applied sciences

MDPI

Article

Helmet Wearing Detection Algorithm Based on YOLOv5s-FCW

Jingyi Liu [1], Hanquan Zhang [2], Gang Lv [2], Panpan Liu [2], Shiming Hu [2] and Dong Xiao [2,*]

[1] College of Sciences, Northeastern University, Shenyang 110819, China; liujingyi@mail.neu.edu.cn
[2] College of Information Science and Engineering, Northeastern University, Shenyang 110819, China; hanquanzhang@yeah.net (H.Z.); 2200872@stu.neu.edu.cn (G.L.); l2206979789@163.com (P.L.); 2370724@stu.neu.edu.cn (S.H.)
* Correspondence: xiaodong@ise.neu.edu.cn

Abstract: An enhanced algorithm, YOLOv5s-FCW, is put forward in this study to tackle the problems that exist in the current helmet detection (HD) methods. These issues include having too many parameters, a complex network, and large computation requirements, making it unsuitable for deployment on embedded and other devices. Additionally, existing algorithms struggle with detecting small targets and do not achieve high enough recognition accuracy. Firstly, the YOLOv5s backbone network is replaced by FasterNet for feature extraction (FE), which reduces the number of parameters and computational effort in the network. Secondly, a convolutional block attention module (CBAM) is added to the YOLOv5 model to improve the detection model's ability to detect small objects such as helmets by increasing its attention to them. Finally, to enhance model convergence, the WIoU_Loss loss function is adopted instead of the GIoU_Loss loss function. As reported by the experimental results, the YOLOv5s-FCW algorithm proposed in this study has improved accuracy by 4.6% compared to the baseline algorithm. The proposed approach not only enhances detection concerning small and obscured targets but also reduces computation for the YOLOv5s model by 20%, thereby decreasing the hardware cost while maintaining a higher average accuracy regarding detection.

Keywords: lightweighting; target detection; improved YOLOv5; attention mechanism

Citation: Liu, J.; Zhang, H.; Lv, G.; Liu, P.; Hu, S.; Xiao, D. Helmet Wearing Detection Algorithm Based on YOLOv5s-FCW. *Appl. Sci.* **2024**, *14*, 9741. https://doi.org/10.3390/app14219741

Academic Editor: Sungho Kim

Received: 15 September 2024
Revised: 15 October 2024
Accepted: 23 October 2024
Published: 24 October 2024

1. Introduction

The construction industry is known to have a significantly higher accident rate than most other industries [1]. Within the construction industry, head injuries result in the highest disability rate, making it crucial to prioritize measures that reduce the likelihood of such injuries. One effective measure to enhance workplace safety is the mandatory utilization of appropriate protective equipment, such as helmets. These helmets serve to buffer and disperse the kinetic energy from falling objects upon impact, thereby mitigating the risk of direct head injuries. To achieve this goal, enhancing the supervision of helmet usage can be instrumental in enhancing the protection of construction workers and reducing the likelihood of workplace mishaps. During the initial stages of construction site supervision, helmet-wearing compliance is typically assessed through manual observation, which involves inspecting individual workers to determine whether they are wearing helmets. However, due to the high mobility of construction personnel, it is often challenging for inspectors to fully monitor the safety compliance of every unit. This approach may lead to increased labor costs and could potentially heighten the risk of safety hazards. Therefore, the development of advanced methods to detect helmet-wearing compliance among construction personnel is crucial and holds significant potential for improving overall safety and reducing workplace accidents.

Machine learning (ML) techniques are increasingly being used to monitor safety equipment in the workplace, especially the use of helmets. Currently, there are two main algorithms used for this kind of monitoring: an image recognition method based on

traditional ML and an image recognition algorithm based on deep learning (DL). Both methods are dedicated to improving the safety standards of working environments and effectively preventing safety accidents. The algorithms commonly used to monitor the wearing of safety helmets (SH), on the other hand, focus on identification by analyzing color and shape characteristics. Liu and Ye [2] proposed a method for face localization through skin color by intercepting the square region of the face and using the extracted Hu moments as features, training these Hu moment features using SVM, and finally constructing a classifier capable of recognizing helmets. Shrestha et al. [3] used Haar-like features to detect faces, an edge detection algorithm to find helmet contour features, and an over fused skin color detection technique and support vector machine (SVM) to develop a hybrid approach for safety helmet recognition. However, there are limitations in this method, as the detection is not effective when the worker is not facing the camera or the contrast between helmet and background is low, and the rapid movement of the worker may also lead to detection failure. Park et al. [4] proposed the use of histogram for orientation gradient (HOG) features in the pedestrian detection phase and color histograms in the HD phase to determine whether a helmet is worn or not by using the spatial matching relationship between a human and a helmet. Feng et al. [5] proposed a method for foreground detection using a hybrid gaussian model followed by processing regarding the connected domain, which uses a model-based technique to recognize the human body and combines SIFT features and color statistical features to detect helmets. Rubaiyat et al. [6] utilized discrete cosine transform (DCT) with an HOG algorithm to extract frequency domain information from the image and used a support vector machine (SVM) classifier to distinguish between human and non-human targets. FE combines color and circular hough transform (CHT) to identify helmets with specific colors and circular contours. Although these traditional ML-HD methods have some advantages in terms of detection speed, these methods rely on manually designing features and training classifiers for a specific object. Due to a more limited feature set, these methods do not have enough generalization ability and often struggle to cope with complex built environments, and they thus may not be able to accurately identify a target in a changing environment.

