field of target detection. Next, we analyze the backbone and head to find out what can be improved.

B. Backbone Network YOLOv7

The backbone feature extraction network part of the YOLOv7 algorithm model is composed of four CBS structures, ELAN structures, and MP1 structures. The convolution layer, batch normalization layer, and activation function layer are coupled together to form a CBS structure. The four CBS architectures have distinct strides. They utilize convolutions with a stride for feature extraction and convolutions with two strides for downsampling. The ELAN architecture separates the input feature matrix into two channels. One of the channels passes through a single CBS module. Another channel is combined with five additional CBS modules. Two channels will flow through a final CBS module. ELAN structure is an exceptionally efficient network architecture. It is capable of regulating the shortest and longest gradient paths. This capability enables the network model to learn more features and increases its robustness. MP1 structure performs sufficient downsampling of the image through two branches. One approach involves passing the data through a maximum pooling layer followed by a 1×1 convolution to adjust the number of channels. The other method uses a 1×1 convolution followed by a 3×3 convolution for downsampling. To achieve more image features without adding parameters to the network model, we need to modify the convolutional layer of the backbone network.

C. Head Network YOLOv7

The head part of the YOLOv7 algorithm model converges the feature maps extracted from the backbone feature network via SPPCSPC, CBS, MP2, and ELAN-R. The output terminal utilizes the RepVGG structure, producing prediction results in three distinct sizes. The SPPCSPC module divides the input feature matrix into two branches. One branch goes through three CBS modules that perform pooling and splicing operations at core dimensions of 5×5, 9×9, and 13×13, respectively. Upon traversing two CBS modules, the second branch proceeds to a CBS module dedicated to channel fusion. The ELAN-R module is highly similar to the ELAN module, differing only in the number of channels. The integration occurs in the channel fusion module following the completion of a single CBS module on the second branch. The MP-2 module resembles the MP-1 module but has a double number of channels. YOLOv7's head architecture bears similarity to the YOLOv5 within the family system model. However, it remains anchor-based, lacking decoupled heads for classification and prediction. Consequently, the unique fusion effect for detection is underemphasized. Thus, enhancing the head network is essential to improve small object detection efficacy.

This section delves into the strengths and weaknesses of the YOLOv7 network model. YOLO models are designed for speed. It is necessary to strengthen the capability of the backbone network for efficient feature extraction and refine the head network to enhance the accuracy of target detection. These are crucial aspects to be tackled in this study. Specifically, the network model will be optimized to better cater to smaller targets and as a result, enhance the detection efficiency of safety operation types of equipment for workers.

III. YOLOv7-DSE

The main structural improvements proposed for YOLOv7 are as follows: Depthwise separable convolutions are used to replace the standard convolutions in the MP1 and ELAN components of the backbone. The Head network introduces space-to-depth convolutions to enhance small object detection. The proposed detector's structure, YOLOv7-DSE, is illustrated in Fig 2.

A. Backbone Network MP1-DSC and ELAN-DSC

In our selected baseline network model, the backbone network is essential for extracting features from images. We have improved the backbone to reduce the volume of the model. We have configured the network application background to be in the power operation site. Consequently, our goal is to reduce the model's parameter count and computational complexity. As a result, we have enhanced the depthwise separable convolutional layer [21] within the

