

Article

Efficient and Lightweight Neural Network for Hard Hat Detection

Chenxi He ^{1,2}, Shengbo Tan ^{1,2}, Jing Zhao ^{1,2}, Daji Ergu ^{1,2}, Fangyao Liu ^{1,2,*}, Bo Ma ^{1,2} and Jianjun Li ^{3,*}

¹ College of Electronic and Information, Southwest Minzu University, Chengdu 610093, China; hxcy1869@163.com (C.H.); tanshengbo123@163.com (S.T.); 15383414924@163.com (J.Z.); ergudaji@163.com (D.E.); maphd1969@163.com (B.M.)

² Key Laboratory of Electronic and Information Engineering, State Ethnic Affairs Commission, Chengdu 610041, China

³ College of Information, Sichuan Vocational College of Finance and Economics, Chengdu 610101, China

* Correspondence: fliu028@163.com (F.L.); jianjunlisvcfe@163.com (J.L.)

Abstract: Electric power operation, as one of the key fields in the world, faces particularly prominent safety issues. Ensuring the safety of operators has become the most fundamental requirement in power operation. However, there are some safety hazards in power construction. These hazards are mainly due to weak safety awareness among staff and the failure to standardize the wearing of safety helmets. In order to effectively address this situation, technical means such as video surveillance technology and computer vision technology can be utilized to monitor whether staff are wearing helmets and provide timely feedback. Such measures will greatly enhance the safety level of power operation. This paper proposes an improved lightweight helmet detection algorithm named YOLO-M3C. The algorithm first replaces the YOLOv5s backbone network with MobileNetV3, successfully reducing the model size from 13.7 MB to 10.2 MB, thereby increasing the model's detection speed from 42.0 frames per second to 55.6 frames per second. Then, the CA attention mechanism is introduced into the backbone network to enhance the feature extraction capability of the model. Finally, in order to further improve the detection recall rate and accuracy of the model, a knowledge distillation of the model was carried out. The experimental results show that, compared with the original YOLOv5s algorithm, the average accuracy of the improved YOLO-M3C algorithm is improved by 0.123, and the recall rate is the same. These results verify that the algorithm YOLO-M3C has excellent performance in target detection and recognition, which can improve accuracy and confidence, while reducing false detection and missing detection, and effectively meet the needs of helmet-wearing detection.



Citation: He, C.; Tan, S.; Zhao, J.; Ergu, D.; Liu, F.; Ma, B.; Li, J. Efficient and Lightweight Neural Network for Hard Hat Detection. *Electronics* **2024**, *13*, 2507. <https://doi.org/10.3390/electronics13132507>

Academic Editor: Daniel Riccio

Received: 14 May 2024

Revised: 6 June 2024

Accepted: 20 June 2024

Published: 26 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: hard hat; knowledge distillation; lightweight network; attention mechanism; object detection

1. Introduction

With the promotion of artificial intelligence in society, industrial intelligence can effectively improve production efficiency while ensuring the safety of workers during the production process. In this progress, computer vision technology plays an indispensable role. In industries such as construction and mining, workers are required to wear safety helmets during operations to ensure their safety. Therefore, helmet-wearing inspections have become a critical task in production safety management. However, due to the hazardous nature of work environments like construction sites, real-time monitoring by human personnel is not feasible. Consequently, real-time monitoring of correct helmet usage has become an important application scenario for intelligent embedded device development and research. Through intelligent technology, it is possible to develop devices capable of monitoring workers' helmet usage. These devices can utilize image recognition technology to determine whether workers are wearing safety helmets. If a worker is detected not wearing a helmet, the system can promptly issue an alert and take appropriate preventive measures. The development of such intelligent helmet monitoring systems provides an efficient and precise safety management tool for industries like construction and mining. This will help ensure the safety of workers and reduce the likelihood of accidents.

In this study, it was found that there are still shortcomings in YOLOv5's detection of safety helmets when conducting experiments using YOLOv5 for safety-helmet-wearing detection. Due to the large number of model parameters, the model cannot be deployed for mobile devices, and the effect of the model is not ideal when detecting small-size safety hats. Therefore, this paper proposes a lightweight helmet-wearing detection model, YOLO-M3C. The YOLOv5s backbone network is replaced by MobileNetV3 for feature extraction, which greatly reduces the network computation, making its deployment on mobile devices possible. Then, CA (coordinate attention) is added to the network to strengthen the attention to the detection target. Finally, knowledge distillation is applied to the model to enhance the effectiveness of knowledge transfer. The experimental results show that the model performs well in target detection. At the same time, the model can also effectively reduce the cost of hardware, making it possible to deploy on low computing power platforms. Through this combination of high performance and low cost, the YOLO-M3C model meets the needs of target detection algorithms in practical application scenarios, providing a more flexible and feasible solution for various applications.

The main contributions of this paper are:

- 1: The backbone network of Mobilenetv3 is used to replace the backbone network of YOLOv5s to carry out feature extraction, greatly reducing the amount of network computation and making the network more lightweight.
- 2: A CA attention mechanism is added to the network to obtain cross-channel information and capture direction perception and position perception information so that the model can locate and identify the target more accurately.
- 3: Distillation of the improved YOLOv5s network to improve the performance of lightweight models. In this process, we use an integrated distillation method that combines three scales: classification, confidence, and bounding box for the transfer of output feature knowledge.

2. Related Work

At present, many people detect hard hats using a detection algorithm that relies on the artificial extraction of image features. Li et al. [1] used head positioning, color space transformation, and color feature recognition to achieve the detection of wearing helmets based on the detection results of pedestrians. Wu et al. [2] used the improved YOLO-Dense backbone for helmet detection to improve feature resolution. Long et al. [3] used an SSD (Single Shot multi-box Detector) [4] to detect helmet wearing. Chen et al. [5] introduced the K-means++ clustering algorithm to cluster the size of the helmet in the image and then used the improved Faster-RCNN algorithm to detect helmet wearing. Park et al. [6] extracted the HOG features of hard hats, and then input the extracted features into a support vector machine (SVM). Mneymneh et al. [7] matched the spatial information of workers and helmets, and judged whether workers were wearing helmets according to the matching points of the features. The above traditional methods all use manually extracted features, which have the characteristics of poor robustness and low real-time performance and that can only achieve helmet detection in specific scenarios. However, the field monitoring of images is complex and the light is changeable, so the method of artificial feature extraction can not meet the needs of safety helmet detection.

