processes

MDPI

Article

A Lightweight Safety Helmet Detection Algorithm Based on Receptive Field Enhancement

Changpeng Ji *, Zhibo Hou and Wei Dai

School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China; houzhibo1112@163.com (Z.H.); daiwei0084@126.com (W.D.)

* Correspondence: ccp@lntu.edu.cn

Abstract: Wearing safety helmets is an important way to ensure the safety of workers' lives. To address the challenges associated with low accuracy, large parameter values, and slow detection speed of existing safety helmet detection algorithms, we propose a receptive field-enhanced lightweight safety helmet detection algorithm called YOLOv5s-CR. First, we use a lightweight backbone, a high-resolution feature fusion network, and a small object detection layer to improve the detection accuracy of small objects while substantially decreasing the model parameters. Next, we embed a coordinate attention mechanism into the feature extraction network to improve the localization accuracy of the detected object. Finally, we propose a new receptive field enhancement module (RFEM) to substitute the SPPF module in the original network, enabling the model to acquire features under multiple receptive fields, thereby enhancing the detection precision of multi-scale objects. Using the Safety Helmet Detection dataset for validation, in contrast to the initial YOLOv5s, the parameters of the improved algorithm were reduced by 62.8% to 2.61 M, and P, R, and mAP$_{0.5}$ were increased by 1.5%, 1.2%, and 2.0%, respectively. The detection speed can reach 149FPS on the RTX3070 GPU, which satisfies the accuracy and real-time requirements for detecting safety helmets.

Keywords: safety helmet detection; YOLOv5s; attention mechanism; lightweight; receptive field enhancement module

check for updates

Citation: Ji, C.; Hou, Z.; Dai, W. A Lightweight Safety Helmet Detection Algorithm Based on Receptive Field Enhancement. *Processes* **2024**, *12*, 1136. https://doi.org/10.3390/pr12061136

Academic Editor: Olympia Roeva

Received: 2 May 2024
Revised: 21 May 2024
Accepted: 28 May 2024
Published: 31 May 2024

1. Introduction

With the continuous development of science and technology, and the acceleration of the industrialization process, all walks of life are developing rapidly, corresponding with a variety of potential security risks that are also emerging. In industrial production, especially in high-risk areas such as construction and manufacturing, accidents may lead to serious personal injury and property losses [1]. To ensure the life safety of workers to the greatest extent, wearing safety helmets has become an essential safety measure [2]. Safety helmets can alleviate the instantaneous impact force to reduce head injury caused by the impact force [3]. However, in many construction sites, there are still many serious production accidents caused by construction personnel not wearing safety helmets, so safety helmet-wearing detection technology has gradually become a research hotspot; the research in this field aims to improve the accuracy, real-time capacity and automation of detection, to inject more advanced and intelligent solutions for site safety management.

The birth of the safety helmet has a history of more than one hundred years. As important personal protective equipment, safety helmets can effectively protect the heads of workers from accidental injury. In 1993, the Golden Gate Bridge project in the United States made it clear that workers must wear safety helmets during construction. Under such a requirement, the number of casualties in the project was reduced by three-quarters compared to other projects in the same period [4]. To urge production personnel to wear protective equipment, the state has also formulated a series of standards to ensure construction safety, and each unit also has its safety norms. The act of not wearing a safety helmet on the construction site has been expressly prohibited. However, most construction workers have high labor intensity, weak safety awareness, and inadequate implementation of the relevant management systems, resulting in the occurrence of the non-wearing of safety helmets, which leads to serious production accidents [5]. Therefore, it is necessary to strengthen the safety education of construction personnel and gradually improve their independent safety awareness, as well as to do an excellent job in detecting helmet-wearing.

The traditional safety helmet-wearing detection method mainly relies on manual inspection, and the site management personnel need to inspect the wearing situation of workers regularly. However, the efficiency of manual inspection is low, especially in large construction sites. It is difficult to achieve comprehensive monitoring, and it is easy to have a problem with missing and false inspections [6]. With the rapid development of deep learning technology, especially the progress in the field of target detection, the safety helmet-wearing detection scheme based on deep learning has gradually emerged. Compared with the traditional way, the deep learning algorithm can learn more complex features through large-scale data to improve the accuracy of the identification of the helmet-wearing situation, avoid the leakage and false detection that easily occurs in the traditional method, and to also ensure uninterrupted detection 24 h a day, so as to improve the efficiency and accuracy of safety production management [7].

Based on the above aspects, the research on safety helmet-wearing detection algorithms has important practical significance in the current social background. With the help of deep learning technology, we can more comprehensively and accurately monitor the safety helmet-wearing situation of site personnel, and provide more intelligent and efficient safety protection for industrial production. The results of this research will not only inject new vitality into industrial safety management, but also promote the application and development of deep learning technology in this field.

At present, object detection algorithms based on deep learning are mainly divided into one-stage algorithms represented by the You Only Look Once (YOLO) series [8–11], and two-stage algorithms represented by the R-CNN series [12–14]. Researchers mostly improve these two types of object detection algorithms to realize the detection of safety helmets. Li introduced PANet with a skip connection and CBAM attention mechanism into the YOLOv3 network, and adopted CIoU loss to improve the average accuracy of safety helmet detection [15]. Zhu combined ResNet101 with FPN to improve the feature extraction network of Faster R-CNN and adjust the size of the prior box. The improved model has a certain generalization ability and robustness [16]. Ding added the ECA-Net attention mechanism to the neck of YOLOX, and used CIoU to calculate the loss. The improved algorithm has a high accuracy while ensuring real-time detection [17]. Liu introduced the RepVGG module to the lightweight backbone network of YOLOv5s, used Soft-NMS to reduce the missed detection rate of occlusion targets, and used the mix-up method to enhance and expand the dataset, which provided a practical reference for substation safety helmet detection [18]. Zhao introduced the transformer self-attention module in the backbone network of YOLOv5, and used the Ghost module and EIoU Loss to reduce the model's parameters and improve the detection accuracy. The algorithm is more effective than the original algorithm at detecting safety helmets [19]. Deng proposed a safety helmet detection method based on improved YOLOv4. By collecting a self-made data set of on-site construction site videos, the K-means algorithm clusters the data set, and a multi-scale training strategy is used in the network training process to improve the model's adaptability to different detection scales. The model mAP value reached 92.89%. The detection speed reached 15 f/s [20].

Table 1 shows the comparison of the above safety helmet detection algorithms, from which it can be seen that although the above safety helmet detection algorithms improve the detection performance, they still have problems, such as large model parameters and a slow detection speed.

