sensors

MDPI

Article

Research on Safety Helmet Detection Algorithm Based on Improved YOLOv5s

Qing An [1], Yingjian Xu [2,*], Jun Yu [3], Miao Tang [1], Tingting Liu [1] and Feihong Xu [1]

[1] School of Artificial Intelligence, Wuchang University of Technology, Wuhan 430223, China; 120160450@wut.edu.cn (Q.A.); 120160311@wut.edu.cn (M.T.); 120160804@wut.edu.cn (T.L.); 120160287@wut.edu.cn (F.X.)
[2] School of Safety Science and Emergency Management, Wuhan University of Technology, Wuhan 430079, China
[3] USTC iFLYTEK Co., Ltd., Hefei 230088, China; junyu@iflytek.com
* Correspondence: xyj971230@163.com

Abstract: Safety helmets are essential in various indoor and outdoor workplaces, such as metallurgical high-temperature operations and high-rise building construction, to avoid injuries and ensure safety in production. However, manual supervision is costly and prone to lack of enforcement and interference from other human factors. Moreover, small target object detection frequently lacks precision. Improving safety helmets based on the helmet detection algorithm can address these issues and is a promising approach. In this study, we proposed a modified version of the YOLOv5s network, a lightweight deep learning-based object identification network model. The proposed model extends the YOLOv5s network model and enhances its performance by recalculating the prediction frames, utilizing the IoU metric for clustering, and modifying the anchor frames with the K-means++ method. The global attention mechanism (GAM) and the convolutional block attention module (CBAM) were added to the YOLOv5s network to improve its backbone and neck networks. By minimizing information feature loss and enhancing the representation of global interactions, these attention processes enhance deep learning neural networks' capacity for feature extraction. Furthermore, the CBAM is integrated into the CSP module to improve target feature extraction while minimizing computation for model operation. In order to significantly increase the efficiency and precision of the prediction box regression, the proposed model additionally makes use of the most recent SIoU (SCYLLA-IoU LOSS) as the bounding box loss function. Based on the improved YOLOv5s model, knowledge distillation technology is leveraged to realize the light weight of the network model, thereby reducing the computational workload of the model and improving the detection speed to meet the needs of real-time monitoring. The experimental results demonstrate that the proposed model outperforms the original YOLOv5s network model in terms of accuracy (Precision), recall rate (Recall), and mean average precision (mAP). The proposed model may more effectively identify helmet use in low-light situations and at a variety of distances.

Keywords: detection; YOLOv5; SIoU; combinatorial attention mechanisms; K-means++; knowledge distillation

Citation: 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. https://doi.org/10.3390/s23135824

Academic Editor: Paweł Pławiak

Received: 11 May 2023
Revised: 19 June 2023
Accepted: 20 June 2023
Published: 22 June 2023

1. Introduction

Construction workers frequently incur injuries due to their failure to wear safety helmets [1]. Based on accident data from 2015 to 2018, out of 78 construction-related accidents, 53 incidents, or 67.95 percent of all accidents, were brought on by failure to wear safety helmets [2]. Consequently, it is crucial to verify that safety helmets are being. Currently, detecting whether employees are wearing safety helmets depends on human monitoring, which has the disadvantages of being expensive and failing to provide real-time detection [3]. Numerous image-based detection methods have been proposed to

address this issue, both conventional image detection techniques and methods for detecting objects using deep learning technology [4,5].

The Viola–Jones algorithm, histograms of oriented gradients (HOG) + support vector machine (SVM), and others are examples of traditional image detection techniques for safety helmet identification and the development of deformable parts models (DPM). Viola–Jones (VJ) can process targets in real time with high accuracy [6]. This method comprises three structures, namely, the feature type and evolution, learning algorithm, and cascade structure. Moreover, it requires a high detection rate for a single classifier, such that each classifier must obtain a detection rate of 99.7% in order to attain an overall detection rate of 90% [7]. Some researchers conducted a detailed analysis of feature-oriented classification storage and a feature matching query utilizing HOG feature extraction and SVM to improve target recognition and classification efficiency while considering the different granularity displayed by the test images [8]. DPM is an extension of HOG that trains a gradient model of the object using SVM, matches the model with the target, and achieves target classification. In the early stages of image recognition, conventional target detection methods required the extraction of a large number of target features, which had two main disadvantages: a large number of logical operations were needed to generate sufficient candidate regions, and the complexity of the characteristics prevented the identification speed and accuracy from meeting the objectives [9].