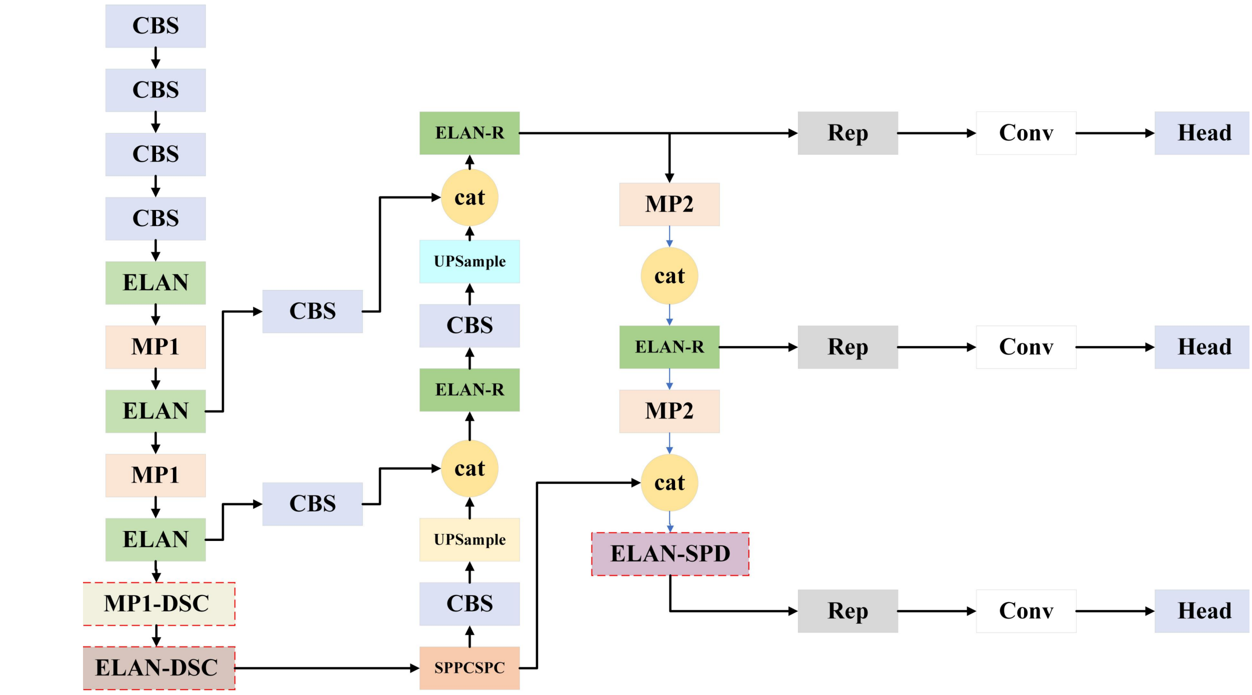


Fig. 2. YOLOv7-DSE Network Structure

backbone network.

Depthwise separable convolution is composed of two separate stages: depthwise and pointwise convolutions. Depthwise separable convolution extends the standard convolutions by applying them independently to each input channel. This stage is succeeded by pointwise convolution, which maps the depthwise convolution's output channels into a new channel space. The depthwise separable convolution process is illustrated in Fig. 3.

To showcase the enhanced performance, we must calculate and compare the parameters and computational costs of standard convolution against depthwise separable convolution. This enables us to reach well-informed conclusions. W denotes the width of the convolutional kernel, while H signifies its height. C_{in} represents the number of input channels, and C_{out} represents the number of output channels. The parameters P can be expressed as follows:

$$P = W \times H \times C_{in} \times C_{out} \tag{1}$$

W' represents the width of the image, H' represents the height of the image, and the calculation result C can be expressed as:

$$C = W \times H \times (W' - W + 1) \times (H' - H + 1) \times C_{in} \times C_{out} \tag{2}$$

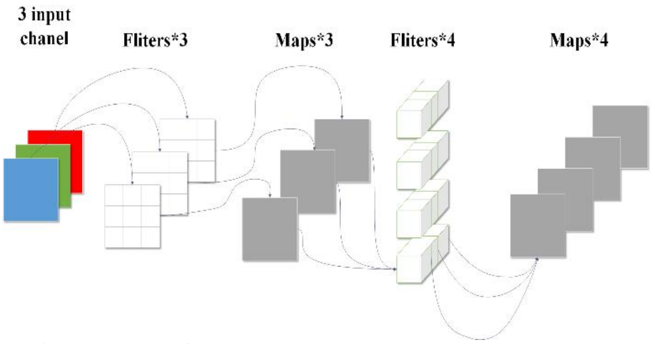


Fig. 3. Process of DSC

Taking a 5×5 image input with 3 channels as an example. We achieve a 3×3×4 feature map by regular convolution requires a 3×3×3×4 convolution kernel. The parameter quantity of its convolutional layer is:

$$P_{con} = 3 \times 3 \times 4 \times 3 = 108 \tag{3}$$

The computation quantity of its convolutional layer is:

$$C_{con} = 3 \times 3 \times (5 - 3 + 1) \times (5 - 3 + 1) \times 3 \times 4 = 972 \tag{4}$$

When applying depth-separable convolution, begin by completing the depthwise convolution process. Then, proceed to a filter of just a 3×3 convolutional kernel. Finally, use pointwise convolution to generate a feature map of 3×3×4, for which a 1×1×3×4 convolutional kernel size is necessary. The parameter quantity of the DSC convolutional layer is:

$$\begin{aligned} P_{dsc} &= P_{depthwise} + P_{pointwise} \\ &= 3 \times 3 \times (5 - 3 + 1) \times (5 - 3 + 1) \times 3 \times 4 = 972 \end{aligned} \tag{5}$$

The computation quantity of its convolutional layer is:

$$\begin{aligned} C_{dsc} &= C_{depthwise} + C_{pointwise} \\ &= 3 \times 3 \times (5 - 2) \times (5 - 2) \times 3 + 1 \times 1 \times 3 \times 3 \times 3 \times 4 = 351 \end{aligned} \tag{6}$$

Depthwise separable convolution uses fewer computations and parameters compared to standard convolutional layers for producing feature maps of equivalent size. Consequently, given the same computational budget and parameter constraints, depthwise separable convolution can be applied to deepen the neural network layers. We utilize it on the backbone network to extract more advanced image features, thereby enhancing the performance of the baseline network model. Section V presents ablation experiments that underscore the effectiveness of depthwise separable convolution in feature extraction.