At this stage, DL algorithms are widely used in helmet wearing detection and have become a mainstream technique. These algorithms are mainly categorized into two types, which are one-stage and two-stage methods. Two-stage methods (e.g., RCNN series) use the generation of region suggestions and FE to achieve accurate positioning, but the detection speed is slow; meanwhile, one-stage methods (e.g., SSD and YOLO) use an end-to-end architecture and directly output the position and category of the target frame, causing the speed of the detection to increase dramatically. Yu-Wen Huang et al. [7] improved the LeNet model based on a parallel network structure for human body detection and subsequently identified the presence of helmets by analyzing color features. Fang et al. [8] utilized Faster R-CNN to generate candidate regions through RPN, extract image features, and then classify and detect whether a worker is wearing a helmet or not; this is inadequate for real-time detection and has a large computational burden when dealing with large-scale video data, in spite of its high detection accuracy. Bo et al. [9] proposed an HD algorithm that combines OpenPose-v1 with Faster R-CNN, using OpenPose to accurately recognize the human head and neck regions in an image, but which may produce false detections when the worker is holding a helmet or the helmet is partially obscured. Deng et al. [10] employed a K-means clustering algorithm to optimize a priori frames in 2020 and performed dimensional clustering concerning target frames on a homemade dataset by adjusting the size of a priori frames so that the model could more accurately obtain edge information about the target object, especially in a helmet-wearing scenario, and optimizing the YOLOv4 model by using a multiscale training strategy; however, the detection accuracy of the improved model is still deficient in highly dense scenarios and is not applicable on resource-constrained devices. Zhou et al. [11] and Kisaezehra et al. [12] applied the YOLOv5 model to detecting SH, and the detection results had excellent performance in terms of speed and could reach a good balance with detection accuracy. However, due to

the more limited samples in the real scenario of a construction site, the model lacks sufficient generalization ability. Tan et al. [13] added an energetic detection scale to YOLOv5 to detect smaller targets. Jin et al. in 2021 [14] used K-means++ to improve the size matching for a priori anchor frames to make the model more accurate in selecting the bounding box (BB), thus accelerating the model convergence, and the use of a depthwise coordinate attention mechanism (DWCA) had strengthened the attention to the important channels, but it still could not deal with the fact that the worker part is occluded or that, when the target is far away and small, the system may miss detection or misjudge. Chen et al. [15] enhanced the model's ability to capture multi-scale feature details by applying a combined PP LCNet and collaborative attention mechanism (AM) module to the YOLOv4 model. The method significantly reduces the number of parameters and computations in the model, thus improving the detection speed and real-time performance for the model. Based on YOLOv5, Li et al. [16] introduced anchor-free SimOTA and SIoU to further improve the regression accuracy.

Liu et al. [17] chose YOLOv5 as the main body of the algorithm and integrated FasterNet, CBAM, and Wise-IoU to build the proposed network, which was faster in processing speed and lower in number of parameters and computations than the baseline method. Chen et al. [18] integrated the attention mechanism and concat Module in YOLOv5s and replaced the path aggregation network (PANet) with the weighted bidirectional feature pyramid network (BiFPN). In addition, the GIoU Loss function was replaced by the CIoU Loss function, which slightly improved the effect of the improved algorithm. An et al. [19] proposed an improved version of the YOLOv5s network. The network first integrates the Global Attention mechanism (GAM) and the Convolutional Block Attention Module (CBAM) and uses SIoU (SCYLLA-IoU LOSS) as a bounding box loss function. Assisted lightweight treatment using knowledge distillation technology was carried out. The network can recognize helmets in low light and at different distances. Li et al. [20] proposed a new and efficient detector model for the helmet detection of small targets and obscured targets. The YOLOv5 algorithm is improved by adding an attention mechanism, the CIoU (Complete Intersection Over Union) loss function, and the Mish activation function. Compared with the YOLOv5 algorithm, the result of the comparison experiment on the self-made helmet data set is 1.9% higher.

While the aforementioned approach has refined and enhanced the algorithm, it suffers from a large number of parameters and computations that hinders the use of terminal devices and hampers the detection of targets that are dense and small. Because of the shortcomings of the prior art, this study suggests a lightweight helmet-wearing detection model, YOLOv5s-FCW. In order to extract features, the YOLOv5 backbone network is substituted with FasterNet, and the original convolutional layer is replaced with a depth-separable convolution to drastically decrease network computation. Secondly, to decrease the number of parameters and computations while maintaining high detection accuracy, the enhanced CBAM was implanted into focus on the detection target. Lastly, the WIoU_Loss loss function was used as a substitute for the GIoU_loss loss function, improving the model's convergence. In contrast to the conventional detection technique, the improved algorithm reduces the model's parameter count, while sustaining high accuracy reduces hardware costs and satisfies the performance standards for construction site helmet-wearing detection.

2. YOLOv5s

The YOLO family of algorithms has been continuously developed and has produced several algorithms from, YOLOv1 to 10 [21], with continuously improved performance. Because of its various benefits and its excellent accuracy, speed, and understanding, YOLOv5 [22] is now the most popular among them. The YOLOv5 model is the smallest, uses the fewest resources, and trains and infers at the quickest rate. Combined with the comprehensive analysis of the current situation of helmet-wearing detection, YOLOv5 was used as the original model for improvement. YOLOv5 comprises four distinct network

architectures, which share the same underlying principles but differ in network depth, and width. Specifically, this paper focuses on YOLOv5s, which features the simplest network structure and divides it into four key components: input, backbone, neck, and output, as illustrated in Figure 1.

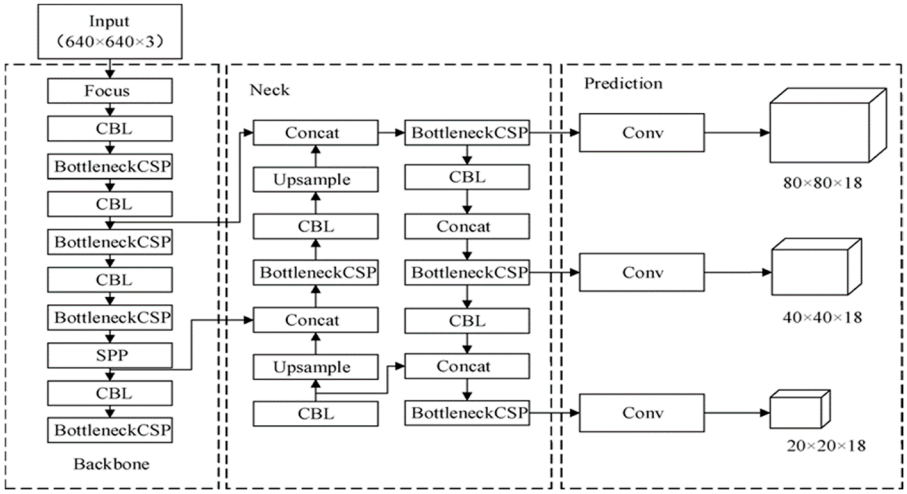

Figure 1. YOLOv5 network structure.