Girshick et al. introduced a target detection method that utilized a deep convolutional neural network, which effectively overcame the limitations of existing target detection technologies. Deep learning target identification approaches now fall into two main categories: one-stage methods based on regression and two-stage methods based on candidate region selection [10–12]. Girshick's region-convolutional neural network (R-CNN) is one of the two-stage techniques used to extract picture information [13]. However, R-CNN faces difficulties when generating candidate frames in complex backgrounds, and scaling and cropping during feature extraction may result in the loss of image information. In [14], Ross et al. proposed the famous fast region with CNN (Fast R-CNN) network, which replaces the spatial pooling layer of SPP-Net, simplifying the network model and saving computing resources. However, region pruning relies on a selective search method to generate regions of interest and cannot be accelerated by GPU. The same year saw the introduction of faster regions with CNN (Faster R-CNN) by Renetal et al., which leverages a region prediction network (RPN) to replace the conventional region prediction algorithm and uses a fully connected layer to enhance image robustness [15]. However, Faster R-CNN cannot share the parameters of multiple related regions in the second stage, which increases computational overhead. In addition, using the fully connected layer may lead to information loss [16,17].

The single shot detector (SSD) and you only look once (YOLO) algorithms are part of the one-stage target identification algorithm, which is a quicker technique. Redmon et al. introduced the YOLO [18] model, which transforms the challenging two-step detection procedure into an abstract regression issue. A multi-scale-based detection technique called SSD, which can effectively find several small objects, was proposed by Liu et al. [19]. However, the preprocessing of minor things could be optimized after the SSD algorithm performs depth convolution. YOLOv2, developed by Redmon et al., employs DarkNet-19 as a novel fundamental model and permits end-to-end training [20]. They also presented the YOLOv3 [21] network, which considerably improves the model's ability to recognize objects of varying sizes by fusing three feature layers of different sizes using feature pyramid network (FPN) technology. Park et al. proposed two-step real-time night-time fire detection in urban environments using Static ELASTIC-YOLOv3 [22]. YOLOv4 was introduced by Bochkovski et al. and leverages the Cross-Stage Partial (CSP) Darknet-53 as the backbone network and replaces the FPN algorithm in the YOLOv3 network [23]. As a result, the model's detection precision was significantly increased. Lin et al. proposed utilizing the improved YOLOv4 to perform defect detection on stitched images of rotating tools, and it has demonstrated good performance [24]. YOLOv5 is based on YOLOv4,

which was proposed by Glenn [25]. YOLOv4 introduced a focus module to boost detection speed and accuracy, and it is now a model with excellent target recognition accuracy. Wang et al. proposed an improved YOLOv5 cotton foreign fiber detection and classification based on polarization imaging, which improved the accuracy and speed of detection [26]. Based on the YOLOv5 model, this research suggests an enhanced YOLOv5s hard hat detection approach. The prediction box was first adjusted by strengthening the YOLOv5 adaptive anchor. In order to improve the feature extraction of small targets and, hence, increase the detection accuracy of the targets, we incorporated a combined attention method. Finally, we improved the regression box's accuracy by using SIoU Loss as the bounding box loss function. The experiments the authors conducted with their hard hat dataset show that the enhanced YOLOv5s model can effectively extract characteristics from small targets. The approach put forth in this research can be used in the field of helmet detection to reduce worker fatalities.

2. Proposed Method

The technical roadmap of this paper is shown in Figure 1.