B. Head Network ELAN-SPD

The main function of the network model's head component

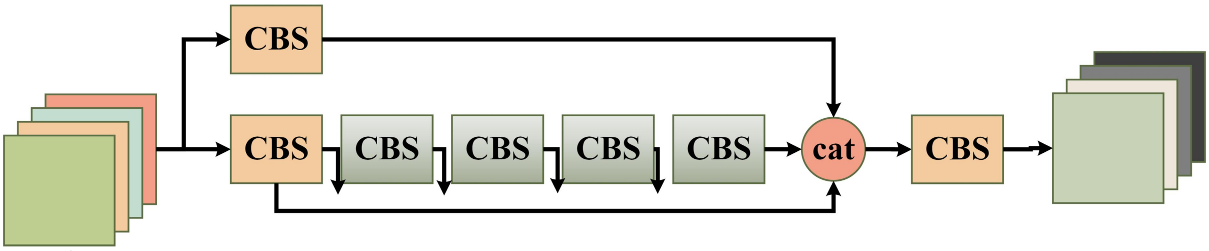


Fig. 4. Structure of ELAN

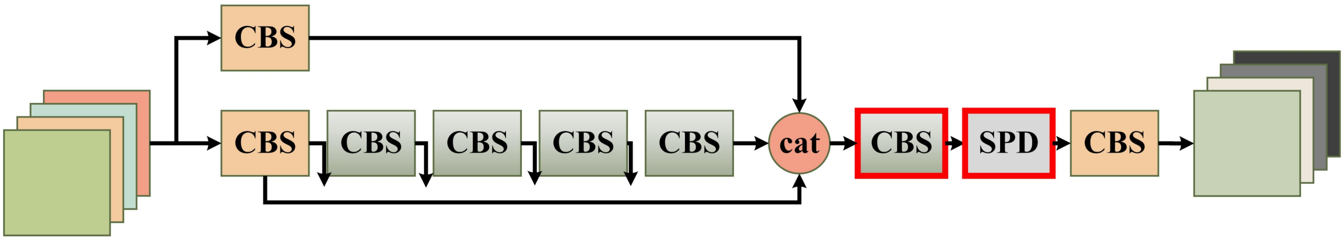


Fig. 5. Structure of ELAN-SPD

is to identify target categories and positions in feature maps generated by the backbone network. Traditional CNN networks generally perform well with high-resolution images and reasonably sized detection targets.However, this method often produces a large amount of redundant pixel information. Small targets usually provide limited information for the network model to learn and are characterized by lower resolutions. Detecting small targets continues to be an extremely challenging task. In many cases, small and large targets coexist in the same images. This coexistence can affect the network model's deep learning process, potentially leading to undetected small targets.

To address this challenge, we incorporated the SPD (space-to-depth) convolutional layer [22] into the head network. ELAN-SPD is an adaptation of the ELAN module in YOLOv7, enhanced by the addition of the SPD module to improve small object detection performance. The main differences between ELAN-SPD and ELAN lie in the CBS layer. The ELAN structure is shown in Fig. 4. And the improved module structure is demonstrated in Fig. 5. SPD convolution addresses the limitations of typical CNN models and enhances them. It effectively prevents information loss due to stride convolution and pooling operations. It ensures crucial data retention for the model to learn. Using convolution with one stride can retain more fine-grained features and prevent the feature information loss associated with larger convolution strides.

P represents the proportion of the original image, S refers to the sub-feature map, O denotes the original feature map, M stands for the intermediate feature map, A represents the length and width of the feature map, K_1 signifies the depth of the feature map, and K_2 represents the filter. The operational principle of SPD is as follows: The original image is partitioned into several sub-feature maps. If the original feature map $O(A, A, K_1)$ is uniformly divided based on component O(P, P), then (x/P × y/P) sub-feature maps are obtained. Downsampling the segmented sub-feature maps, when P=2, four sub feature maps $S_{0,0}$, $S_{1,0}$, $S_{0,1}$, $S_{1,1}$ can be obtained, and the size of each sub feature map is $(A/2, A/2, K_1)$. At the same time, the original feature map O underwent downsampling at a multiple of 2. The sub-feature maps $S_{0,0}$, $S_{1,0}$, $S_{0,1}$, $S_{1,1}$ downsampled and concatenated into the intermediate feature map M based on channel dimensions. The spatial dimension of M has been changed to 2 and the number of channels has been doubled.These feature map changes can be expressed by the following formulas as:

$$
\begin{aligned}
S_{0,0} &= O[0:A:P, 0:A:P] \\
S_{1,0} &= O[1:A:P, 0:A:P] \\
&\dots \\
S_{P-1,0} &= O[P-1:A:P, 0:A:P] \\
S_{0,1} &= O[0:A:P, 1:A:P] \\
S_{1,1} &= O[1:A:P, 1:A:P] \\
&\dots \\
S_{P-1,1} &= O[P-1:A:P, 1:A:P] \\
&\dots \\
S_{P-1,P-1} &= O[P-1:A:P, P-1:A:P]
\end{aligned}
\tag{7}
$$

After the SPD feature layer, a non-strided convolutional layer with K_2 filters is incorporated. The futher transformation of original feature is from $O(A, A, K_1)$ to $F(A/P, A/P, K_2)$. And the implementation of non-strided convolutional layer can retain maximum discriminative feature information and avoid the loss of feature information caused by convolution stride size. The process of SPD can be illustrated in Fig. 6.

We enhanced the ELAN of the YOLOv7 head network by implementing the SPD convolution module to create the ELAN-SPD. Targets with small resolutions are divided into multiple sub-feature maps, which are merged into intermediate feature maps. Discriminative information is isolated using the K_2 filter. Finally, feature information is extracted from the original image.