Helmet detection algorithms can be roughly divided into two types of methods based on sensors and computer vision, and computer-vision-based helmet detectors can be divided into traditional helmet detection algorithms and deep-learning-based helmet detection algorithms. Sensor-based helmet detection algorithms mainly rely on positioning technology; during the detection process, workers need to wear a physical tag or sensor. However, this requirement may have an impact on the normal operation of workers. Therefore, the sensor-based approach has some limitations in its application. In contrast, traditional helmet detection algorithms mainly include the circular Hough transform method [8], the background difference method [9] and the DPM [10] (Deformable Part Model) algorithm. The training steps of these methods are similar. Firstly, the candidate

region is extracted through the active window, then the selected feature region is extracted, and finally, content classification is carried out to determine whether the worker is wearing a safety helmet. In this process, commonly used feature extraction methods include Haar-like features [11], Local Binary Pattern (LBP) [12], and Histogram of Oriented Gradient (HOG) [13]. Then, AdaBoost [14], a support vector machine [15] and other methods are used for classification. However, the traditional helmet detection algorithms are mostly based on low-level features set manually and lack the ability to express multi-class targets. In addition, the region selection based on the sliding window algorithm has redundancy, resulting in high complexity of the algorithm, which cannot effectively balance the relationship between precision and efficiency.

Traditional helmet detection methods mainly rely on multi-feature fusion, and obtain skin color, head and face information through image processing technology, and then carry out corresponding detection. For example, Waranusast et al. [16] proposed a K-nearest neighbor-based automatic detection system for motorcycle helmets. Li et al. [17] used Hough transform to determine the shape of the helmet and used the corresponding histogram to train the support vector machine for detection. Filatov et al. [18] combined SqueezeDet and MobileNets on the basis of YOLOv4 to significantly improve the accuracy of the helmet detection algorithm. Based on YOLOv5, Li et al. [19] proposed a hierarchical positive sample selection mechanism to improve the fit of the model. The literature [20] improves the recognition rate of the model by introducing an attention mechanism and multilevel features. Yifan Zhao et al. [21] designed a new infrared pedestrian detection framework that utilizes temperature information in infrared images to improve detection accuracy and robustness. This method can also be applied to helmet detection.

3. Method

Aiming to address the shortcomings of the existing technology, this paper proposes a safety helmet detection model YOLO-M3C based on YOLOv5s. First, MobileNetV3 is integrated into the backbone network of the model to replace the original backbone layer. Specifically, the module after the convolution layer in the original backbone is replaced with the MobileNetV3_InvertedResidual module, so that it can greatly reduce the network computing amount. In order to obtain better model accuracy, a CA attention mechanism is introduced in the neck part to improve the ability of the neck layer to capture and process features.

Finally, the performance of the lightweight model is improved by knowledge distillation. We use YOLOv5m as the teacher network and the improved YOLOv5s as the student network. In this process, based on the YOLO model, object detection is divided into three output categories—classification, confidence, and bounding box. We use an integrated distillation method that combines three scales: classification, confidence, and bounding box for the transfer of output feature knowledge. The student network can learn the output rules of the teacher network in the logits layer, so that the student network can learn knowledge from the teacher network output in the training process. This allows you to establish dependencies between target features to improve the effectiveness of knowledge transfer and bring the student network closer to the level of the teacher network. Figure 1 shows the overall framework structure of this paper.

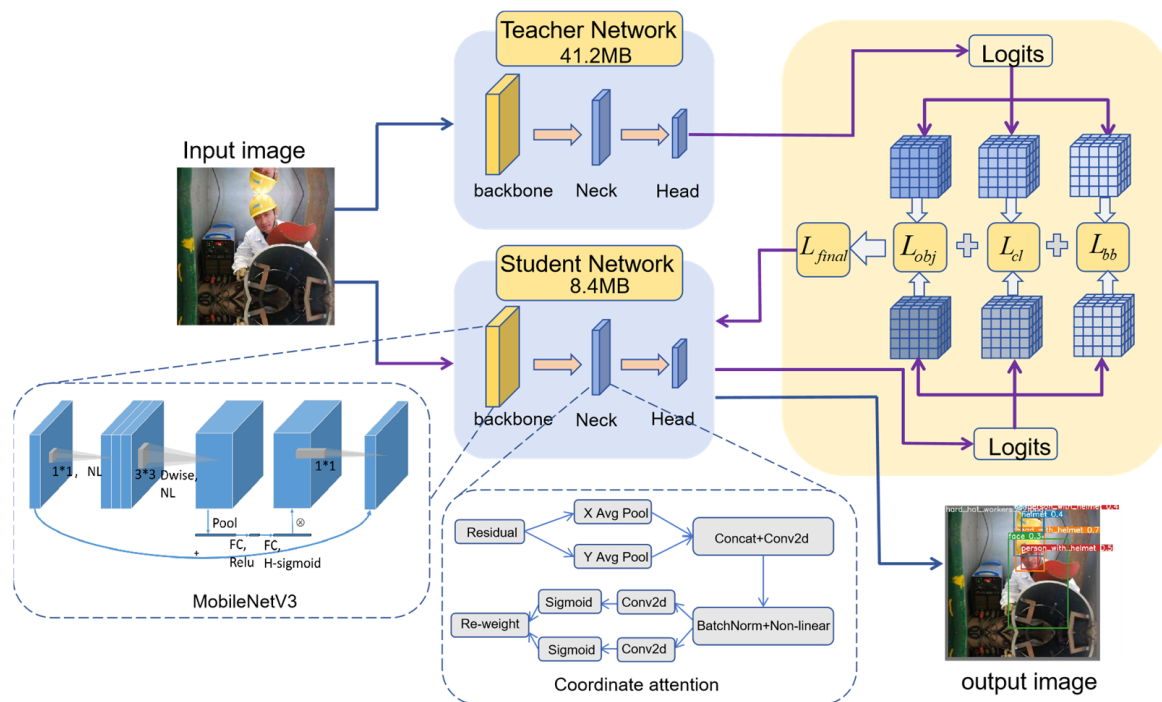


Figure 1. Overview of the proposed method. Replace the backbone module in the student network with the MobilenetV3_InvertedResidual module and introduce coordinate attention (CA) in the neck part of the network. Finally, an integrated distillation method is used, combining three scales—classification, confidence, and bounding box—for the transfer of output feature knowledge.

3.1. YOLO-M3C

YOLOv5 carries out two CSP (Cross Stage Partial) [22] network designs identical to YOLOv4 for the network structure and applies them to the backbone network and the middle layer. In the middle layer, a combination of FPN (Feature Pyramid Network) [23] and PAN (Perceptual Adversarial Network) [24] is used.