The input plays an essential function in pre-processing images. To boost the model's capability of identifying objects, YOLOv5 incorporates an adaptive anchor frame calculation and mosaic data enhancement mechanism into it.

The backbone network's primary function is to extract features from images, and it comprises three modules: focus, convolution, and spatial pyramid polling (SPP). The feature map is sliced by the slicing structure, which results in the conversion of a 640 × 640 × 3 picture into a 320 × 320 × 12 feature map. After the convolution operation, a twice-down sampled feature map is obtained. The slicing structure's primary objective is to raise the model's object detection (OD) speed and decrease the number of floating decimal operations per second. Spatial pyramid pooling can boost the perceptual field with minimal speed reduction and aid in feature layer and anchor frame alignment.

The neck network is mainly composed of a feature pyramid network (FPN) and path aggregation network (PAN), which can be used to target multiple scales. FPN is accountable for fusing different features from high and low layers, whereas PAN communicates positioning characteristics from the bottom of the tower upwards, so that the positioning information from the bottom level is transmitted to the top level. Employing both FPN and PAN structures can enhance detection presentation for dense targets.

The prediction side of YOLOv5, also referred to as the output side, consists of two components: the BB loss function and NMS. GIOU_Loss is adopted as the loss function, which addresses the issue of BB non-reconvergence and significantly enhances the velocity and precision of BB prediction. To improve the recognition of multiple targets and obscured targets, a weighted NMS approach is employed to identify the optimal OD location.

3. YOLOv5s-FCW

The YOLOv5 target detection algorithm is known for its high precision and fast processing speed in target identification. However, the current algorithm for detecting helmet wearing has a large number of parameters, a complex neural network structure, and requires high computation resources, making it unsuitable for deployment in embedded and other devices. Additionally, it struggles with detecting small-sized objects and has lower recognition accuracy. To overcome these challenges, improvements have been made to the YOLOv5 algorithm.

3.1. Improvement for Backbone Extraction Network

The backbone network concerning YOLOv5s is replaced with a backbone network regarding FasterNet for FE. FasterNet [23] is a new family of neural networks that runs faster on multiple processing platforms and outperforms networks such as MobileVit [24].

Group convolution (G Conv) and depthwise convolution (DW Conv) are used by networks like ShuffleNet, GhostNet, and MobileNet to distinguish between spatial features. Nevertheless, although DW Conv is effective in reducing FLOPs, it usually requires point-wise convolution (PW Conv) as a subsequent operation and cannot simply substitute the conventional Conv, and this can result in a considerable drop in accuracy. Therefore, in practical applications, channel quantities c about DW Conv is raised to c0 (c0 > c) to offset for the accuracy decline; for example, the width of DW Conv in the inverted residual block is expanded by six times. However, this leads to higher memory access, which may cause significant delays and reduce overall computational speed. Therefore, partial convolution (P Conv) is proposed, which only performs conventional Conv to a subset regarding input channels, leaving the other channels unaffected. P Conv can reduce memory access and computational complexity while maintaining accuracy, thus improving overall computational speed, especially for I/O-bound devices. Essentially, P Conv's FLOPs are lower than those of a regular Conv, and P Conv's FLOPs are only 1/16th about those concerning a regular Conv. In addition, P Conv's memory accesses are smaller. The P Conv structure diagram is shown in Figure 2.

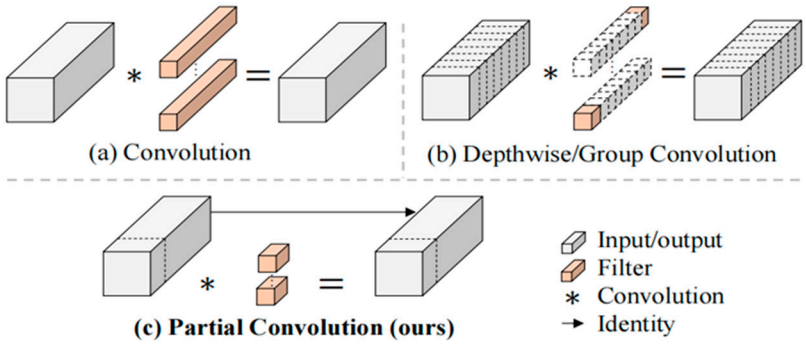

Figure 2. P Conv structure diagram.

The FasterNet neural network is built on the P Conv and is composed of four hierarchical levels. Before every level, there is an embedding layer that uses a 4 × 4 convolution with a step size of 4 or a merging layer that uses a 2 × 2 convolution with a step size of 2 for channel number expansion and spatial shrinking. Each stage has a bunch of FasterNet blocks. Every FasterNet block contains a P Conv layer, subsequently dual 1 × 1 Conv layers. The FasterNet structure diagram is shown in Figure 3.

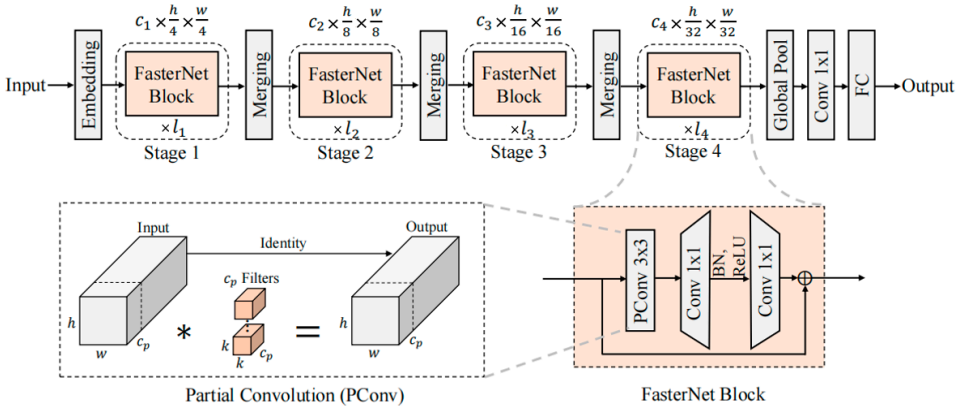

Figure 3. FasterNet structure diagram. ∗ denotes Convolution.