C. Loss Function EIOU

The baseline model in this paper employs the CIOU loss

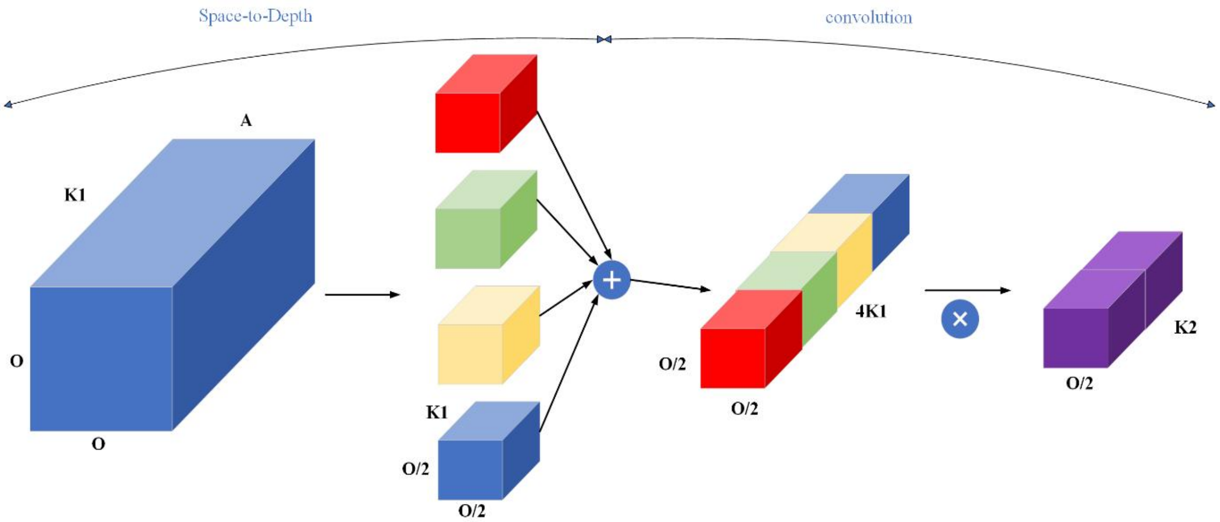


Fig. 6. YOLOv7-DSE Network Structure

function [23]. It takes into account three crucial geometric factors: the overlapping area, distance between center points, and transverse longitudinal ratio. Set the prediction box to P, the target box to P^{gt}, and c and c^{gt} represent the center points of the prediction box and the tagets box respectively.

E represents the Euclidean distance between two center points, and d represents the diagonal length of the minimum bounding box that can cover the prediction box and the target box. And w, w^{gt} represent the widths of the prediction and target boxes, res, while h and h^{gt} represent the heights of the prediction and target boxes, respectively. β represents the weight function, where v measures the difference in the aspect ratio. It can be represented by the following formul as:

$$v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \tag{8}$$

$$\beta = \frac{v}{(1 - IOU) + v} \tag{9}$$

CIOU loss function can be expressed by the following formula:

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

Prior experiments have shown that the CIOU loss function significantly improves convergence and detection rates over its predecessor. However, the definition of variable v, based solely on the aspect ratio difference, can occasionally hinder the effective optimization of model similarity. Consequently, we refined the baseline loss function by adapting it to the EIOU loss function [24], introducing an innovative loss computation method. The EIOU calculation method is shown in Fig. 7.

The loss function is comprised of three components: IOU loss, distance loss Ldis, and positioning loss Lasp. The values of the width and height of the smallest frame covering the predicted and target frames are denoted by hw and hc, respectively. The EIOU was calculated as:

$$\begin{aligned} L_{EIoU} &= L_{IoU} + L_{\text{dis}} + \mathrm{L}_{\text{asp}} \\ &= 1 - IoU + \frac{E^2(c, c^{gt})}{(w^c)^2 + (h^c)^2} + \frac{E^2(w, w^{gt})}{(w^c)^2} + \frac{E^2(h, h^{gt})}{(h^c)^2} \end{aligned} \tag{11}$$

This approach allows for the retention of the advantageous properties of the CIOU loss function while also allowing for the direct minimization of the disparity between the width and height of the predicted and intended bounding boxes through the application of EIOU loss. It can enhance the localization accuracy and convergence speed of anchor boxes in object detection. These effects enable model to percorm more accurately in complex scenriors of work space. And the detection performance can be improved.

IV. Experiments

This section provides an elaboration of the experimental dataset, and details the data analysis and data preprocessing procedures. Therefore it introduces com used evaluation metrics for object detection and conducts comparative experiments as well as ablation experiments.

A. *Data Sets*

In deep learning, the quality of our dataset often determines the effectiveness of our training. To guarantee successful training, we compiled a private dataset of representative images from diverse sources. It encompasses images from power worksites and internet. The dataset includes 2613 high-definition images, each with a pixel value of 4032×3024. We utilized LabelImg to annotate our target categories and location information for each image in the dataset.And we converted them into the XML format needed

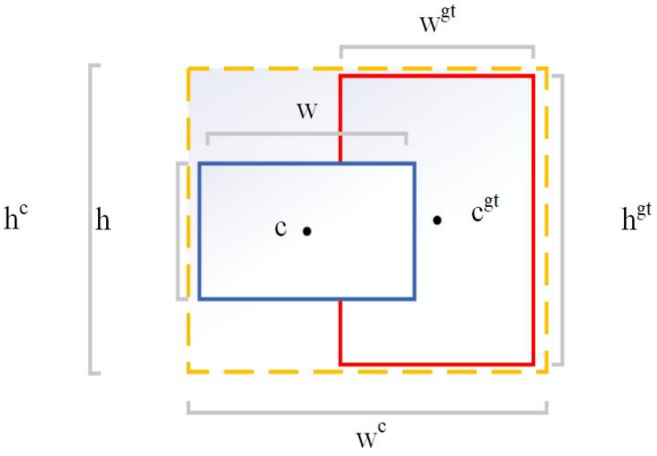
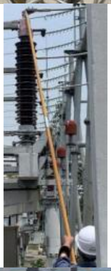