To solve the above problems, we take the YOLOv5s algorithm as the baseline to make improvements. The main contributions of this paper are as follows:

(1) Redesign the overall structure of the network, thereby significantly reducing network parameters and improving the detection accuracy of small objects.
(2) The coordinate attention mechanism is added to the feature extraction part of the network to enhance the ability of the network to locate the object.
(3) Use the Max-Pooling module and three dilated convolutions with different dilation rates to design a new Receptive Field Enhancement Module (RFEM) to make the network have richer receptive fields and improve the detection accuracy of multi-scale objects.

Table 1. The comparison of different safety helmet detection algorithms.

Model	Improvements	$\mathbf{mAP_{0.5}}$	Parameters	FPS
YOLO3-CPANet-CIoU [15]	CPANet, CIoU	0.920	298 M	21
Improved Faster R-CNN [16]	ResNet101, Anchor	0.909	-	-
Improved YOLOX [17]	ECA-Net, CIoU	0.917	-	71.9
Improved YOLOv5s [18]	RepVGG, Soft-NMS, Mixup	0.804	10.6 M	83.3
Improved YOLOv5s [19]	Detect layer, Self-attention, Ghost module, EIoU	0.960	13 M	-
Improved YOLOv4 [20]	Anchor, Multi-scale training strategy	0.929	-	15

2. Improved YOLOv5s Model

2.1. Network Structure Improvements

With the deepening of the layers of the original YOLOv5s backbone network, the feature map's size continues to shrink, and the output channels multiply, which leads to the original smaller objects occupying too few pixels in the feature map, which affects the precision in detecting small objects, and leads to the excessive amounts of parameters in the last few layers of the network. Therefore, we redesigned the network structure of YOLOv5s. The convolution layer and down-sampling layer at the tail towards the end of the backbone network were removed so that the down-sampling rate of the network was reduced from 32 to 16, significantly reducing the network parameters while retaining more detailed information. The improved high-resolution feature fusion network obtains 40 × 40, 80 × 80, and 160 × 160 high-resolution feature maps after the 40 × 40 feature map is passed through FPN [21] and PAN [22]. It can retain richer semantic information than the original YOLOv5s feature fusion network. The large object detection layer with a size of 20 × 20 in the original YOLOv5s detection head is removed, and the small object detection layer with a size of 160 × 160 is added to obtain the YOLOv5s-r model. Its network structure is illustrated in Figure 1 (consider an input image of size 640 × 640). Since the 20 × 20 detection layer has a larger receptive field, and is used to detect large-size targets, according to the proportion of the width and height of the helmet target in the training dataset in Figure 2, it can be seen that the width and height of most helmet targets are within 0.2 of the whole picture size, which belongs to small size targets. Therefore, removing this detection layer will not affect the detection accuracy of the helmet, and can greatly reduce the number of network parameters, as shown in Table 2.

Table 2. Comparison of network parameters before and after.

Model	Parameters	GFLOPs	Head Size
YOLOv5s	7.02 M	15.8	20/40/80
YOLOv5s-r	2.03 M	13.9	40/80/160

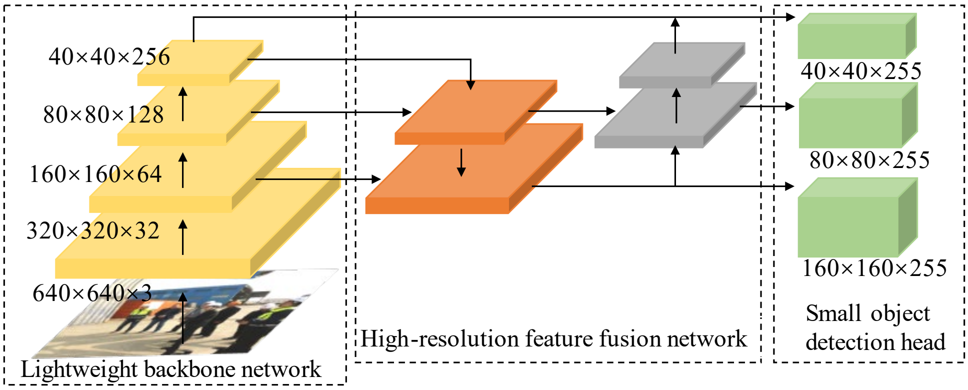

40×40×256
80×80×128
160×160×64
320×320×32
640×640×3
Lightweight backbone network
High-resolution feature fusion network
40×40×255
80×80×255
160×160×255
Small object detection head

Figure 1. YOLOv5s-r network structure. It consists of three parts: a lightweight backbone network, a high-resolution feature fusion network, and a small object detection head.

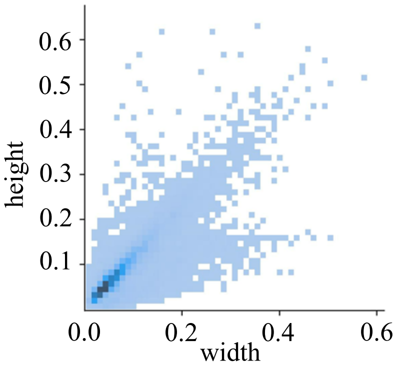


Figure 2. Training dataset label width and height distribution. Where the color is darker, the samples are more densely distributed.

2.2. Coordinate Attention Mechanism

By adaptively creating different weight parameters for the information in the image, the attention mechanism can strengthen the attention to key information while weakening the attention to useless information [23], thereby improving the neural network's performance. For deep learning, usually the deeper the model and the more parameters, the stronger the learning ability of the model, but the more information and computation are required, which can easily lead to information overload. The attention mechanism is introduced to make the model focus on important information, reduce the processing of redundant information, and improve the model's estimation speed and accuracy [24]. The attention mechanism is more like a weight vector, giving a higher weight to important content. Common attention mechanisms include SE [25], ECA [26], CBAM [27], etc. However, the general attention mechanism usually ignores the location information, resulting in the network being unable to learn the object's coordinate information. At the same time, the increase in parameters brought by most attention mechanisms greatly impacts lightweight networks. The coordinate attention mechanism is an attention mechanism that almost does not increase network parameters [28]. It enhances the representation of the region of interest by incorporating location information into the channel attention, and its structure is illustrated in Figure 3.