To deploy algorithms on mobile devices in the future, it is necessary to reduce the size of the models. Kang et al. [25] introduced the MobileNetV2 [26] network as a replacement for the backbone of the YOLO v5s algorithm, achieving a combination of high accuracy and low complexity. Lisang Liu et al. [27] addressed the issue of low localization accuracy in YOLOv4-tiny by optimizing its backbone feature extraction network using MobileNetV3. The backbone network in this paper adopts the MobileNetV3 network structure to replace the original backbone layer. This structure is mainly developed for embedded and mobile devices, so that the model can be better deployed on mobile devices that are lightweight. In YOLOv5s, the module behind the convolution layer in the backbone has been replaced with the MobileNetV3_InvertedResidual module. The parameters of MobileNet V3 were obtained through Network Architecture Search (NAS), and it also employs the SE channel attention mechanism [28]. Therefore, MobileNet V3 was born from the synthesis of three model ideas: the depth-wise separable convolution of MobileNet V1 [29], the inverted residual structure of MobileNet V2, and lightweight attention models. Additionally, MobileNet V3 redesigns time-consuming layer structures, reducing the number of layers compared to previous designs. This adjustment almost does not affect the model's accuracy but significantly improves the inference speed, saving approximately 7 milliseconds of inference time. This 7-millisecond time saving constitutes about 11% of the total inference time. Overall, the module adopts a reciprocal residual structure to reduce the complexity of the model through depth-separable convolution and linear bottleneck operations. At the same time, h-switch is used as the activation function of the model, and the input features are spatially processed by 3*3 convolution. Finally, a single pixel convolution layer is used

to restore the number of channels to the original level, achieve dimensionality reduction, and add residual connections to the output.

Then, the CA [30] mechanism is introduced into the neck part to improve the neck layer's ability to capture and process features. This mechanism is designed for lightweight networks. It embeds location information into channel attention. Unlike traditional channel attention mechanisms, this approach decomposes channel attention into two 1D feature encoding processes, aggregating features along different directions. The advantage of this design lies in its ability to capture long-range spatial dependencies while preserving precise location information. Subsequently, the generated feature maps form a pair of direction-aware and position-sensitive feature maps after encoding separately. These feature maps can be applied complementarily to the input feature maps, enhancing the representation of the target of interest.

Compared to previous lightweight network attention methods, coordinate attention has several advantages. Firstly, it not only captures cross-channel information but also captures direction-aware and position-aware information, aiding the model in more accurately locating and identifying the target of interest. Secondly, coordinate attention is highly flexible, lightweight and easily embeddable into classic modules. Finally, as a pre-trained model, coordinate attention can bring significant benefits to downstream tasks based on lightweight networks, especially in scenarios with dense prediction tasks. These characteristics mean that coordinate attention has significant potential for applications in lightweight networks. Figure 2 shows the network architecture of YOLOv5. Figure 3 shows the improved network architecture. Figure 4 shows the MobileNetV3 structure.

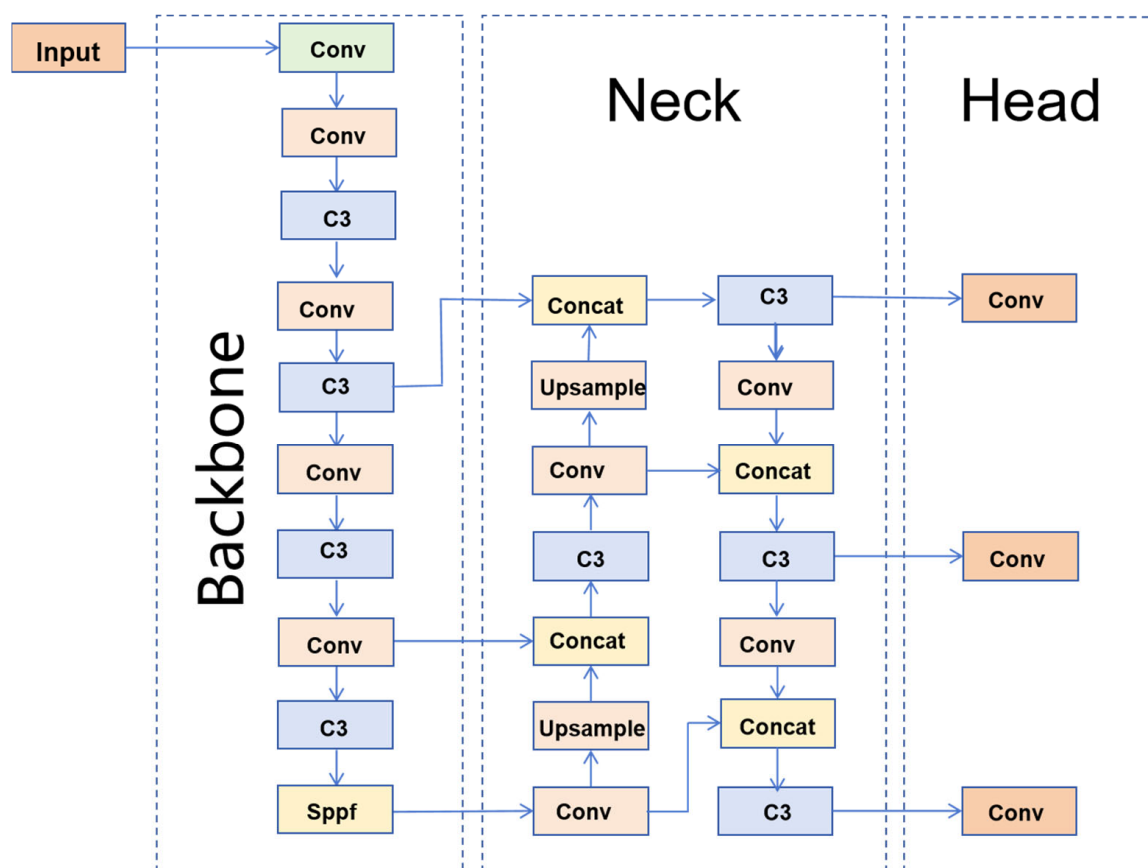


Figure 2. The network architecture of YOLOv5.