3.2. CBAM Module

Due to the fact that small targets such as SH occupy a relatively small percentage of the picture and are effortlessly impacted by components such as background, the initial YOLOv5 framework may fail to capture the important characteristics of these small objects during convolutional operation, generating unsatisfactory detection results for small targets in the image. To overcome this challenge, this study introduces an AM, which aims to inform the model which content and position to focus on, effectively extracting feature information from small and dense targets and further improving detection accuracy.

In 2018, Woo et al. suggested the CBAM [25] as an expansion of SeNet [26]. We have improved CBAM; the improved CBAM structure is shown in Figure 4. This module enhances the feature encoding ability of YOLOv5. The channel attention (CA) and spatial attention (SA) make up CBAM. These modules add weights to the channel and spatial layers, respectively, improving the representation of each. The first part of the CBAM attention module is the channel attention module. The second part of the CBAM attention module is the spatial attention module. In YOLOv5, CBAM may be included into the residual network and enhanced when used in conjunction with the C3 or Conv modules.

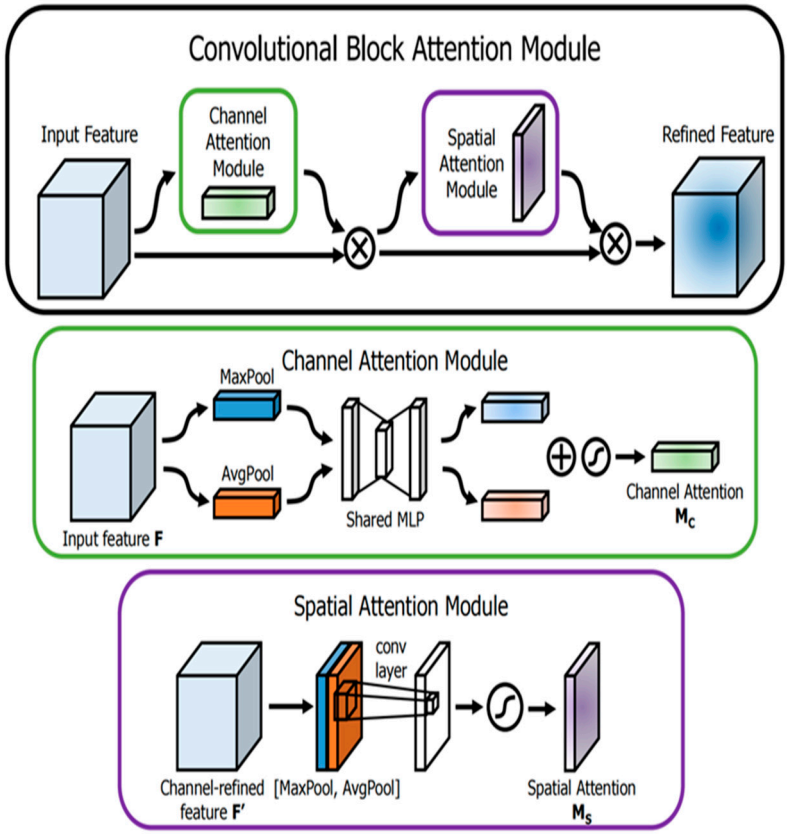

Figure 4. CBAM module structure diagram.

3.3. Loss Function Improvement

The design of the loss function determines detection performance concerning target detection, a fundamental issue in computer vision. A well-structured BB loss function is a crucial component of the OD loss function, and it can considerably enhance the capability of the OD model. In recent years, most studies have presumed that the training data contain high-quality examples and have focused on bolstering the BB loss's fitting capacity. However, we observe that the OD training set also includes low-quality examples, which can adversely affect the model's detection performance if the BB regression for these examples is continuously emphasized. Focal-EIoUv1 attempted to resolve this issue, but the non-monotonic focusing mechanism's potential was not completely used by its static

focusing mechanism. With this in mind, we introduce Wise-IoU (WIoU) [27]. "Outliers" are used by the dynamic non-monotonic focus mechanism in place of IoU to assess the anchor frame quality and offer an intelligent gradient gain allocation strategy. The function that calculates loss for BB in the YOLOv5 model is GIoU_Loss [28], and the GIoU_Loss equation is:

$$L_{(GIOU)} = 1 - U_{i,o} + \left(\frac{|C - B \cup B^{gt}|}{|C|} \right) \quad (1)$$

where $B = \{x, y, w, h\}$ denotes the prediction box dimensions, $B^{gt} = \{x^{gt}, y^{gt}, w^{gt}, h^{gt}\}$ demonstrate the dimension of the real box, and the minimum area of B and B^{gt} is C.

As the anticipated BB's distance from the ground truth BB increases, the value of C also increases. At the same time, the area of the projected BB and the ground truth BB differs from C, and this discrepancy also grows, gradually approaching 1. In this study, WIoU_Loss is selected as the function to quantify BB loss in the improved YOLOv5 model due to the slow and unstable convergence regarding the GIoU loss function; the WIoU_Loss formula is:

$$L_{WIOU} = R_{WIOU} L_{WIOU} \quad (2)$$

$$R_{WIOU} = \exp \left(\frac{(x - x_{gt})^2 - (y - y_{gt})^2}{\left(W_g^2 + H_g^2\right)^*} \right) \quad (3)$$

where W_g and H_g denote the width and height of the minimum enclosing frame, separately. To prevent convergence issues caused by gradients generated by R_{WIOU}, W_g and H_g are removed from the computational graph using superscript * to indicate this operation.

This method effectively removes convergent elements without the need to introduce additional metrics, such as element ratios. To prevent R_{WIOU} from generating gradients that prevent convergence, the computational graph is divided into W_g and H_g (this operation is indicated by the superscript *). No additional measure, like element ratio, is added since it essentially removes obstacles to junction.