The coordinate attention mechanism's operation is as follows: for the input x of size $C \times H \times W$, a pooling kernel with dimension $(H,1)$ or $(1,W)$ is first used to encode each channel along the horizontal and vertical coordinate directions. The output of the c-channel with height h can be stated as:

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

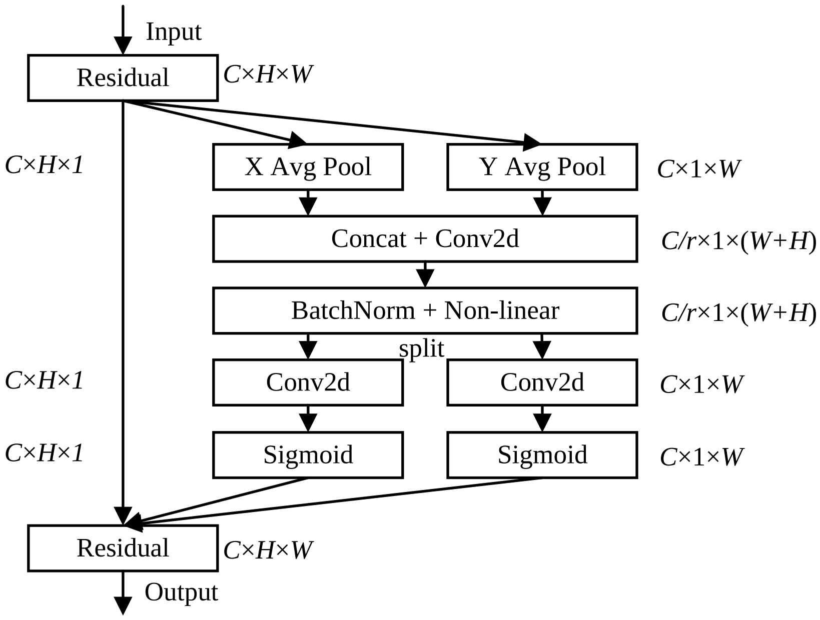


Figure 3. Coordinate attention mechanism. It is mainly composed of horizontal and vertical Avg-Pooling, Conv2d, BatcNorm, and non-linear layers, which can enhance the ability of the network to locate the target to be detected.

Likewise, the output of the c-channel with width w can be expressed as:

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

Then, the encoded output is first concatenated and then transformed by a 1 × 1 convolution transformation F_1:

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

where [·, ·] represents the concatenation operation in the spatial dimension, δ is the hard Swish activation function, and $f \in R^{C/r \times (H+W)}$ is the intermediate feature when spatial information is encoded horizontally and vertically. Then, f is divided in the spatial dimension to obtain two independent tensors: $f^h \in R^{C/r \times H}$ and $f^w \in R^{C/r \times W}$. Finally, two 1 × 1 convolution transforms F_h and F_w are used to adjust f^h and f^w to the same channel:

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

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

where σ represents the sigmoid activation function, while g^h and g^w are the attention weights. Finally, the output y of the attention mechanism is achieved by multiplying the input feature map with the attention weights:

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

The CA attention mechanism is a plug-and-play module, which is generally embedded into the feature extraction part of the network to improve the feature extraction effect of the network on the key part. We propose two ways of embedding:

(a) Integrate CA attention into each C3 module of the backbone network;
(b) Add the CA attention mechanism to the end of the backbone network.

After experiments, the second method has higher detection accuracy, and is compared with the detection accuracy of three attention mechanisms where SE, CBAM, and ECA (as shown in Table 3) were inserted. It can be found that inserting the CA attention mechanism in the b mode can effectively improve the detection accuracy, and $\mathrm{mAP}_{0.5}$ reaches the highest value (93.1%).

Table 3. Comparison of four different attention mechanisms.

Model	P	R	$\mathrm{mAP}_{0.5}$	Parameters	GFLOPs
YOLOv5s	0.917	0.877	0.920	7.02 M	15.8
+SE	0.925	0.865	0.922	7.05 M	15.8
+CBAM	0.924	0.879	0.925	7.06 M	15.9
+ECA	0.923	0.861	0.914	7.02 M	15.8
+CA(a)	0.924	0.864	0.923	6.71 M	15.2
+CA(b)	0.925	0.882	0.931	7.03 M	15.8

The heatmap effect after adding various attention mechanisms is shown in Figure 4. It can be easily seen from the visualization results that the heat map after adding the coordinate attention mechanism is brighter on the safety helmet target, indicating that the coordinate attention mechanism is more accurate for the positioning of the safety helmet target, which can effectively avoids the interference caused by the complex environment.

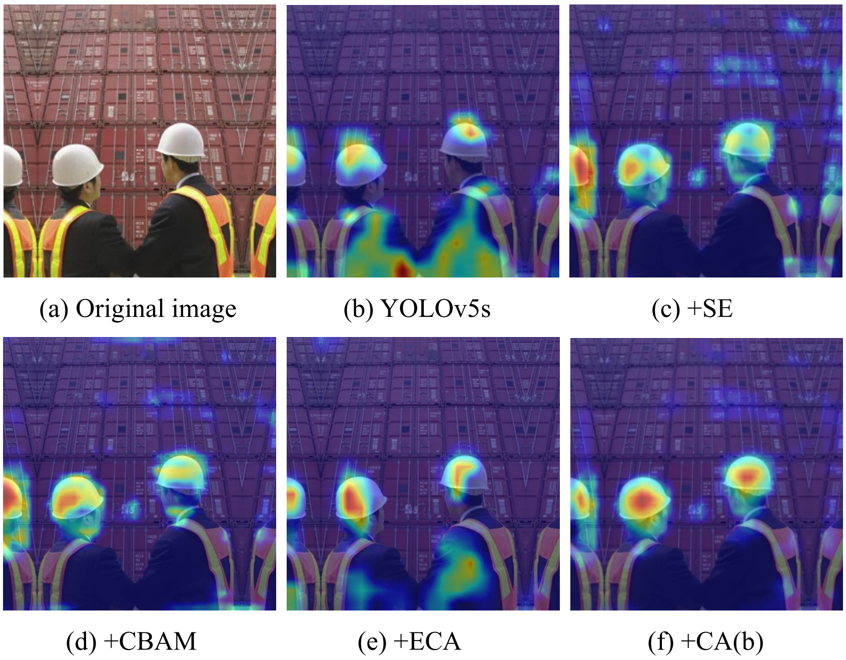

(a) Original image
(b) YOLOv5s
(c) +SE
(d) +CBAM
(e) +ECA
(f) +CA(b)

Figure 4. Comparison of heatmaps of various attention mechanisms. The more concentrated the dark color is on the safety helmet targets, the better the effect of the attention mechanism is.