Fig. 7. EIOU Calculation Method

for training. Our dataset features six label types: Badge (monitoring armband), Person (all present), Glove (insulated gloves), Wrong Glove (non-insulated gloves), Operating Bar (insulated operating bar), and Power Checker (test pen). Table I displays examples of each dataset label.

TABLE I
LABEL SAMPLE DIAGRAM

Name of Label	Description	Image Style
Badge	Supervisors at the work site wear red armbands	
Glove	Insulating gloves worn by workers at the work site	
Wronggloves	Lack of insulated gloves	
Operating bar	Insulated operating bar	
Person	All present	
Power checker	High voltage testing pen	

Using original high-resolution images in the training set may hinder the detection of smaller targets. Therefore, this study preprocesses images by applying enhancements like rotation, flipping, cropping, and scaling. This not only boosts training efficiency, but also enhances small target detection and the model' s generalization capability.On the one hand, it improves our training efficiency.

B. Experimental Evaluation Factors

In deep learning, we applied the confusion matrix for evaluation using precision, recall, and mAP values were used as indicators of experimental evaluation. R denotes the recall rate, the percentage of correctly predicted samples by the model in the original set. It can be expressed as:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{12}$$

P represents accuracy, which refers to the proportion of correct samples to all samples in the original sample. It can be expressed as:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{13}$$

In the experiment, another parameter is needed to combine recall and accuracy, namely average accuracy mAP. To measure the performance of the network model, which can be expressed as:

$$mAP = \frac{1}{k}\sum_{k=1}^{n} AP_k X_i \tag{14}$$

C. Experimental Configuration

The computer operating system used in the experiment was Windows 10, with a CPU model of Intel (R) Core (TM) i7-10700F CPU @ 2.90GHz, a GPU model of Geforce RTX3070, 16GB of graphics memory, and 16GB of memory. The model is based on Python 1.8, using CUDA version 11.0 and Cudnn version 8.0.4. Utilizing pre-trained weights from YOLOv7 and the Adamax optimizer, we addressed data insufficiency and significantly improved the learning speed and accuracy of our network model. We configured the batch size to 16 and the learning rate to 0.001 to prevent model overfitting, closely tracking the loss value of the worker safety equipment detector' s bounding boxFigure 8 illustrates that after 100 epochs, the training and validation loss values of our detector converged.

We diligently monitored recall and accuracy to maintain optimal performance of the network model during training. Table II presents variations in recall and accuracy over various epochs, indicating that our model reached peak recall and precision after 100 epochs.

D. Comparison of Other Models

In the study, to showcase the improved performance of our network model, we conducted horizontal comparative experiments utilizing the same dataset and experimental settings. Compare the method proposed in this study with those proposed in seven other sources from the literature, including YOLOv4 by Alexey Bochkovskiy et al. [25], and Faster RCNN by Shaoqing Ren et al. [26]. RCNN by Ross

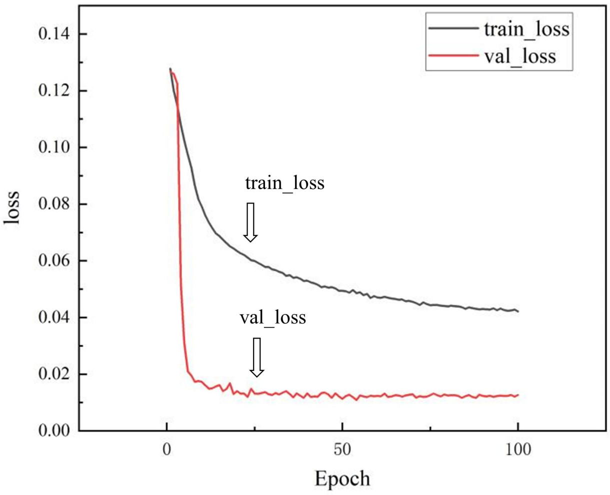


Fig. 8. Training and Validation Losses for YOLOv7-DSE

Girshick et al. [27], YOLOv5 by Glenn Jocher et al. CEAM-YOLOv7 by S Liu et al. [28], YOLOv7-RAR by Y Zhang [29], and YOLOv7 by Chien Yao Wang et al. [30]

Figure 9 illustrates that the YOLOv7 algorithm achieves a precision of 86.23% for detecting all categories when the confidence threshold is set to 0.5. Figure 10 demonstrates that the improved YOLOv7-DSE algorithm achieves a higher precision of 90.41% for detecting all categories at the same confidence threshold of 0.5. YOLOv7-DSE has improved the precision by 4.18%.

Figure 11 portrays that the YOLOv7 algorithm achieves a recall of 90.38% for detecting all categories when the confidence threshold is set to 0.5. Figure 12 demonstrates that the improved YOLOv7-DSE algorithm achieves higher recall of 93.72% for detecting all categories at the same confidence threshold of 0.5. YOLOv7-DSE has improved the recall by 3.34%.