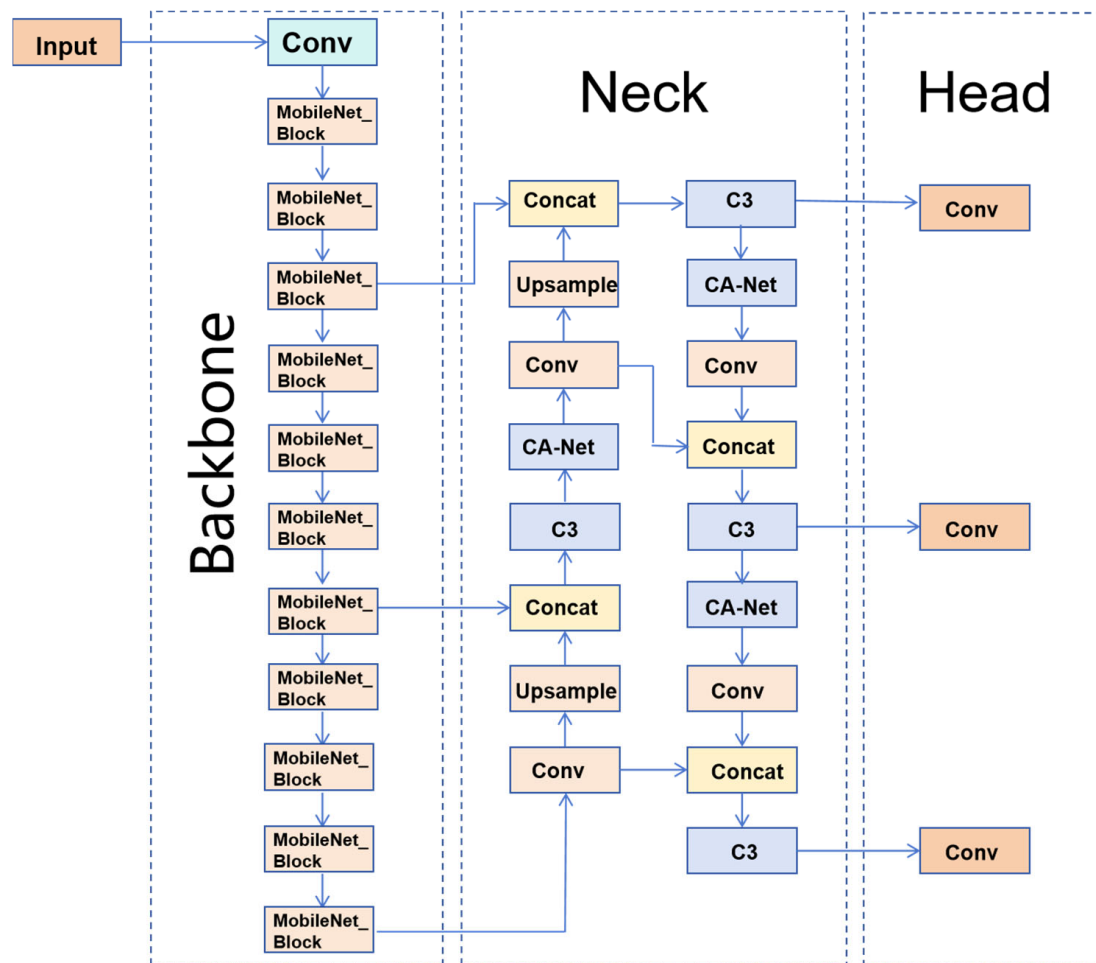


Figure 3. Improved network architecture. The module after the convolution layer in the backbone is replaced with the MobilenetV3_InvertedResidual module and the CA attention mechanism is introduced into the neck part.

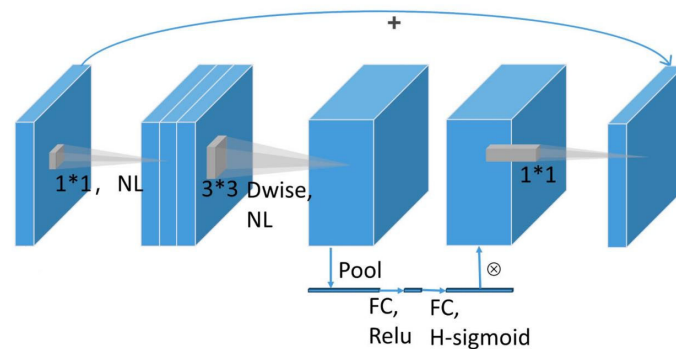


Figure 4. MobileNetV3 structure. MobileNetV3 integrates depth-wise separable convolutions, inverted residual structures, and linear bottleneck layers to enhance the model's representational power and computational efficiency.

3.2. YOLO's Multi-Scale Output Feature Distillation Design

In order to further improve the detection performance of the experiment, we applied knowledge distillation [31] to the model. Knowledge distillation is a model compression method based on the “teacher-student network” training approach, widely used in the industry due to its simplicity and effectiveness. This method follows the teacher–student paradigm: a complex and large model serves as the teacher, while the student model has a simplified structure. The teacher model plays a role in assisting the training of the

student model. With strong learning capabilities, the teacher model transfers its learned knowledge to the relatively weaker student model, thereby enhancing the generalization ability of the student model. In practical deployment, the complex and large teacher model is not used for online tasks but acts as a mentor, while the flexible and lightweight student model performs the actual prediction tasks. This approach aims to reduce computation and storage costs for online deployment while ensuring model performance. Yang et al. [32] applied KD to YOLOv5, using YOLOv5m as the teacher and YOLOv5n as the student for bell detection, which improved the YOLOv5n *mAP* by about 2%. Martin Aubard et al. [33] combined the YOLOX-ViT model with knowledge distillation technology, successfully maintaining high detection accuracy while reducing the model size. This paper adopts the YOLOv5m model as the teacher model and YOLOv5s as the student model for distillation to enhance the performance of the student model. The knowledge distillation process is shown in Figure 5.

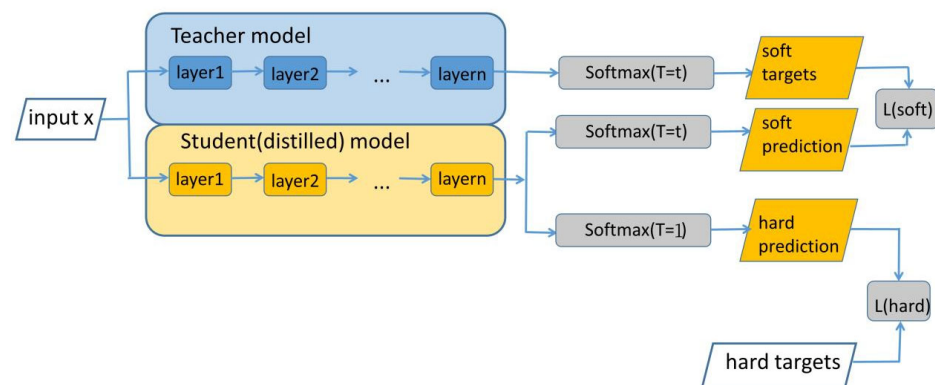


Figure 5. Knowledge distillation process. The total loss of knowledge distillation typically consists of the loss $L(\text{hard})$ between the student model's predictions and the true values, and the distillation loss $L(\text{soft})$ between the teacher model and the student model.

This paper uses the output characteristics of the teacher network to guide students' network training. The output characteristic knowledge usually refers to the output of the last layer of the teacher network, including the transmitted logical unit and soft target knowledge. The core idea of output feature distillation is to guide the student network to mimic the final prediction results of the teacher network in order to achieve similar prediction performance. Considering the characteristics of the YOLO network and the requirements of the target detection task, the complete output feature knowledge includes three aspects: classification loss, objectness loss and bounding box loss. Therefore, we propose an integrated distillation method that combines three scales of classification, confidence and bounding box for the transmission of output feature knowledge.

By extracting the classification knowledge from the output features of the teacher network, the student network obtains the soft target knowledge of category probability to improve the classification accuracy. By using the confidence in the output features of the teacher network, the student network can gain insight into the reliability of the prediction box from the output features of the teacher, thus improving the confidence. The boundary box is extracted from the output features of the teacher network, so that the student network can learn the size and specific location information of the target from the output features of the teacher network, so as to improve the localization accuracy of the student network.