2.3. *Improved Receptive Field Enhancement Module*

In convolutional neural networks, the receptive field can be defined as the size of the mapped area of each pixel in the output feature map of each layer of the network on the input feature map [29]. The richer the receptive field is, the more global and local information can be obtained.

Spatial pyramid pooling [30] is a commonly used receptive field enhancement method. It can avoid image distortion caused by image area cropping and zoom operation. In YOLOv5s, the structure of SPP is improved, and a fast version of spatial pyramid pooling–SPPF, is proposed; its structure is shown in Figure 5. CBS modules are added before and after SPP, the parallel structure of SPP is changed to a serial structure, and the pooling kernel of max pooling is uniformly set to 5 × 5. In total, SPPF can obtain receptive fields of 5 × 5, 9 × 9, and 13 × 13 scales. In addition, the commonly used receptive field enhancement methods include dilated convolution, ASPP [31], RFB [32], PPM [33], etc.

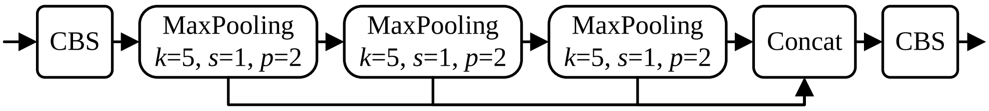


Figure 5. SPPF structure. It consists of three Max-Pooling modules and two CBS modules.

In deep learning, the convolutional layer is the basic module in the convolutional neural network, whose role is to extract features from the input data. The convolution layer can be regarded as a sliding window operation, and the convolution operation is performed between the convolution kernel and the input data to obtain the output data. In a traditional convolution operation, each element in the convolution kernel is multiplied once with each element in the input data, and these results are then added to produce an element in the output data. Therefore, if the convolution kernel size is 3 × 3, then a convolution operation will involve nine input data elements. If the convolution kernel size increases, the amount of computation required for the convolution operation increases accordingly. In practical applications, the data sets that need to be processed are usually large. If the size of the convolution kernel is too large, the computational amount required for the convolution operation will be too large, which will affect the training and prediction efficiency of the model.

Dilated convolution is a method to solve the above problem. Dilated convolution can increase the network receptive field without reducing the resolution, increasing parameters, and increasing the depth of the network [34]. Therefore, the number of parameters in the convolution layer can be reduced by dilated convolution, thus reducing the size and complexity of the model and improving the efficiency of the model. This is done by introducing dilation into the convolution kernel. In dilated convolution, every element in the convolution kernel is not multiplied by every element in the input data, but by elements separated by a certain distance in the input data. This spacing distance is called the dilation rate and can also be understood as the step size of the element in the convolution kernel [35]. In dilated convolution, when the dilated rate is 1, it is a traditional convolution operation. As an essential technique in convolutional neural networks, dilated convolution has been widely used in image processing and other fields.

The structure of dilated convolution is shown in Figure 6. Among them, the outer-most box represents the input image, the black dot represents the convolution kernel, and the blue area represents the convolution receptive field. Figure 6a is an ordinary convolution process, the dilation rate is 1, and the convolution receptive field is 3; Figure 6b is a dilated convolution with dilation rate 2, and the receptive field after convolution is 5; Figure 6c is a dilated convolution with dilation rate 3, and the receptive field after convolution is 7.

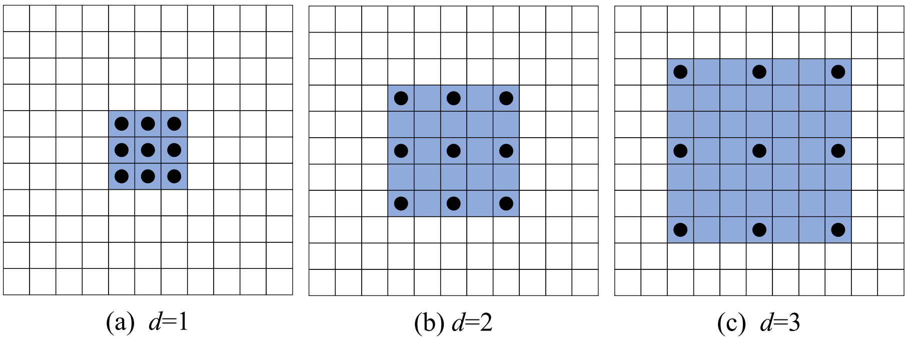


Figure 6. Dilated convolution structure. The blue area represents the convolution receptive field.

As shown in Equation (7), for a convolution kernel of size $k \times k$, a, a dilated convolution of size $k_i \times k_i$ can be obtained after the dilation operation with dilation rate d :

$$k_i = k + (k-1) \times (d-1) \tag{7}$$

For example, in a traditional convolution operation, when the convolution kernel size is 3×3 (that is, $k = 3, d = 1$), the effective receptive field is 3×3. When the dilation rate is 2, multiplication is performed between each element in the convolution kernel and the input data by one pixel, and the size of the effective receptive field of each element in the convolution kernel is 5×5, which is larger than the effective receptive field of the traditional convolution operation. Similarly, when the dilation rate is 3, the effective receptive field size of each element in the convolution kernel is 7×7, and so on. Therefore, by adjusting the dilation rate, the effective receptive field of the convolution kernel can be expanded, and the receptive field size of the model can be improved; that is, multi-scale information can be obtained, and features in the input data can be better captured. In practice, the size of the target under detection is usually very rich; at this time, only relying on the SPPF module will not be able to fully deal with the multi-scale changes in the detection target. Therefore, we use the Max-Pooling module and dilated convolution to design a new receptive field enhancement module called RFEM.

Since the resolution of the output feature map of the YOLOv5s-r feature extraction network is 26×26 (the input is 416×416), we, respectively, use dilated convolution with 3, 8, and 13 dilation rates. According to Equation (7), they can obtain receptive fields with scales of 7×7, 17×17, and 27×27, which can adapt to the size of the output feature map.