TABLE II
TRAINING PROCESS OF YOLOV7-DSE

Epoch	Recall(%)	Precsion (%)
0	0	0
10	82.6±0.05	74.2±0.05
20	85.8±0.03	76.1±0.04
30	86.3±0.05	77.3±0.03
40	90.1±0.04	79.4±0.04
50	91.2±0.03	80.7±0.03
60	91.4±0.05	82.9±0.05
70	91.8±0.05	84.5±0.05
80	92.5±0.04	86.4±0.04
90	93.1±0.03	88.9±0.02
100	**93.7**±**0.03**	90.4±**0.02**

Table III showcases the mAP performance metrics for each model. Horizontal comparison tests revealed that our improved network model surpasses the performance of other models on our custom dataset. Results demonstrate a 2.7% mAP improvement of our model over the YOLOv7 baseline. The optimized YOLOv7-DSE algorithm has a model size of only 73.1MB, which is a 22.4% reduction compared to the baseline network YOLOv7. This demonstrates that the model has achieved a lightweight effect. Furthermore, Figure 13 shows mode enhanced detection of small objects.

TABLE III
THE COMPARISON OF MODELS

Models	mAP	F1	Model size	Miss rate
RCNN	52.67%	0.54	514.2MB	0.1%
Faster-RCNN	58.96%	0.60	523.6MB	0.08%
YOLOv4	72.94%	0.74	103.2MB	0.06%
YOLOv5	75.75%	0.74	101.6MB	0.06%
YOLOv7	79.74%	0.80	94.2MB	0.05%
YOLOv7-DSE	**82.38%**	**0.84**	**73.1MB**	**0.04%**
CEAM-YOLOv7	80.44%	0.81	96.4MB	0.04%
YOLOv7-RAR	79.86%	0.79	84.3MB	0.05%

E. Ablation Study

This paper introduces four modules aimed at enhancing performance. The ablation study seeks to validate the performance enhancements contributed by each module indi-

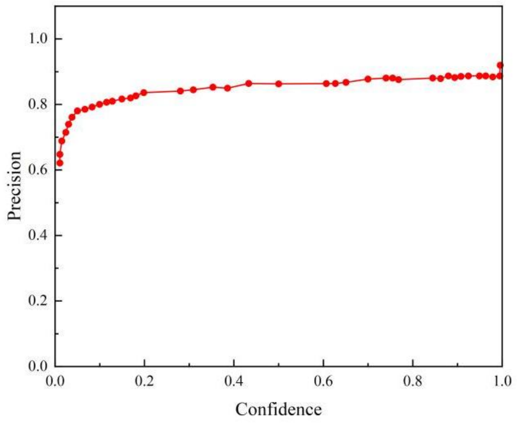


Fig. 9. Precision for YOLOv7

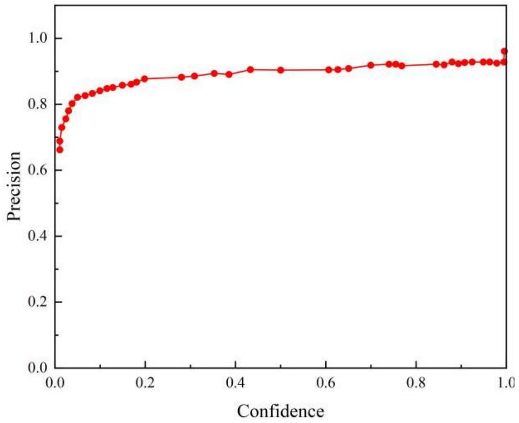


Fig. 10. Precision for YOLOv7-DSE

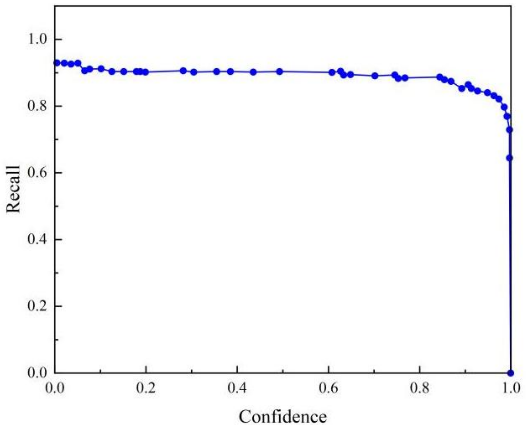


Fig. 11. Recall for YOLOv7

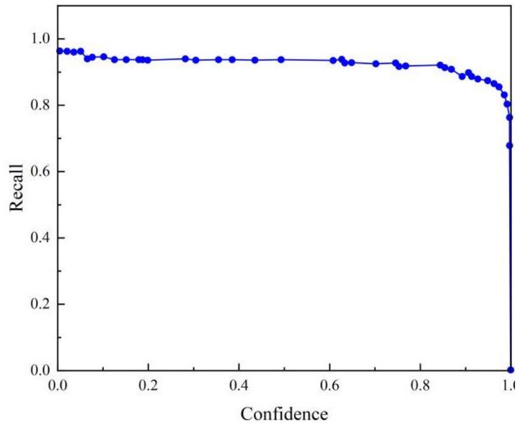


Fig. 12. Recall for YOLOv7-DSE

TABLE IV
ABLATION STUDY

Baseline YOLOv7	MP1-DSC	ELAN-DSC	ELAN-SPD	Epoch	Batchsize	mAP	Model size
√				100	16	79.74%	94.2MB
	√			100	16	80.25%	81.7MB
	√	√		100	16	80.57%	70.3MB
	√	√	√	100	16	82.38%	73.1MB

-vidually to the entire model. Modules were sequentially integrated into the original algorithm and tested at each stage using mAP and model size as the metric. The ablation experiment was conducted on YOLOv7, and the modules included in the sequence are MP1-DSC, ELAN-DSC, and ELAN-SPD. Experimental results are detailed in Table IV.