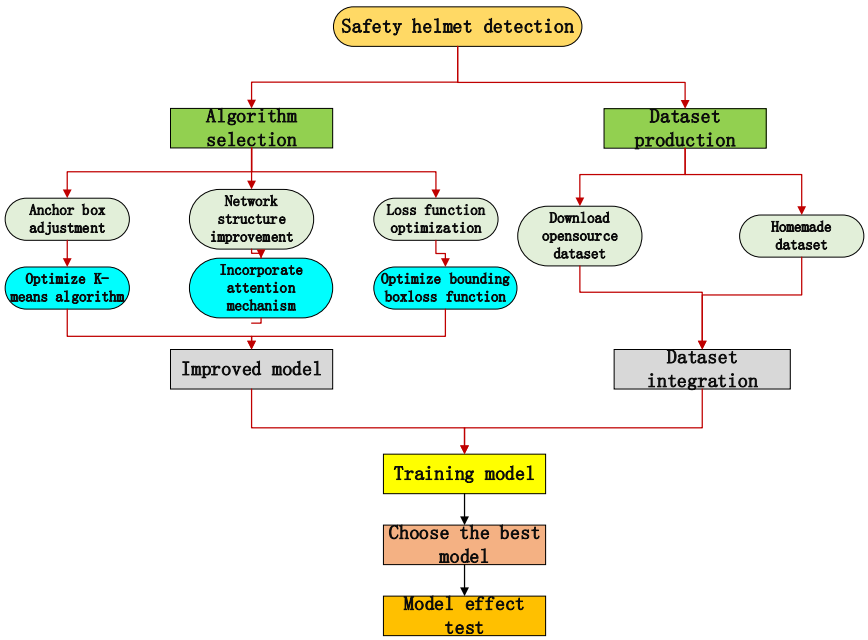

Figure 1. Overall framework diagram of the proposed method.

In order to further select the most accurate algorithm, the actual detection effect of the current mainstream target detection algorithm model was utilized in real images of small targets with complex backgrounds, as shown in Figure 2. In Figure 2a, we show the actual detection effect of the SSD model, where the confidence of person is 0.69, and the confidence of hat is 0.70. Figure 2b shows the actual detection effect of the Fast R-CNN model Figure 1, where the confidence of person is 0.65, and the confidence of hat is 0.68. Figure 2c is the actual detection effect diagram of the Faster R-CNN model, where the confidence of person is 0.71, and the confidence of hat equals 0.72. Figure 2d demonstrates the actual detection effect diagram of the YOLOv5s model, where the confidence of person is 0.73, and the confidence of hat is 0.73. Compared with other mainstream target detection algorithms, the experimental results illustrate that the YOLOv5s target detection algorithm has a higher detection accuracy than the SSD model, Fast R-CNN model, and Faster R-CNN model, while maintaining a light weight.

Figure 2. Different target detection algorithms are used for small targets with complex backgrounds in real images and actual detection effect pictures. (**a**) The detection effect diagram of the SSD model. (**b**) The detection effect diagram of the Fast R-CNN model. (**c**) The detection effect diagram of the Faster R-CNN model. (**d**) The detection effect diagram of the YOLOv5s model.

In order to create an object detection algorithm that can effectively detect small objects, especially helmets, this paper proposes an improved YOLOv5s network model based on the YOLOv5s model. This method can effectively enhance the ability to extract helmet target features.

A real-time target identification system called YOLOv5 offers four network models with varying degrees of depth: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The lightweight YOLOv5s network structure, shown in Figure 3, is made up of four parts: input (Input), backbone network (Backbone), neck network (Neck), and detecting head (Prediction). This research focuses on enhancing this structure. The focus module, the CBL convolutional layer, and the CSP1_X module are the components of the YOLOv5s backbone network. A $640 \times 640 \times 3$ picture is fed into the focus structure, followed by slice processing, and a convolution operation yields a $320 \times 320 \times 64$ feature map. The CBL convolutional layer and the CSP1_X module are then used to create a rich feature map with semantic information. The neck network implements two upsampling operations using CSP2_X and FPN+PAN models to combine shallow and high-level semantic features, realizing the fusion of multi-scale receptive fields and enhancing the feature fusion ability. For the prediction, we used the regression + classification method, dividing the input image into three different sized grids: 80×80, 40×40, and 20×20, thereby identifying large, medium, and small targets. Furthermore, YOLOv5 applies adaptive picture scaling, adaptive anchor frame computation, and mosaic data improvement to the input. The backbone network receives the focus and CSP structures, whereas the neck network receives the FPN+PAN structure [27]. The target detection frame in the output terminal employs GIoU_Loss as its loss function. We also suggest the NMS non-maximum suppression approach. The YOLOv5s algorithm not only increases detection accuracy when compared to the conventional two-stage detection approach, but also significantly reduces training time.