As shown in Figure 7, The process of the RFEM module is as follows: the input feature map M_0 is passed through three dilated convolutions D_i with dilation rate i to obtain their output N_i:

$$N_i = D_i(M_0) \; (i = 3, 8, 13) \tag{8}$$

At the same time, the input feature map M_0 is passed through three Max-Pooling modules MP in series to obtain their output M_i :

$$M_j = MP(M_{j-1}) \; (j = 1, 2, 3) \tag{9}$$

Finally, N_i, M_i, and the feature map obtained after 1×1 convolution F_1 are concatenated, and then the channel is adjusted by 1×1 convolution F_2 as the output Out of the whole module:

$$Out = F_2([F_1(M_0), N_i, M_j]) \; (i = 3, 8, 13; j = 1, 2, 3) \tag{10}$$

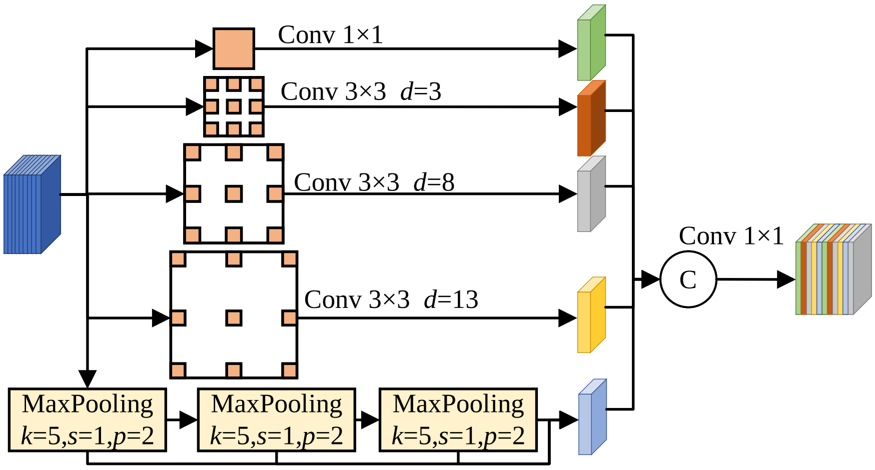

Conv 1×1
Conv 3×3 *d*=3
Conv 3×3 *d*=8
Conv 3×3 *d*=13
Conv 1×1
C
MaxPooling *k*=5,*s*=1,*p*=2
MaxPooling *k*=5,*s*=1,*p*=2
MaxPooling *k*=5,*s*=1,*p*=2

Figure 7. Receptive field enhancement module structure. It is mainly composed of three Max-Pooling modules and dilated convolutions with dilated rates of 3, 8, and 13.

In this way, the RFEM module can obtain receptive fields of six different scales: 5 × 5, 9 × 9, 13 × 13, 7 × 7, 17 × 17, and 27 × 27, so that the network can cope with multi-scale changes in the target under detection and enhance the precision of detection. The coordinate attention mechanism is inserted into the tail of the YOLOv5s-r backbone, and the RFEM module is used to replace the original SPPF model to obtain the improved final model YOLOv5s-CR, whose structure is shown in Figure 8.

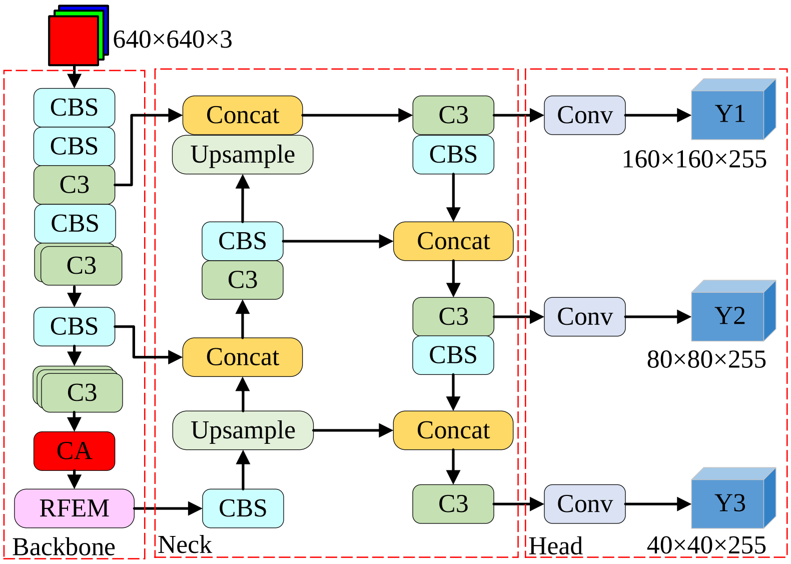


Figure 8. YOLOv5s-CR network structure. Its backbone is a more lightweight backbone embedded with coordinate attention mechanism and RFEM, and it also has a small object detection layer.

3. Experimental Results and Analysis

3.1. Datasets and Experimental Environments

Using Safety Helmet Detection, a public dataset of safety helmet detection on Kaggle (https://www.kaggle.com/), the dataset contains 5000 pictures of workers wearing safety helmets and not wearing safety helmets in various scenarios and randomly divides them into training sets, validation sets, and test sets in the ratio of 7:1.5:1.5.

To evaluate the effectiveness of the YOLOv5s-CR algorithm, the following evaluation indicators are selected: Precision, Recall, mean Average Precision, Parameters, GFLOPs, and FPS. The system used in the experimental equipment was Windows 10, the CPU was an Intel (R) Core(TM) i7-11800H@ 2.30 GHz (MSI Technology Co., Ltd., Shanghai, China), the memory was 16 GB, the GPU was an NVIDIA RTX3070(MSI, China), and the deep learning framework was Pytorch v1.11.

3.2. Ablation Experiments

To verify the effectiveness of the improvements in this paper, we use the 6.2 version of the YOLOv5s model as a baseline for ablation experiments. The input image has dimensions of 416 × 416, and the batch size is configured as 32, and each model is trained for 150 epochs. Experiment with each improvement step and record the results. The experimental outcomes are presented in Table 4.

Table 4. Ablation experiment.

Model	**P**	**R**	**mAP$_{0.5}$**	**Parameters**	**GFLOPs**	**FPS**
YOLOv5s	0.917	0.877	0.920	7.02 M	15.8	138
YOLOv5s-r	0.926	0.878	0.932	2.03 M	13.9	158
YOLOv5s-r+CA	0.928	0.883	0.934	2.04 M	14	153
YOLOv5s-r+RFEM	0.928	0.888	0.936	2.61 M	15.8	156
YOLOv5s-CR	0.932	0.889	0.94	2.61 M	15.8	149

Table 4 reveals that, in comparison to the initial YOLOv5s version, the P, R, and $mAP_{0.5}$ of the YOLOv5s-r after the redesigned network structure are increased by 0.9%, 0.1%, and 1.2%, respectively, under the condition that the parameters are reduced by 71% and the GFLOPs are reduced by 12%. Using YOLOv5s-r as the baseline, after adding the coordinate attention mechanism, although the parameters and GFLOPs are slightly increased, P, R, and $mAP_{0.5}$ are increased by 0.2%, 0.5%, and 0.2%, respectively. After introducing the RFEM module to replace the SPPF module in YOLOv5s-r, P, R, and $mAP_{0.5}$ are increased by 0.2%, 1.0%, and 0.4%, respectively. The above two improvements are integrated into YOLOv5s-r to obtain the improved final model, YOLOv5s-CR. Compared with the baseline YOLOv5s, the P, R, and $mAP_{0.5}$ of YOLOv5s-CR are increased by 1.5%, 1.2%, and 2.0%, respectively, and the detection speed is increased by 11FPS under the condition that the parameters are reduced by 62.8% and the GFLOPs are almost unchanged.