The total distillation loss includes student loss and distillation loss.

The student loss for classification loss is expressed as $f_{obj}(o_i^{gt}, \hat{o}_i)$, the student loss for confidence loss is expressed as $f_{cl}(p_i^{gt}, \hat{p}_i)$, and the student loss for bounding box loss is expressed as $f_{bb}(b_i^{gt}, \hat{b}_i)$.

$\hat{o}_i$ is the classification of student network prediction, $\hat{p}_i$ is the confidence of the student network prediction, $\hat{b}_i$ is the bounding box of student network prediction. $o_i^{gt}, p_i^{gt}, b_i^{gt}$ are their ground truth. The loss function is L2 loss function:

$$L = \frac{1}{n} \sum_{i=1}^n (O_i^t - O_i^s)^2. \quad (1)$$

However, there are some problems with the introduction of distillation loss directly into the YOLO algorithm. For YOLO algorithms, most of the bbox generated will be background. If a large number of background regions are passed to the student network, it will cause the network to constantly return the coordinates of these background regions and classify these background regions, so it is difficult to converge the model in training. Therefore, this paper limits the output objectness of the teacher network to a certain extent, and only the bbox with a higher output objectness of the teacher network will contribute to the final loss function of the student network.

First, the objectness loss is:

$$f_{obj}^*(o_i^{gt}, \hat{o}_i, o_i^T) = f_{obj}(o_i^{gt}, \hat{o}_i) + \lambda_D f_{obj}(o_i^T, \hat{o}_i). \quad (2)$$

This formula consists of two parts: the first part is student loss and the second part is distillation loss. One of the inputs of this part is no longer the ground truth, but the output of the teacher network, which is used to balance two losses, and the default is 1.

Classification loss is

$$f_{cl}^*(p_i^{gt}, \hat{p}_i, p_i^T, \hat{o}_i^T) = f_{cl}(p_i^{gt}, \hat{p}_i) + \hat{o}_i^T \lambda_D f_{cl}(p_i^T, \hat{p}_i). \quad (3)$$

This formula is also composed of two parts. In the second part, the objectness output of the teacher network is added, indicating the probability that each bbox contains an object. Therefore, if a bbox is background, the probability value will be very low. Then, the distillation loss part is basically absent, so as to prevent students from mislearning such background information on the Internet.

Bounding box loss and classification loss is

$$f_{bb}^*(b_i^{gt}, \hat{b}_i, b_i^T, \hat{o}_i^T) = f_{bb}(b_i^{gt}, \hat{b}_i) + \hat{o}_i^T \lambda_D f_{bb}(b_i^T, \hat{b}_i). \quad (4)$$

The final total loss function is as follows:

$$L_{final} = f_{obj}^*(o_i^{gt}, \hat{o}_i, o_i^T) + f_{cl}^*(p_i^{gt}, \hat{p}_i, p_i^T, \hat{o}_i^T) + f_{bb}^*(b_i^{gt}, \hat{b}_i, b_i^T, \hat{o}_i^T). \quad (5)$$

4. Experimental Results and Comparative Analysis

4.1. Experiment Settings

This experiment was built, trained and verified by the Pytorch deep-learning framework under a NVIDIA GeForce RTX 3050 Laptop GPU graphics card. In this paper, two datasets were used: Safety Helmet Wearing Dataset 1 [34] and Safety Helmet Wearing 2 [35]. The first dataset contains 5000 images with a total of 75,578 labels, and the second dataset contains 7851 images. Both datasets were split into training and validation sets in an 8:2 ratio. The image resolution size was 640*640, the initial learning rate was 0.003, the IoU threshold was set to 0.5, and the mixup was set to 0.5. All reference models trained 300 epochs according to this parameter.

4.2. Evaluation Index

In this paper, the ablation experiment and comparison experiment are carried out to prove that each improvement point of the network can improve the performance of the model. In the ablation experiment, recall, *mAP*, parameter number, calculation amount and model size were used to compare the performance of each model. The average precision

(mAP), parameter number and model size were used to compare the performance of each model.

Recall, mAP is calculated as follows:

$$recall = \frac{TP}{TP + FN} \quad (6)$$

$$mAP = \frac{1}{Q} \sum_{q \in Q} AP(q). \quad (7)$$

The calculation formula of average precision is shown as follows:

$$AP = \frac{TP + TN}{TP + TN + FP}. \quad (8)$$

TP : True Positive, the predicted result of the classifier is positive samples, and the actual samples are positive samples, that is, the number of positive samples correctly identified. FP : False Positive, the predicted result of the classifier is positive samples, but the actual sample is negative samples, that is, the number of negative samples that are falsely positive. TN : True Negative, the predicted result of the classifier is negative samples, but it is actually negative samples, that is, the number of negative samples correctly identified. FN : False Negative, the predicted result of the classifier is negative samples, but the actual sample is positive samples, that is, the number of positive samples missed. AP is the average precision of all possible values of recall rates, and mAP is the average of AP values taken under all categories.

4.3. Ablation Study

In this paper, MobileNetv2 and MobileNetv3 replace the original YOLOv5s backbone network, respectively, for performance test and comparison. After testing the test set, the results are shown in Table 1:

Table 1. Different versions of the MobileNet experiment.

Network	Recall	mAP	Detection Speed (Frames/s)	Params/ 10^6
YOLOv5s	0.90	0.699	42.0	7.0
YOLOv5s+ MobileNetv2	0.84	0.634	62.5	5.5
YOLOv5s+ MobileNetv3	0.88	0.678	55.6	5.3

The detection speed of MobileNetv2-YOLOv5s reached 62.5 frames/s at the fastest, which was 20.5 frames/s higher than that of the original YOLOv5s target detection model. The minimum number of model parameters of MobileNetv3-YOLOv5s was about 5.2 million, which is 1.9 million less than the original YOLOv5s. The optimized YOLOv5s model can meet the real-time detection standard of safety helmet wearing in power plants. Therefore, this paper ultimately chose the MobileNetv3 structure as the backbone network structure of the improved YOLOv5s.