4. Experimental Results and Analysis

4.1. Dataset

In this study, various methods such as video screenshots, photography, and online search were used in the experiment to collect pictures of helmet wearing, covering various scenes such as construction sites and daily life. At the same time, this study also uses the public safety helmet dataset SHDW to increase the sample richness. The helmet dataset created in this study contains a variety of helmet targets at different scales, including large, medium and small-scale targets, as well as multiple scale targets in dense scenarios. In addition, the dataset includes targets in low-light and foggy conditions, as well as targets with occlusion and background blur. For consistency, all input images are adjusted to a size of 640 × 640 pixels. In total, we collected about 3241 images related to safety helmets.

Then, we randomly process the data set according to an 8:1:1 partitioning criterion and subdivide the dataset into a training set, verification set, and test set. To visualize the diversity of the dataset, some sample images are presented in Figure 5.

In order for the model to learn the mapping between the input data and the output labels, the data need to be labeled. This process not only requires a sufficient amount of picture data but also requires labeling. In order to complete the experiment in this paper, a powerful annotation tool named LabelImg was adopted to manually annotate images. In this experiment, we annotate two categories, namely "person" and "hat". These targets in the dataset were labeled with "hat" to indicate the correct usage of helmets and "person" to indicate the non-usage of helmets.

Figure 5. Some sample images of the dataset. (**a**) Obstructed positive sample targets, (**b**) Small targets, (**c**) Fuzzy positive sample targets, (**d**) Interference targets, (**e**) Intensive targets, (**f**) Complex targets.

In order to improve the performance of the model and increase the diversity of the dataset, we enhanced the marked data set by randomly rotating, randomly flipping, changing brightness, adjusting contrast, adding Gaussian noise, and other methods to expand the data set. Overfitting is effectively reduced during training, and network generalization and robustness are improved. The overall effect is shown in Figures 6–8.

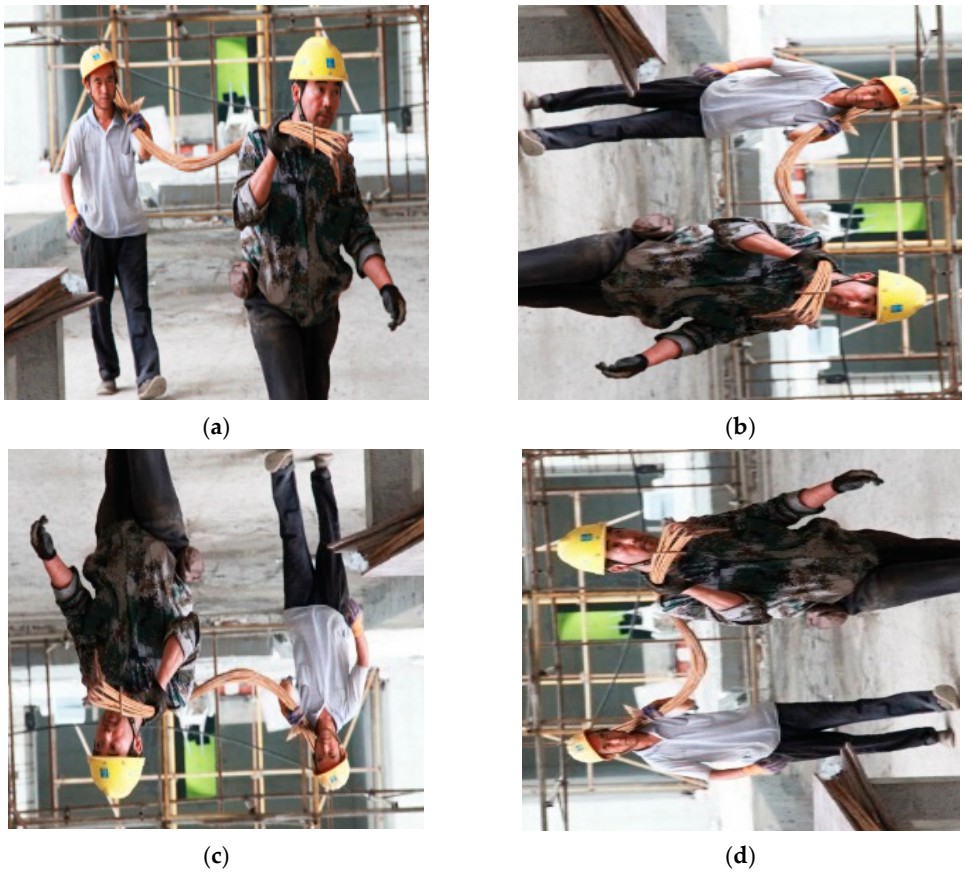

Figure 6. Random rotation angle. (**a**) Original image, (**b**) Rotation 90°, (**c**) Rotation 180°, (**d**) Rotation 270°.

(**a**) (**b**)

Figure 7. Fogging of images. (**a**) Original image, (**b**) After fogging.

(**a**) (**b**)

Figure 8. Changing the brightness. (**a**) Original image, (**b**) After fogging.

4.2. Setup

For this research, we utilized Windows 10 and hardware with an i5-13400 processor (Intel Corporation, Santa Clara, CA, USA) operating at a refresh rate of about 2.50 GHz, as well as a single 3060Ti GPU (NVIDIA, Santa Clara, CA, USA) with 8 GB of video memory. The improved method is based on YOLOv5s, the smallest network among the four versions of the YOLOv5 model. The Pytorch (Version 1.12.1) framework and the Python programming language were employed to develop the proposed algorithm. GPU acceleration was achieved using CUDA 11.6. In order to ensure the fairness of the experimental comparison, the other algorithms and the proposed method use the same experimental setup as in this paper.