Results indicate that each module contributes to the overall model performance enhancement. Notably, ELAN-SPD provides the most significant improvement to the overall model. In YOLOv7, ELAN-SPD enhances the mAP of the model by 1.29%. Finally, the MP1-DSC module and ELAN-DSC module increased mAP by 0.51% and 0.32%, respectively. When the optimized depthwise separable convolution is frozen, the model's parameter count increases by 22.4%.

To illustrate the enhancement of the SPD convolutional layer in the head network for detecting small objects, the mAP value was greatly improved by substituting the ordinary convolutional layer with SPD convolutional layer. Table V indicates that the detection capability of all small targets was substantially enhanced. This largely resolves issues related to inaccuracies and missed detections of small targets.

V. CONCLUSION

This study introduces an enhanced algorithm using YOLOv7-DSE to detect workers's safety equipment. Comparative and ablation experiments were conducted. The experiments showed a 2.64% increase in mAP over the baseline YOLOv7 model. The optimized YOLOv7-DSE algorithm has a model size of 73.1MB, which is a 22.4% reduction compared to the baseline network YOLOv7. The enhanced YOLOv7-DSE model detects small objects with greater accuracy than competing algorithms. The detection results show our refined model effectively corrects false positives and false negatives.

However, the model's comprehensive network layers require advanced hardware and substantial memory use during detection. This hinders its deployment in standard embedded systems. Therefore, future research will focus on streamlining the model for lightweight network construction without compromising accuracy. Additionally, we will investigate advanced object detection networks to further address this challenge. Real-time object detection in video streams is also one of the future research directions.

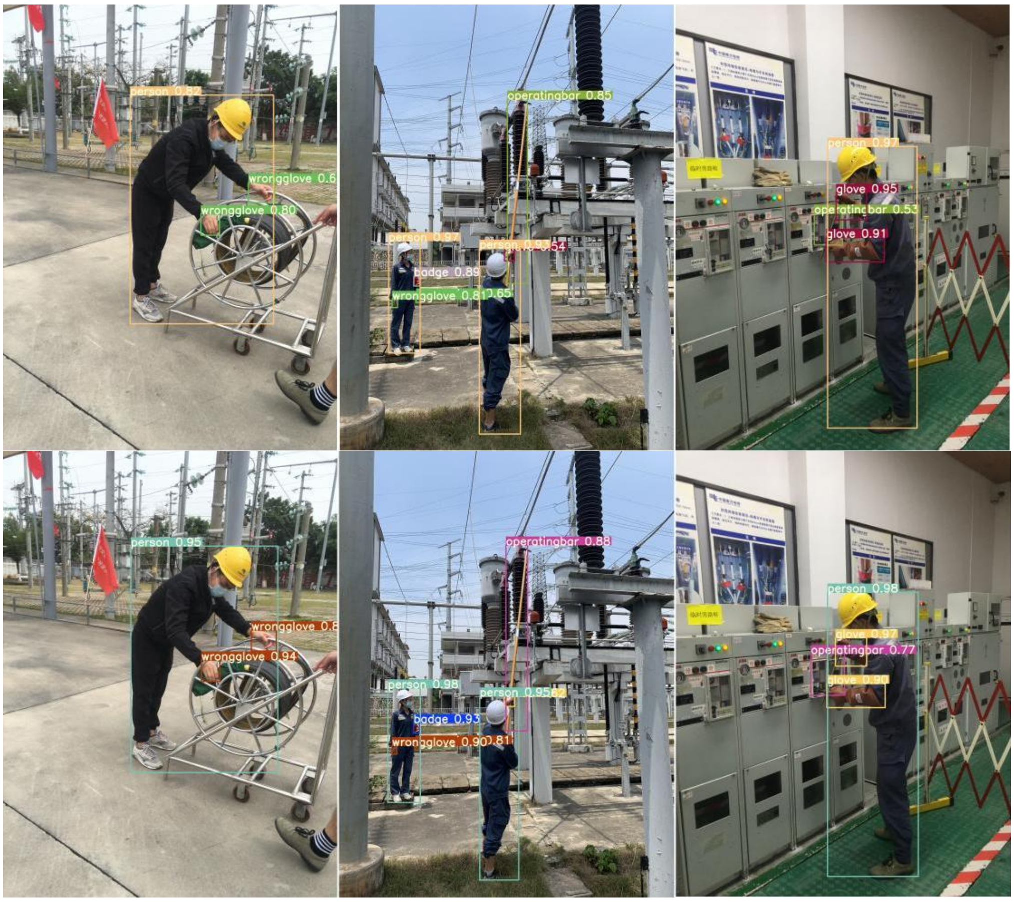


Fig. 13. Detection Results for YOLOv7-DSE

TABLE V
PERFORMANCE COMPARISON OF THE YOLOV7 MODELS FOR EACH CATEGORY ON THE TEST SET

MODEL	glove	wrongglove	operatingbar	person	powerchecker	badge
YOLOv7	78.03%	63.62%	62.3%	90.1%	93.1%	91.2%
YOLOv7-DSE	**80.88%**	**70.45%**	**67.85%**	90.2%	93.5%	91.4%