Figure 9a is the P-R curve of YOLOv5s, and Figure 9b is the P-R curve of YOLOv5s-CR. As can be seen from the figure, the area surrounded by the P-R curve of YOLOv5s-CR is larger, which shows that the accuracy, recall rate, and robustness of the model are better. Moreover, the $mAP_{0.5}$ of YOLOv5s-CR reaches a higher value of 0.940, indicating that the improved algorithm has better detection performance.

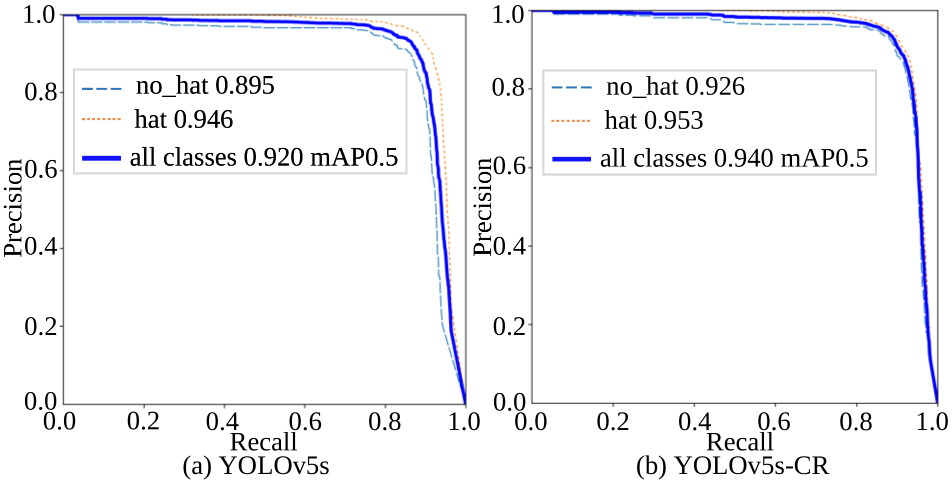
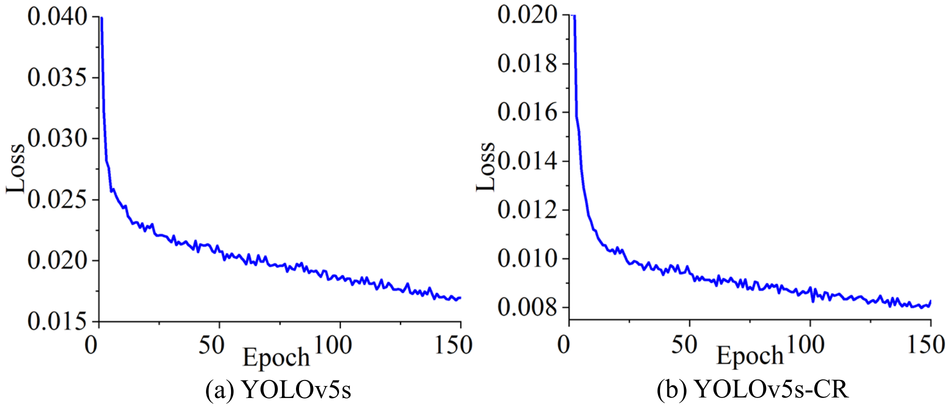


Figure 9. YOLOv5s Precision-Recall curve and YOLOv5s-CR Precision-Recall curve. The larger area enclosed by the curve indicates the better robustness of the model.

Figure 10 shows the change in the confidence loss of YOLOv5s and YOLOv5s-CR during training. It can be seen from the figure that the loss value of YOLOv5s-CR decreases faster in the training process; that is, the model converges faster. The loss of YOLOv5s drops from 0.04 to 0.02 at about 70 epochs. The confidence loss of YOLOv5s-CR dropped to 0.02 in the first few epochs, and to around 0.08 at 150 epochs, which is lower than that of YOLOv5s, indicating that the model can achieve higher detection accuracy.

In summary, we prove the effectiveness of the improvements, and the improved YOLOv5s-CR model has better performance.

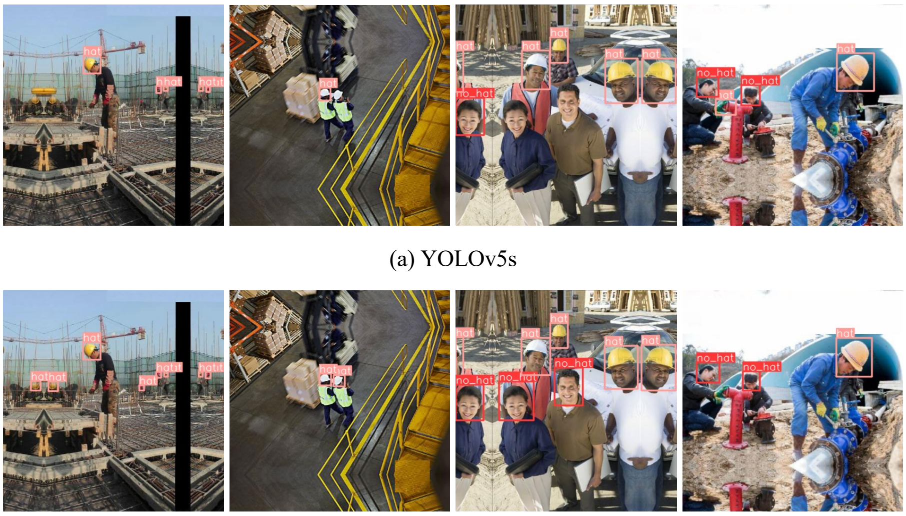
Figure 11 shows the detection effect comparison of YOLOv5s and YOLOv5s-CR. In the first picture with occlusion, YOLOv5s-CR correctly detected all targets, while YOLOv5s had two missed targets. The second and third pictures are from a top-down angle; YOLOv5s-CR can correctly detect all the targets, while the YOLOv5s still have undetected phenomena. In the last picture, the YOLOv5s have a false detection, while YOLOv5s-CR has no false detections. Therefore, in the actual situation, the detection effect of the improved YOLOv5s-CR algorithm is better than that of the original YOLOv5s algorithm to a large extent.