On the basis of improving the YOLOv5s backbone network by using MobileNetv3, the attention mechanism CA was added to YOLOv5s, and knowledge distillation was carried out to increase the feature extraction capability of the model. In order to verify the detection effect of the YOLO-M3C target detection model. In this paper, the YOLO-M3C is compared with the unimproved model, as shown in the Table 2:

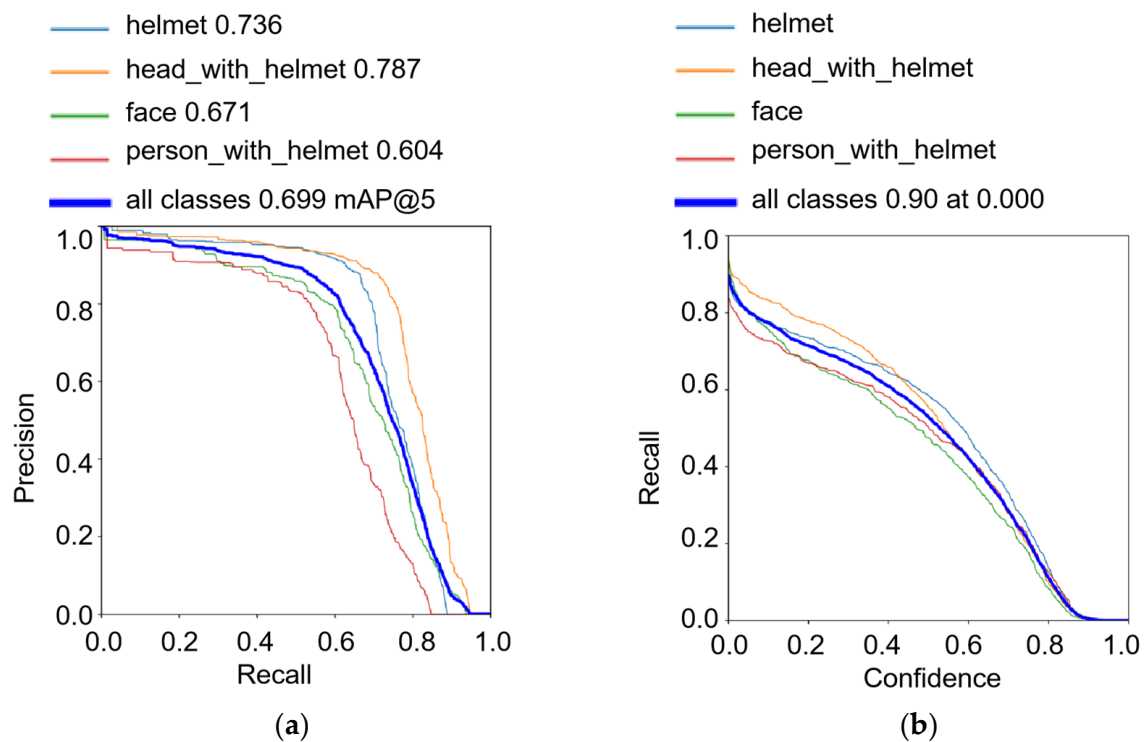
Table 2. Comparison before and after algorithm improvement.

Network	Recall	<i>mAP</i>	Params/ 10^6	FLOPs/ 10^9	Model Size/MB
YOLOv5s	0.90	0.699	7.0	16.0	13.7
YOLOv5s+ MobileNetv3	0.88	0.678	5.3	10.0	10.2
YOLOv5s+ MOBILENetv3+CA	0.89	0.762	4.2	9.6	8.4
YOLOv5s-M3C (ours)	0.90	0.822	4.2	9.6	8.4

The YOLOv5s model has a size of 13.7 MB and a computational capacity of 16×10^9 . After the CA attention mechanism was further added for optimization, the size of the model was reduced to 8.4 MB, the computation amount was 9.6×10^9 , the average accuracy value of the model was improved compared with that before the attention mechanism was added, reaching 0.762, and the number of model parameters was only 4.2 million. The *mAP* of the improved YOLOV5S-M3C model is only 12 percentage points higher than that of YOLOv5s, whereas the recall rate is equal to that of YOLOv5s. It can be seen that with these improvements, a significant reduction in computation, number of parameters, and model size has been achieved, while maintaining a high level of *mAP*. It meets the deployment conditions of edge computing terminal for real-time detection of safety helmet wearing in power plants.

The mean average accuracy curve of the YOLOv5s and YOLO-M3C training results on the helmet data set is shown in Figure 6.

In addition, in order to more intuitively feel the detection difference between the improved algorithm and YOLOv5s, situation images of obscure and easily confused targets were selected for detection and comparison, and the results are shown in Figure 7.

**Figure 6.** Cont.

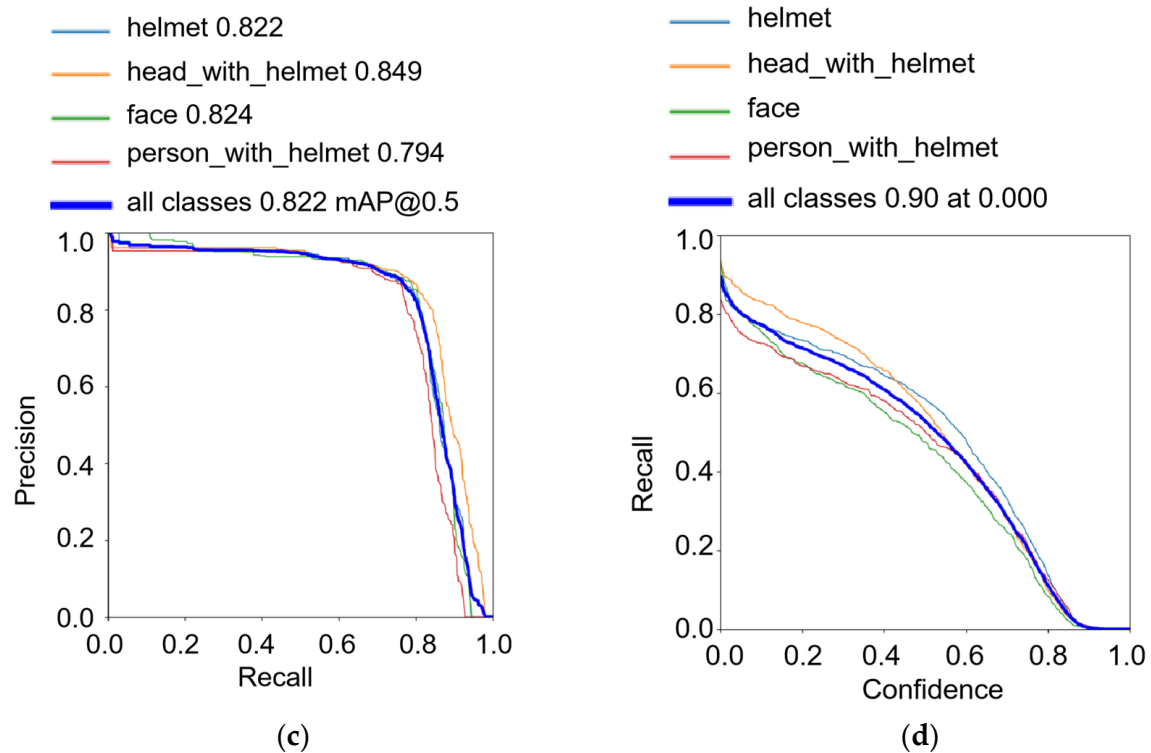


Figure 6. (a) YOLOv5s average accuracy chart; (b) YOLOv5s recall rate chart; (c) YOLO-M3C average accuracy chart; (d) YOLO-M3C recall rate chart.