REFERENCES

[1] E. S. Marquez, S. Jonathon, and N. Mahesan, "Deep cascade learning," *IEEE transactions on neural networks and learning systems*, pp. 5475-5485, 2018.

[2] B. Sun, W. Li, et al. "Obstacle Detection of Intelligent Vehicle Based on Fusion of Lidar and Machine Vision," *Engineering letters,* vol. 29, no. 2, pp. 722-730, 2021.

[3] Y. Tan, J. Wang, "A support vector machine with a hybrid kernel and minimal Vapnik-Chervonenkis dimension," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 4, pp. 385-395, 2004.

[4] A. Md, et al. "Performance Evaluation and Comparative Analysis of Different Machine Learning Algorithms in Predicting Cardiovascular Disease," *Engineering Letters,* vol. 29, no. 2, pp.731-741, 2021.

[5] J. Deng, W. Dong, R. Socher, et al. "Imagenet: A large-scale hierarchical image database," *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 248-255, 2009.

[6] N. Iandola, S. Han, W. Moskewicz, et al. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size," arXiv:1602.07360, 2016.

[7] R. Girshick, J. Donahue, T. Darrell, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* pp. 580-587, 2014.

[8] X. Sun, W. Cui, Y.Tao, et al. "Flame Image Detection Algorithm Based onComputer Vision," *International Journal of Computer Science*, vol. 50, no. 4, pp. 2-18, 2023.

[9] W. Liu, D. Anguelov, D. Erhan, et al. "SSD: single shot multibox detector," *European Conference on Computer Vision*, pp. 21-37, 2016.

[10] J. Redmon, A. Farhadi, "YOLO9000: Better, faster, stronger," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263-7271, 2017.

[11] Y. Cao, J. Zhao, "Fast EfficientDet: An Efficient Pedestrian Detection Network," *Engineering Letters*, vol.30, no.2, pp. 537-545, 2022.

[12] H. Law, J. Deng, "Cornernet: Detecting objects as paired keypoints," *Proceedings of the European Conference on Computer vision,* pp. 734-750, 2018.

[13] S. Ren, K. He, R. Girshick, et al."Faster R-CNN:Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, 2017.

[14] K. He, G. Gkioxari, P. Dollár, et al. "Mask R-CNN," *Proceeding of the International Conference on Computer Vision*, pp. 2980-2988, 2017.

[15] Z. Cai, N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 6154-6162.

[16] Y. Wen, H. Chiu, J. Liaw, et al. "The safety helmet detection for ATM's surveillance system via the modified Hough transform," *IEEE 37th Annual 2003 International Carnahan Conference on Security Technology*, pp. 364-369, 2003.

[17] M. Rubaiyat, T. Toma, M. Kalantari-Khandani, et al. "Automatic detection of helmet uses for construction safety," *2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops* , pp. 135-142, 2016.

[18] K. Li, X. Zhao, J. Bian, et al. "Automatic safety helmet wearing detection," arXiv:1802.00264, 2018.

[19] J. Redmon, A. Farhadi, “YOLOv3: An incremental improvement,” arXiv:1804.02767, 2018.

[20] Y. Wang, A. Bochkovskiy, M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7464-7475, 2023.

[21] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1800-1807, 2017.

[22] R. Sunkara, T. Luo, "No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects," *Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer Nature Switzerland*, pp. 443-459, 2022.

[23] H. Rezatofighi, N. Tsoi, Y. Gwak, et al. "Generalized intersection over union: A metric and a loss for bounding box regression," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 658-666, 2019.

[24] F. Zhang, W. Ren, Z. Zhang, et al. "Focal and efficient IOU loss for accurate bounding box regression," *Neurocomputing*, vol. 506, pp. 146-157, 2022.

[25] B. Alexey, W. Chien-Yao, L. Hong-Yuan, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv:2004.10934, 2020.

[26] S. Ren, K. He, R. Girshick, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, pp. 28-37, 2015.

[27] R. Girshick, J. Donahue, T. Darrell, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580-587, 2014.

[28] S. Liu, Y. Wang, Q. Yu, et al. "CEAM-YOLOv7: Improved YOLOv7 based on channel expansion and attention mechanism for driver distraction behavior detection," *IEEE Access*, vol. 10, pp. 129116-129124, 2022.

[29] Y. Zhang, Y. Sun, Z. Wang, et al. "YOLOv7-RAR for Urban Vehicle Detection," *Sensors*, vol. 23, no. 4, pp. 1801-1812, 2023.

[30] S. Li, W. Liu, "Small Target Detection Model in Aerial Images Based on YOLOv7X+," *Engineering Letters*, vol. 32, no. 2, pp. 436-443, 2024.