(a) YOLOv5s

(b) YOLOv5s-CR

Figure 10. YOLOv5s and YOLOv5s-CR loss curve. It is clear that YOLOv5s-CR converges faster and ends up with a lower loss.



(a) YOLOv5s

(b) YOLOv5s-CR

Figure 11. Comparison of detection results between YOLOv5s and YOLOv5s-CR. Obviously, the improved YOLOv5s-CR can correctly detect all targets.

3.3. Comparative Experiments

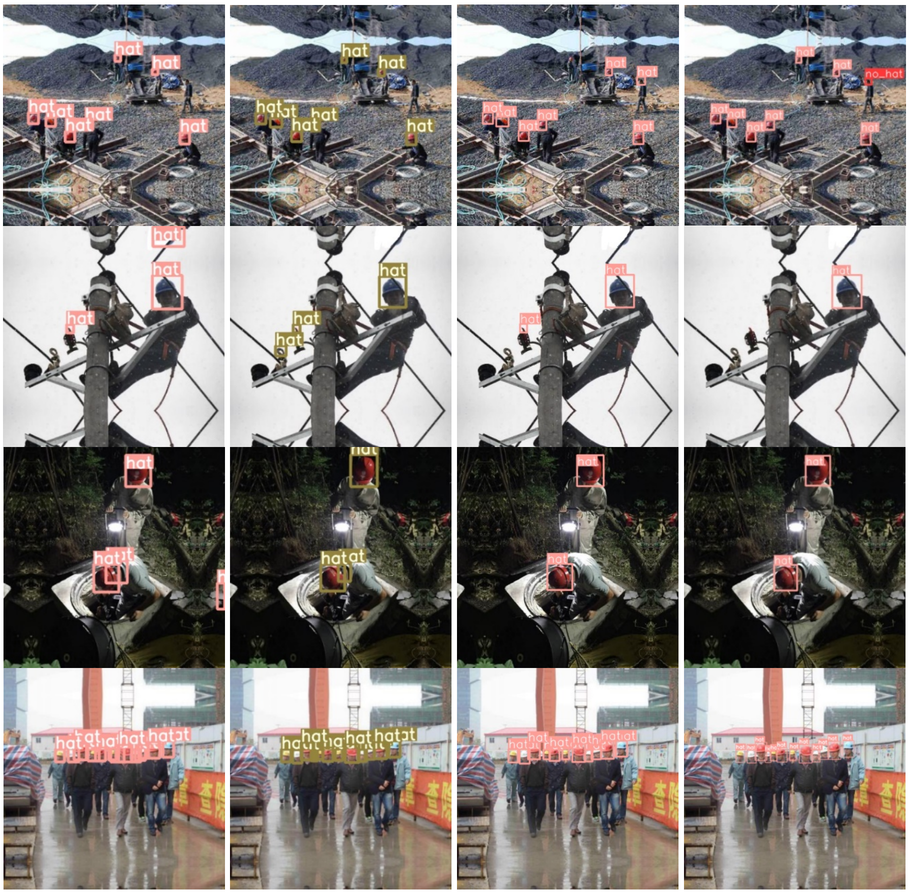
To further validate the effectiveness of the YOLOv5s-CR algorithm in helmet detection, we select some representative object detection algorithms for comparative experiments. Under identical experimental conditions, the algorithm in this paper is compared with SSD, Faster-RCNN, YOLOv3-tiny, YOLOv7-tiny, and YOLOv8n object detection algorithms using the same data set; data set division method and model training strategy. The results of the experiments are presented in Table 5.

Table 5. Comparative experiment.

Model	P	R	$mAP_{0.5}$	Parameters	GFLOPs	FPS
SSD	0.592	0.613	0.592	91.1 M	26.3	64
Faster-RCNN	0.752	0.822	0.752	108 M	137	15
YOLOv3-tiny	0.888	0.81	0.87	8.67 M	12.9	178
YOLOv7-tiny	0.922	0.88	0.932	6.01 M	13.9	112
YOLOv8n	0.926	0.876	0.931	3.01 M	9.3	107
YOLOv5s-CR	0.932	0.889	0.94	2.61 M	15.8	149

Since most of the safety helmet objects in this experimental dataset are small objects and dense objects, Table 5 illustrates that SSD and Faster-RCNN algorithms have poor detection effects and slow detection speed; although the YOLOv3-tiny algorithm has a fast detection speed, the detection effect is poor. The YOLOv7-tiny and YOLOv8n algorithms have high detection accuracy, but the detection speed is slow. YOLOv5s-CR has greatly improved detection accuracy and recall when the parameters are only 2.61 M. In comparison to YOLOv3-tiny, the detection performance was improved by 4.4%, 7.9%, and 7%, respectively. Compared with YOLOv7-tiny, it was improved by 1.0%, 0.9%, and 0.8%, respectively. In comparison to YOLOv8n, it was improved by 0.6%, 1.3%, and 0.9%, respectively. Simultaneously, the speed of detection is increased by 37FPS and 42FPS, respectively, compared with YOLOv7-tiny and YOLOv8n. It is evident that although the GFLOPs of the proposed algorithm are slightly increased compared with other algorithms, the parameters are greatly reduced, the detection performance is better, and it can satisfy the requirements for detecting safety helmets.

Figure 12 shows the comparison of the detection effects of the four YOLO algorithms in the above comparison models in different scenarios. The first figure shows the scene containing long-distance small objects; it is evident that YOLOv3-tiny, YOLOv7-tiny, and YOLOv8n all have missed detection phenomenon, while YOLOv5s-CR has not missed detection due to the small object detection layer. The second picture shows the scene of working at height from an elevation perspective. YOLOv3-tiny, YOLOv7-tiny, and YOLOv8n all have false detections. The third picture is the scene when the light intensity is weak, YOLOv3-tiny and YOLOv7-tiny also have false detections, and YOLOv5s-CR has no false detections in both cases. The fourth figure shows the object dense scene, YOLOv3-tiny, and YOLOv7-tiny both have different degrees of false detection, and YOLOv5s-CR can correctly detect all objects. Evidently, the algorithm in this paper has good robustness in various scenarios.