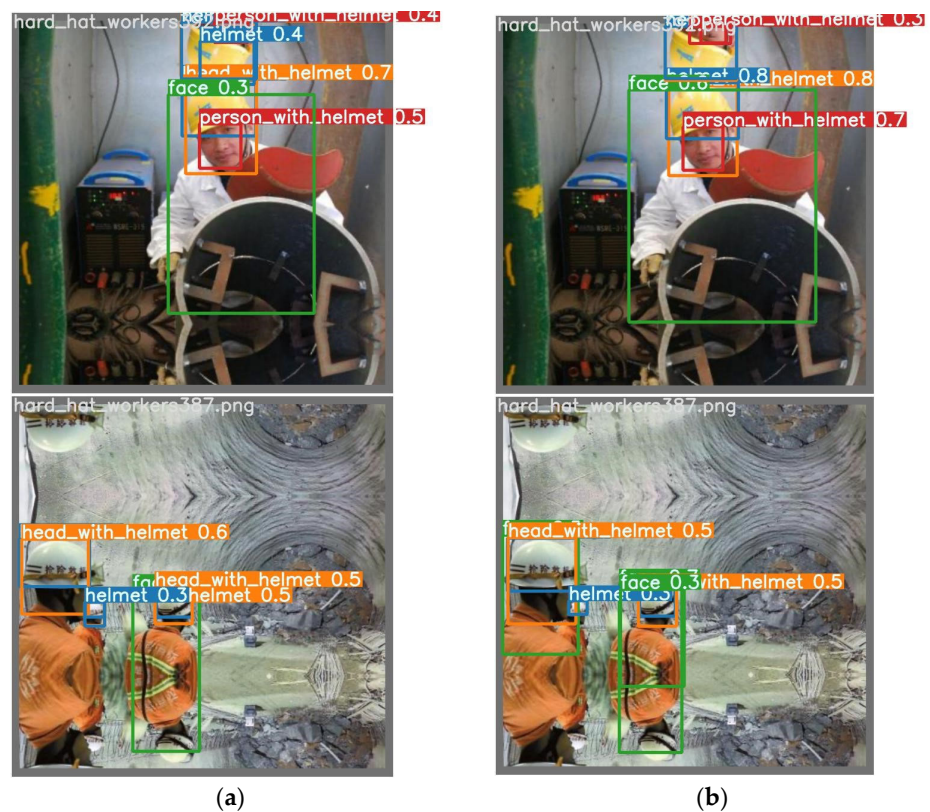


Figure 7. (a) YOLOv5s test result; (b) YOLO-M3C test result.

4.4. Comparison of YOLO-M3C with Other Algorithms

Comparing YOLO-M3C with other mainstream algorithms, the performance of YOLO-M3C was analyzed to further prove the superiority and feasibility of testing YOLO-M3C. Tables 3 and 4 show the results of the experiments conducted using the two datasets:

Table 3. The experimental results of YOLO-M3C compared with other YOLO algorithms using Safety Helmet Wearing Dataset 1.

Network	<i>mAP</i>	Params/ 10^6	Model Size/MB
YOLO-M3C (ours)	0.822	4.2	8.4
ShuffleNetV2-YOLOv5s	0.635	1.4	2.8
GhostNet-YOLOv5s	0.795	3.6	7.3
YOLOv3	0.816	62.0	236.0

Table 4. The experimental results of YOLO-M3C compared with other algorithms using Safety Helmet Wearing Dataset 2.

Network	<i>mAP</i>	Params/ 10^6	Model Size/MB
YOLO-M3C (ours)	0.806	4.2	8.4
Fast R-CNN	0.615	18.6	182
SSD	0.73	25.0	188
SSD-Lite	0.78	3.4	25
MobilNetV2 SSD-Lite	0.412	3.0	23.7

According to the experimental results, compared with MobilNetV2 SSD-Lite, the average precision of YOLO-M3C was improved by 40 percentage points and the model size was greatly reduced. Compared to SSD-Lite, the model's average precision increased by 2 percentage points, and the model size was reduced by 16.6 MB. Compared with ShuffleNetV2-YOLOv5s and GhostNetYOLOv5s, the average precision of YOLO-M3C was improved by 18.7 percentage points and 2.7 percentage points, although the model size and parameter number were slightly larger. Compared with the mainstream detection network models Fast R-CNN, SSD and YOLOv3, the model size and parameter number of YOLO-M3C were significantly reduced, but the average accuracy was significantly improved. Overall, compared to current models based on lightweight improvement algorithms and mainstream detection algorithms, YOLO-M3C shows better performance. It successfully reduces the number of parameters and the size of the model while maintaining an excellent average accuracy level.

5. Conclusions

To address issues where object detection models cannot meet the deployment requirements of edge computing in actual power plant environments and perform poorly in detecting small-sized safety helmets in such environments, this study has improved the object detection model to adapt to the deployment conditions of edge computing.

This study has improved the original YOLOv5s object detection model. The lightweight network MobileNetV3 was used to replace the original model's backbone network, CSP-Darknet53. Through a series of experiments and comparative analyses, we selected MobileNetV3 as the backbone network for the improved model and proposed the MobileNetV3-YOLOv5s object detection model. Compared to the original model, although MobileNetV3-YOLOv5s shows a slight decrease in detection accuracy, it significantly improves the speed of helmet detection while reducing the model size. To further enhance the model's feature extraction capability, this study introduced the CA attention mechanism based on the improved lightweight network model MobileNetV3-YOLOv5s. With this improvement, we aim to meet the real-time and portability requirements for helmet detection in power plant environments while improving the model's detection performance. Finally, to further enhance the model's detection recall and accuracy, knowledge distillation was applied to

the model. Testing results show that compared to the model before improvement, YOLO-M3C has achieved some improvements in detection accuracy and meets the deployment requirements for devices such as embedded systems.

Author Contributions: Formal analysis, S.T.; data curation, J.Z.; writing—original draft preparation, C.H.; writing—review and editing, J.L.; visualization, D.E.; supervision, F.L.; project administration, B.M.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This project is supported by the Southwest Minzu University Graduate Innovative Scientific Research Project (Project No. YB2023658), and in part by the Science and Technology Planning Project of Sichuan Province of China under Grant 2022YFG0377. Also, this research is supported by a grant from the Sichuan Vocational education talent training and education and teaching reform research project (GZJG2022-329), and in part by the National Natural Science Foundation of China under Grant 72174172.