4.3. Model Indicators

In the research field of helmet detection, the main evaluation indicators of algorithm performance include Precision, Recall, AP (Average Precision), and mAP (mean Average Precision). In addition, the evaluation of the algorithm's execution speed is quantified by Frames Per Second (FPS). This paper also follows the standard evaluation system and makes a comprehensive and detailed evaluation of the helmet target detection algorithm. The predicted results are evaluated by calculating proportion for their intersection with true values and are considered true if the resulting value exceeds a predefined threshold. This method yields four possible classification outcomes for prediction results: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), where accuracy

is likelihood regarding correct detection amid all detected targets; thus, its equation is expressed as (4):

$$Precison = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4}$$

The concept of recall refers to the probability of accurately detecting all target instances, and its mathematical representation is given by (5):

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

AP is a measure of model precision at different recall rates, computed as the area enclosed by the Precision–Recall (PR) curve and coordinate axis. For a continuous PR curve, AP can be calculated using integration. The formula is expressed as shown in Equation (6):

$$\mathrm{AP} = \int_0^1 \mathrm{PR} dr \tag{6}$$

The formula for mAP, expressed as Equation (7), represents the average of mean precision values for various target categories. In this paper, n refers to two categories, namely hat and person.

$$\mathrm{mAP} = \frac{\sum_{i=1}^{n} AP_i}{n} \tag{7}$$

In the task of helmet wearing detection, in order to meet the requirements of real-time detection, the calculation speed of the model is also a relatively important index, so this paper also introduced FPS as an evaluation standard in the comparison experiment. FPS is the number of helmet image frames that a model can process in one second. The calculation speed of the model can be judged by comparing the number of helmet images identified per unit time and the time required to process these helmet images.

4.4. Attentional Selection Experiments

To tackle the problem of detecting small targets in the YOLOv5s network, this paper explores three different AM (SE, CBAM, and CA) that are added to its backbone for comparison. Table 1 displays the findings of the conducted test.

Table 1. Experimental results for three AM.

Methods	mAP (IoU = 0.5)	mAP (IoU = 0.5:0.95)	AP	R	FPS
YOLOv5s	0.812	0.571	0.853	0.752	57.8
YOLOv5s + CA	0.812	0.558	0.884	0.74	57.6
YOLOv5s + CBAM	0.824	0.591	0.885	0.758	58.1
YOLOv5s + SE	0.821	0.572	0.838	0.759	58.8

While the CA results in a slight enhancement in AP, it also causes a decrease in mAP (IoU = 0.5:0.95); the SE causes a decrease in AP but results in a small increase in the total mAP of the structure; the CBAM achieves a 1.2% increase in mAP (IoU = 0.5) and a 3.2% increase in AP with a detection speed that is almost constant, so it was finally decided to choose CBAM. Figure 9 shows the comparison of detection results for YOLOv5s and YOLOv5s + CBAM.

(**a**) (**b**)

Figure 9. Comparison of detection results for YOLOv5s and YOLOv5s + CBAM. (**a**) YOLOv5s, (**b**) YOLOv5s + CBAM.

4.5. Comparative Experiments

Figures 10 and 11 present the average accuracy rate regarding the improved model and initial YOLOv5s structure, which were trained in an equivalent arrangement for 100 epochs.

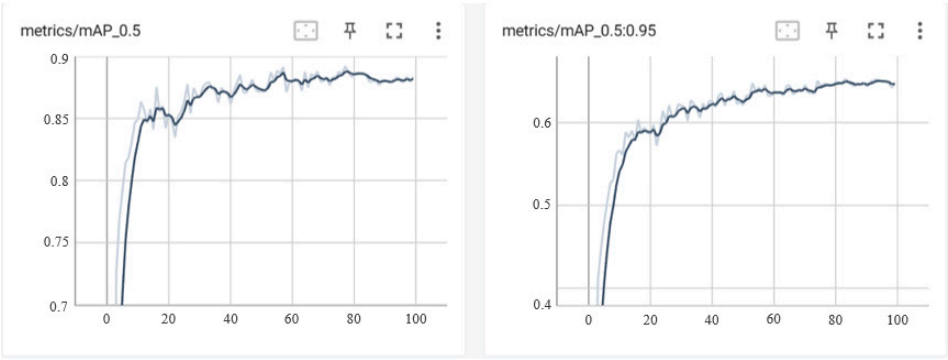

Figure 10. YOLOv5s mAP curve. The dark curve is the result on the verification set, and the bright curve is the result on the training set.

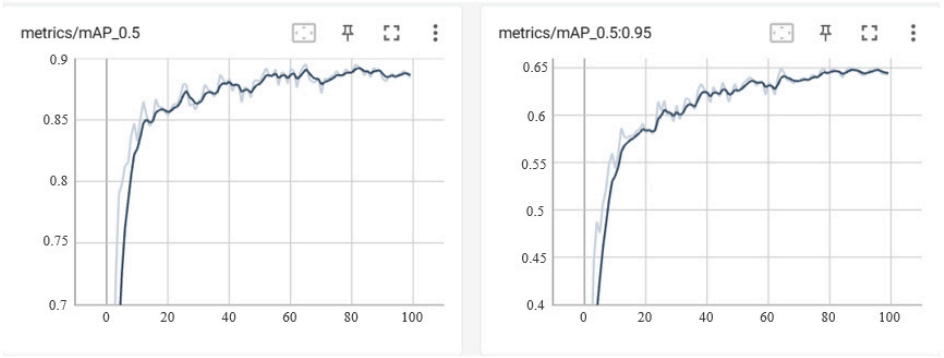

Figure 11. YOLOv5s-FCW curve. The dark curve is the result on the verification set, and the bright curve is the result on the training set.

4.6. Improvement for Rest About Network and Its Ablation Experiments

This research then goes on to compare and examine the detection outcomes after choosing the attention method, and for the problems of many parameters, complex network, and large number of computations of existing algorithms for helmet-wearing detection. FasterNet replaces the YOLOv5s backbone network with FE, lowering the number of parameters and the computing burden on the network. Additionally, to enhance accuracy without sacrificing the detection speed of the model, GIoU Loss was replaced with WIoU

Loss and incorporated into the YOLOv5s network. We added different modules to the baseline model and performed a series of experiments, and the results are shown in Table 2.

Table 2. Experimental results for YOLOv5s-based improvement and its ablation.