(a) YOLOv3-tiny (b) YOLOv7-tiny (c) YOLOv8n (d) YOLOv5s-CR

Figure 12. Comparison of detection results between YOLOv3-tiny, YOLOv7-tiny, YOLOv8n, and YOLOv5s-CR. YOLOv5s-CR has the best detection effect among the four algorithms.

4. Conclusions

Taking version 6.2 of the YOLOv5s algorithm as the baseline, we propose a lightweight YOLOv5s-CR safety helmet detection algorithm. At the same time, we propose a new receptive field enhancement module called RFEM, which enables the network to have richer receptive fields so that the network can cope with multi-scale changes in objects. Using the public dataset Safety Helmet Detection for verification, in contrast to the initial YOLOv5s, the parameters of YOLOv5s-CR are reduced by 62.8%, only 2.61 M, while $\text{mAP}_{0.5}$ is increased by 2.0%, and the detection speed can reach 149FPS. Compared with the commonly used object detection algorithms, the highest mAP can be achieved while having the fewest parameters. Experiments show that our algorithm can meet the needs of safety helmet detection tasks and has strong practicability.

Author Contributions: Conceptualization, C.J. and Z.H.; methodology, C.J. and Z.H.; software, Z.H.; validation, C.J. and Z.H.; investigation, Z.H.; resources, C.J. and W.D.; data curation, Z.H.; writing—original draft preparation, Z.H. and W.D.; writing—review and editing, Z.H. and W.D.; visualization, Z.H.; supervision, C.J. and W.D.; project administration, W.D.; funding acquisition, C.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Research on Key Methods of Internet of Things Security Situational Awareness in Open Pit Mines (LJKMZ20220677).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

References

1. Tang, K.; Chen, L.; Zhang, J.Z.; Zhang, S.Y. Statistical analysis and countermeasures of production safety accidents in construction industry in China. *Constr. Saf.* **2020**, *35*, 40–43.
2. Wang, J.; Kang, W.; Zhou, W.; Huang, F.; Tao, X.; Wu, Q. Helmet Detection Algorithm Based on the Improved YOLOv5 and Dynamic Anchor Box Matching. In Proceedings of the 2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT), Chongqing, China, 22–24 November 2021; pp. 79–83.
3. Li, Q.Y.; Wang, J.B.; Wang, H.W.; Zeng, H.D.; Yang, X.X.; Liu, Y.F. Industrial safety helmet impact resistance study. *J. Saf. Sci. Technol.* **2021**, *17*, 182–186.
4. Xu, D.M.; Wang, D.M. Technology and management innovation in the construction of San Francisco Golden Gate Bridge. *J. Eng. Stud.* **2015**, *7*, 106–115. [CrossRef]
5. Ding, L.; Miu, X.R.; Hu, J.F.; Zhao, Z.P.; Zhang, X.J. Improve YOLOv8s and DeepSORT for miners' hat band detection and personnel tracking. *Comput. Eng. Appl.* **2024**, *60*, 328–335.
6. Zhou, F.; Zhao, H.; Nie, Z. Safety helmet detection based on YOLOv5. In Proceedings of the 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA), Shenyang, China, 22–24 January 2021; pp. 6–11.
7. Qi, Z.Z.; Xu, Y.X. Research on safety helmet wearing detection based on improved YOLOv5s algorithm. *Comput. Eng. Appl.* **2023**, *59*, 176–183.
8. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
9. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
10. Farhadi, A.; Redmon, J. Yolov3: An incremental improvement. In Proceedings of the Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; Springer: Berlin/Heidelberg, Germany, 2018; Volume 1804, pp. 1–6.
11. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–26 June 2023; pp. 7464–7475.
12. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
13. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
14. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv* **2015**, arXiv:1506.01497.

15. Li, T.Y.; Li, D.; Chen, M.J.; Wu, H.; Liu, Y.Q. A High-precision convolutional Neural Network safety Helmet Detection method. *Chin. J. Liq. Cryst. Displays* **2021**, *36*, 1018–1026. [CrossRef]
16. Zhu, Y.H.; Du, J.Y.; Liu, Y.; Jie, Y.P. Research on small object helmet detection algorithm based on improved Faster R-CNN. *Pract. Electron.* **2022**, *30*, 64–66+83.
17. Ding, T.; Chen, X.Y.; Zhou, Q.; Xiao, H.L. Real-time detection of helmet wear based on improved YOLOX. *Electron. Meas. Technol.* **2022**, *45*, 72–78.
18. Liu, Z.X.; Zhang, N.; Lian, T.; Ma, J.; Zhao, Y.; Ni, W. A helmet detection method for lightweight networks. *Meas. Control. Technol.* **2022**, *41*, 16–21+53.
19. Zhao, L.; Zhang, D.; Liu, Y.; Guo, J.; Shi, Z. Improved YOLOv5s Network for Multi-scale safety Helmet Detection. In Proceedings of the 2022 11th International Conference on Communications, Circuits and Systems (ICCCAS), Singapore, 13–15 May 2022; pp. 262–266.
20. Benyang, D.; Xiaochun, L.; Miao, Y. Safety helmet detection method based on YOLOv4. In Proceedings of the 2020 16th International Conference on Computational Intelligence and Security (CIS), Nanning, China, 27–30 November 2020; pp. 155–158.
21. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
22. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
23. Lei, J.Y.; Ye, S.; Xia, M.; Zheng, L.; Zou, J.L. Grape leaf disease detection based on improved YOLOv4. *J. South-Cent. Univ. Natl.* **2022**, *41*, 712–719.
24. Li, M.; Jia, X.R.; Li, S.A. Image superresolution reconstruction algorithm based on mixed attention mechanism. *Comput. Simul.* **2023**, *40*, 236–241.
25. 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.
26. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
27. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
28. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
29. Liu, H.; Wang, X.L. Remote sensing image segmentation model based on adaptive receptive field mechanism. *J. Image Graph.* **2021**, *26*, 464–474.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
31. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
32. Liu, S.; Huang, D.; Wang, Y. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.
33. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
34. Sun, H.; Fang, S.L.; Dan, Z.P.; Ren, D.; Yu, M.; Sun, S.F. Hierarchical feature interaction and enhanced receptive field dual-branch remote sensing image dehazing network. *J. Remote Sens.* **2023**, *27*, 2831–2846.
35. Wang, Z.; Xia, F.; Zhang, C. FD_YOLOX: An improved YOLOX object detection algorithm based on dilated convolution. In Proceedings of the 2023 IEEE 18th Conference on Industrial Electronics and Applications (ICIEA), Ningbo, China, 18–22 August 2023; pp. 1263–1268.

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.