Data Availability Statement: The data used in the research paper are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Li, K.; Zhao, X.; Bian, J.; Tan, M. Automatic Safety Helmet wearing detection. In Proceedings of the 2017 IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), Honolulu, HI, USA, 31 July–4 August 2017. [\[CrossRef\]](#)
- Wu, F.; Jin, G.; Gao, M.; He, Z.; Yang, Y. Helmet detection based on improved Yolo V3 Deep Model. In Proceedings of the 2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC), Banff, AB, Canada, 9–11 May 2019. [\[CrossRef\]](#)
- Long, X.; Cui, W.; Zheng, Z. Safety helmet wearing detection based on Deep Learning. In Proceedings of the 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, 15–17 March 2019. [\[CrossRef\]](#)
- Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision—ECCV 2016, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2016; pp. 21–37. [\[CrossRef\]](#)
- Chen, S.; Tang, W.; Ji, T.; Zhu, H.; Ouyang, Y.; Wang, W. Detection of safety helmet wearing based on improved faster R-CNN. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020. [\[CrossRef\]](#)
- Park, M.; Elsafty, N.; Zhu, Z. Hardhat-Wearing detection for enhancing On-Site safety of construction workers. *J. Constr. Eng. Manag.* **2015**, *141*, 04015024. [\[CrossRef\]](#)
- Mneymneh, B.E.; Abbas, M.; Khoury, H. Automated hardhat detection for construction safety applications. *Procedia Eng.* **2017**, *196*, 895–902. [\[CrossRef\]](#)
- Merlin, P.M.; Farber, D. A parallel mechanism for detecting curves in pictures. *IEEE Trans. Comput.* **1975**, *C-24*, 96–98. [\[CrossRef\]](#)
- Lee, D. Effective Gaussian mixture learning for video background subtraction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 827–832. [\[CrossRef\]](#) [\[PubMed\]](#)
- Felzenszwalb, P.F.; Girshick, R.; McAllester, D.; Ramanan, D. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645. [\[CrossRef\]](#) [\[PubMed\]](#)
- Viola, P.; Jones, M. Rapid object detection using a boosted cascade of Simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Kauai, HI, USA, 8–14 December 2001. [\[CrossRef\]](#)
- Wang, X.; Han, T.X.; Yan, S. An hog-LBP human detector with partial occlusion handling. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009. [\[CrossRef\]](#)
- Cao, X.; Wu, C.; Yan, P.; Li, X. Linear SVM classification using boosting hog features for vehicle detection in low-altitude airborne videos. In Proceedings of the 2011 18th IEEE International Conference on Image Processing, Brussels, Belgium, 11–14 September 2011. [\[CrossRef\]](#)
- Wu, S.; Nagahashi, H. Parameterized AdaBoost: Introducing a parameter to speed up the training of real AdaBoost. *IEEE Signal Process. Lett.* **2014**, *21*, 687–691. [\[CrossRef\]](#)
- Kazemi, F.M.; Samadi, S.; Poorreza, H.R.; Akbarzadeh-T, M.-R. Vehicle recognition using curvelet transform and SVM. In Proceedings of the Fourth International Conference on Information Technology (ITNG'07), Las Vegas, NV, USA, 2–4 April 2007. [\[CrossRef\]](#)
- Waranusast, R.; Bundon, N.; Tingtong, V.; Tangnoi, C.; Pattanathaburt, P. Machine vision techniques for motorcycle safety helmet detection. In Proceedings of the 2013 28th International Conference on Image and Vision Computing New Zealand (IVCNZ 2013), Wellington, New Zealand, 27–29 November 2013. [\[CrossRef\]](#)

17. Li, J.; Liu, H.; Wang, T.; Jiang, M.; Wang, S.; Li, K.; Zhao, X. Safety helmet wearing detection based on image processing and machine learning. In Proceedings of the 2017 Ninth International Conference on Advanced Computational Intelligence (ICACI), Doha, Qatar, 4–6 February 2017. [\[CrossRef\]](#)
18. Filatov, N.; Maltseva, N.; Bakhshiev, A. Development of hard hat wearing monitoring system using deep neural networks with high inference speed. In Proceedings of the 2020 International Russian Automation Conference (RusAutoCon), Sochi, Russia, 6–12 September 2020. [\[CrossRef\]](#)
19. Li, Z.; Xie, W.; Zhang, L.; Lu, S.; Xie, L.; Su, H.; Du, W.; Hou, W. Toward efficient safety helmet detection based on Yolov5 with hierarchical positive sample selection and box density filtering. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–14. [\[CrossRef\]](#)
20. Han, K.; Zeng, X. Deep learning-based workers safety helmet wearing detection on construction sites using multi-scale features. *IEEE Access* **2022**, *10*, 718–729. [\[CrossRef\]](#)
21. Zhao, Y.; Cheng, J.; Zhou, W.; Zhang, C.; Pan, X. Infrared pedestrian detection with converted temperature map. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019. [\[CrossRef\]](#)
22. Wang, C.-Y.; Liao, H.-Y.M.; Wu, Y.-H.; Chen, P.-Y.; Hsieh, J.-W.; Yeh, I.-H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020. [\[CrossRef\]](#)
23. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. [\[CrossRef\]](#)
24. Wang, C.; Xu, C.; Wang, C.; Tao, D. Perceptual adversarial networks for image-to-image transformation. *IEEE Trans. Image Process.* **2018**, *27*, 4066–4079. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Yu, K.; Tang, G.; Chen, W.; Hu, S.; Li, Y.; Gong, H. MobileNet-YOLO v5s: An Improved Lightweight Method for Real-Time Detection of Sugarcane Stem Nodes in Complex Natural Environments. *IEEE Access* **2023**, *11*, 104070–104083. [\[CrossRef\]](#)
26. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520. [\[CrossRef\]](#)
27. Liu, L.; Ke, C.; Lin, H.; Xu, H. Research on pedestrian Detection algorithm based on MobileNet-YOLO. *Comput. Intell. Neurosci.* **2022**, *2022*, 8924027. [\[CrossRef\]](#)
28. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [\[CrossRef\]](#)
29. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861. [\[CrossRef\]](#)
30. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021. [\[CrossRef\]](#)
31. Hinton, G.E.; Vinyals, O.; Dean, J.M. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531. [\[CrossRef\]](#)
32. Yang, Q.; Li, F.; Tian, H.; Li, H.; Xu, S.; Fei, J.; Wu, Z.; Feng, Q.; Lu, C. A new knowledge-distillation-based method for detecting conveyor belt defects. *Appl. Sci.* **2022**, *12*, 10051. [\[CrossRef\]](#)
33. Aubard, M.; Antal, L.; Madureira, A.; Ábrahám, E. Knowledge Distillation in YOLOX-ViT for Side-Scan Sonar Object Detection. *arXiv* **2024**, arXiv:2403.09313. [\[CrossRef\]](#)
34. Gochoo, M. Safety Helmet Wearing Dataset. Mendeley Data, V1. 2021. Available online: <https://data.mendeley.com/datasets/9rcv8mm682/1> (accessed on 19 June 2024). [\[CrossRef\]](#)
35. Peng, D.; Sun, Z.; Chen, Z.; Cai, Z.; Xie, L.; Jin, L. Detecting heads using feature refine net and cascaded multi-scale architecture. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 2528–2533. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.