Methods	mAP (IoU = 0.5)	AP	R	Parameters (10^6)	GFLOPs
YOLOv5s	0.812	0.853	0.752	7.07	16.5
YOLOv5s + CBAM	0.824	0.885	0.758	7.09	16.5
YOLOv5s + FasterNet	0.81	0.837	0.574	5.62	13.1
YOLOv5s + WIOU	0.819	0.862	0.77	7.07	16.5
YOLOv5s + CBAM + WioU	0.815	0.878	0.732	7.09	16.5
YOLOv5s + CBAM+ FasterNet	0.772	0.854	0.713	5.62	13.1
YOLOv5s + FasterNet + WIoU	0.802	0.888	0.721	5.62	13.1
YOLOv5s + FasterNet + WIoU + CBAM(YOLOv5-FCW)	0.837	0.889	0.813	5.62	13.1

We add each module to the base network-YOLOv5s for ablation experiments to confirm the effect of different modules on experimental results. As presented in Table 2, the foregoing divided eight groups of data were trained separately, and the obtained metrics were averaged as the final measure, with the original YOLOv5s without any improvement as the benchmark and + denoting the module mixed improvement. The accuracy, recall, and average accuracy of the original YOLOv5s are 81.2%, 85.3%, and 75.2%, respectively, as observed. Using this as a benchmark, every improvement in each metric will have some improvement, and only the combination of FasterNet and CBAM and FasterNet and WIOU, respectively, will slightly reduce the accuracy rate, but the accuracy rate reaches 85.4% and 88.8%, respectively, a respective increase of 0.1% and 3.5% compared to the benchmark. In a comprehensive view, the best result is achieved by improving all four together. The mAP (IoU = 0.5), P, and R of the proposed method YOLOv5-FCW reached 83.7%, 88.9%, and 81.3%, respectively, which increased by 2.5%, 4.6%, and 6.1% compared with the benchmark. The number of parameters of YOLOv5-FCW is reduced by 20% compared with the benchmark, which further verifies the feasibility of the improved scheme.

The R of Yolov5s-FCW is 0.813, while the R of YOLOv5s + FasterNet + WIoU and YOLOv5s + CBAM + FasterNet are 0.721 and 0.713, respectively. The R of Yolov5S-FCW is increased by 12.76% and 14.03% over YOLOv5s + FasterNet + WIoU and YOLOv5s + CBAM + FasterNet, respectively. This indicates that YOLOv5s-FCW has a greater capability and higher probability of accurately detecting all helmet target instances.

In terms of Parameters (10^6) and GFLOPS, YOLOV5S-FCW achieved almost the same results as YOLOv5s + FasterNet + WIoU and YOLOv5s + CBAM + FasterNet. The network parameters and computing burden of YOLOV5S-FCW are similar to those of YOLOv5s + FasterNet + WIoU and YOLOv5s + CBAM + FasterNet. However, the performance of indicator R is improved by 12.76% and 14.03% compared with YOLOv5s + FasterNet + WIoU and YOLOv5s + CBAM + FasterNet. This proves that compared with YOLOv5s + FasterNet + WIoU and YOLOv5s + CBAM + FasterNet, YOLOV5S-FCW can achieve higher detection accuracy at the same computational cost.

4.7. Analysis of Test Results

The partial visualization of the detection results is presented in Figure 12. The left-hand side of the figure displays the original YOLOv5s detection graph, while the right-hand side displays the improved YOLOv5s detection effect. The images in the figure include (a) dense helmet targets, (b) small helmet targets, and (c) obscured helmet targets.

(**a**)

(**b**)

(**c**)

Figure 12. Comparison of model detection results in different scenarios. (**a**) Comparison of dense helmet detection, (**b**) Comparison of long-range small helmet detection, (**c**) Comparison of obscured helmet detection.

From the visualization of the results in Figure 12, it is evident that the initial YOLOv5s model failed to detect the target without a helmet in the center of Figure 12a, as well as the small helmet target on the left side of Figure 12b and the obscured helmet target in Figure 12c. However, the improved model accurately detected all of these targets. At long ranges and for tiny targets, the original model is even more serious in missing detection, even if the enhanced model can still recognize extremely tiny objects at a great distance with accuracy, and the confidence score has been improved. It is shown that the improved model has better generalization ability in crowded targets, small targets, and occluded target scenarios.

4.8. Comparative Experiments with Other SOTA Methods

In order to further verify the performance of the algorithm, a comparison experiment with YOLOv7, 8, 9, and transformers SOTA methods was set up. The experimental results of the comparison between the proposed method and SOTA methods such as YOLOv7, 8, 9, and transformers are shown in Table 3.

Table 3. Comparison with other SOTA methods.

Methods	mAP (IoU = 0.5)	AP	R	Parameters (10^6)	GFLOPs
YOLOv7	0.769	0.826	0.787	9.3	23.2
YOLOv8	0.798	0.843	0.796	8.6	20.4
YOLOv9	0.828	0.865	0.805	10.4	26.0
Transformers	0.836	0.891	0.798	36.7	84.97
Ours	0.837	0.889	0.813	5.62	13.1

As can be seen from Table 3, our method outperforms YOLOv7, YOLOv8, and YOLOv9 on all indexes. In the AP index, the transformers method is the best. At the same time, transformers' Parameters and GFLOPs are the largest, which means that it will have to pay a lot of computing costs. However, the Parameters and GFLOPs of our method are the least, which shows that the proposed method can achieve relatively better performance with less computational cost.

5. Conclusions

To handle the issues regarding existing safety HD methods, such as high parameter counts, complex neural networks, intensive computation requirements, limited suitability for deployment in embedded and other devices, and insufficient accuracy in detecting small targets, we propose an optimized lightweight algorithm called YOLOv5s-FCW for HD. Firstly, to enhance the efficiency of the YOLOv5s algorithm, we replaced backbone network with FasterNet for FE, resulting in a reduction in both the parameters and the computational complexity of the network; secondly, we integrated a CBAM module into the YOLOv5 model to enhance the model's capacity for detecting small targets, such as SH, by increasing its focus on them; finally, to enhance the convergence of the model, we replaced the GIoU_loss loss function with the WIoU_Loss loss function. Experimental results demonstrate that our proposed YOLOv5s-FCW method enhances mAP by 2.5% and accuracy by 4.6% compared to the baseline algorithm. This optimized algorithm also improves the recognition of small and obscured targets and reduces the computational requirements of the YOLOv5s model by 20% while maintaining a higher average accuracy of detection. As a result, our algorithm reduces hardware costs and fulfills the performance standards for the detection of SH in construction zones.

Although the proposed method improves the detection accuracy and detection speed of the safety helmet target to a certain extent, there is still room for further improvement in the parameters and lightweight of the network. Whether it can be applied to videos of safety helmet scenes still needs to be studied in the future work.

Author Contributions: Conceptualization, J.L. and H.Z.; Data curation, H.Z. and P.L.; Formal analysis, J.L.; Funding acquisition, D.X.; Investigation, D.X.; Methodology, G.L. and P.L.; Resources, G.L. and P.L.; Software, P.L.; Validation, H.Z., S.H. and D.X.; Writing—original draft, G.L. and S.H.; Writing—review & editing, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 52074064.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

References

1. Kurien, M.; Kim, M.K.; Kopsida, M.; Brilakis, I. Real-time simulation of construction workers using combined human body and hand tracking for robotic construction worker system. *Autom. Constr.* **2018**, *86*, 125–137. [CrossRef]
2. Liu, X.-H.; Ye, X.-N. Skin color detection and Hu moments in helmet recognition research. *J. East China Univ. Sci. Technol. (Nat. Sci. Ed.)* **2014**, *3*, 365–370.
3. Shrestha, K.; Shrestha, P.P.; Bajracharya, D.; Yfantis, E.A. Hard-hat detection for construction safety visualization. *J. Constr. Eng.* **2015**, *2015*, 721380. [CrossRef]
4. Park, M.W.; Elsafty, N.; Zhu, Z. Hardhat-wearing detection for enhancing on-site safety of construction workers. *J. Constr. Eng. Manag.* **2015**, *141*, 04015024. [CrossRef]
5. Feng, G.; Chen, Y.; Chen, N.; Li, X.; Song, C. Research on automatic identification technology of the safety helmet based on machine vision. *Mach. Des. Manuf. Eng.* **2015**, *44*, 39–42.
6. Rubaiyat, A.H.M.; Toma, T.T.; Kalantari-Khandani, M.; Rahman, S.A.; Chen, L.; Ye, Y.; Pan, C.S. Automatic detection of helmet uses for construction safety. In Proceedings of the 2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW), Omaha, NE, USA, 13–16 October 2016; pp. 135–142.
7. Huang, Y.; Pan, D. Helmet recognition based on parallel double convolutional neural networks. *Technol. Dev. Enterp.* **2018**, *37*, 24–27.
8. Fang, Q.; Li, H.; Luo, X.; Ding, L.; Luo, H.; Rose, T.M.; An, W. Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. *Autom. Constr.* **2018**, *85*, 1–9. [CrossRef]
9. Bo, Z.; Song, Y.; Xiong, R.; Zhang, S. Helmet-wearing detection considering human joint. *China Saf. Sci. J.* **2020**, *30*, 177–182.
10. Deng, B.; Lei, X.; Ye, M. Safety helmet detection method based on YOLO v4. In Proceedings of the 2020 16th International Conference on Computational Intelligence and Security (CIS), Nanning, China, 27–30 November 2020; pp. 155–158.
11. 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.
12. Kisaezehra; Farooq, M.U.; Bhutto, M.A.; Kazi, A.K. Real-Time Safety Helmet Detection Using Yolov5 at Construction Sites. *Intell. Autom. Soft Comput.* **2023**, *36*, 911–927. [CrossRef]
13. Tan, S.; Lu, G.; Jiang, Z.; Huang, L. Improved YOLOv5 network model and application in safety helmet detection. In Proceedings of the 2021 IEEE International Conference on Intelligence and Safety for Robotics (ISR), Nagoya, Japan, 4–6 March 2021; pp. 330–333.
14. Jin, Z.; Qu, P.; Sun, C.; Luo, M.; Gui, Y.; Zhang, J.; Liu, H. DWCA-YOLOv5: An improved single shot detector for safety helmet detection. *J. Sens.* **2021**, *2021*, 4746516. [CrossRef]
15. Chen, J.; Deng, S.; Wang, P.; Huang, X.; Liu, Y. Lightweight helmet detection algorithm using an improved YOLOv4. *Sensors* **2023**, *23*, 1256. [CrossRef] [PubMed]
16. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
17. Liu, Y.; Jiang, B.; He, H.; Chen, Z.; Xu, Z. Helmet wearing detection algorithm based on improved YOLOv5. *Sci. Rep.* **2024**, *14*, 8768. [CrossRef] [PubMed]
18. Chen, H.; Qi, J.; Wang, M.; Wu, C. Helmet-Wearing Detection Algorithm Based on Improved YOLOv5s. In Proceedings of the 2023 42nd Chinese Control Conference (CCC), Tianjin, China, 24–26 July 2023; pp. 8564–8569.
19. An, Q.; Xu, Y.; Yu, J.; Tang, M.; Liu, T.; Xu, F. Research on Safety Helmet Detection Algorithm Based on Improved YOLOv5s. *Sensors* **2023**, *23*, 5824. [CrossRef] [PubMed]
20. Li, Y.; Zhang, J.; Hu, Y.; Zhao, Y.; Cao, Y. Real-time safety helmet-wearing detection based on improved yolov5. *Comput. Syst. Sci. Eng.* **2022**, *43*, 1219–1230. [CrossRef]
21. 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 (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
22. Han, J.; Liu, Y.; Li, Z.; Liu, Y.; Zhan, B. Safety helmet detection based on YOLOv5 driven by super-resolution reconstruction. *Sensors* **2023**, *23*, 1822. [CrossRef] [PubMed]
23. Chen, J.; Kao, S.; He, H.; Zhuo, W.; Wen, S.; Lee, C.-H. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 12021–12031.
24. Mehta, S.; Rastegari, M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv* **2021**, arXiv:2110.02178.
25. 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.

26. Jie, H.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023.
27. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. *arXiv* **2023**, arXiv:2301.10051.
28. Rezatofighi, H.; Tsoi, N.; Gwak, J.Y.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.