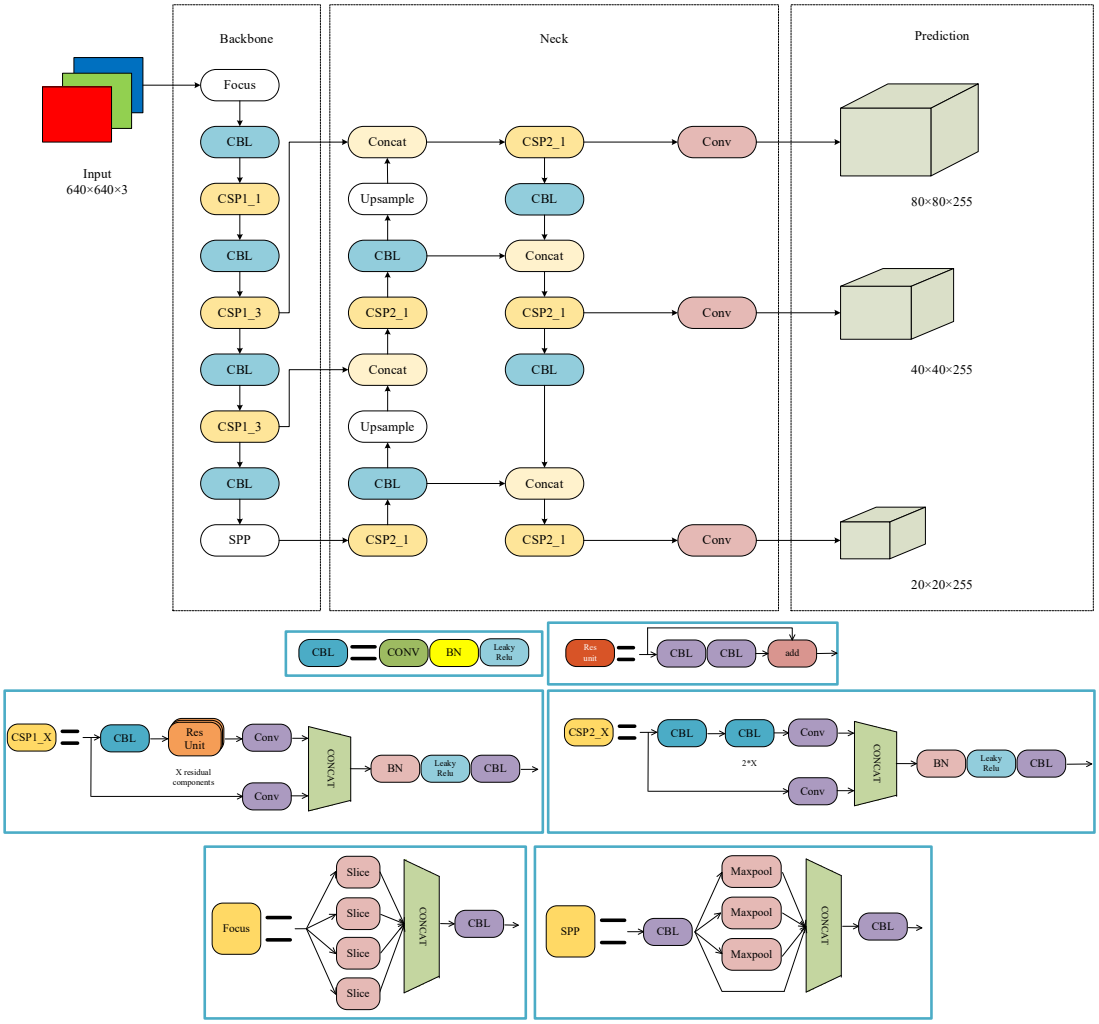

Figure 3. YOLOv5s network structure.

The optimization of the YOLOv5s algorithm for helmet detection can be divided into several aspects:

(1) Anchor box adjustment: The anchor box is adjusted using the K-means++ algorithm to place it closer to the actual target box, thereby improving the initial preselection box.
(2) Network structure improvement: An attention mechanism is incorporated to optimize the network topology, increasing the model's effectiveness and precision.
(3) Bounding box loss function optimization: The prediction box regression is made faster and more accurate by optimizing the loss function for the bounding box at the output end.
(4) The knowledge distillation technology is used to realize the light weight of the network model, thereby reducing the computational workload of the model, increasing the detection speed, and meeting the needs of real-time monitoring.

2.1. Improvement of Adaptive Anchor Frame Mechanism in YOLOv5 Based on K-means++ Algorithm

A core problem in computer vision is object recognition, which entails locating and recognizing items inside an image using bounding boxes. To increase the object recognition models' accuracy, selecting an appropriate prior bounding box during training can be beneficial. The YOLOv5 model incorporates the concept of an anchor box into target recognition. An initial bounding box with a defined size and aspect ratio is known as an anchor box. The anchor box's proximity to the ground-truth bounding box is taken

into account by the model when adjusting the predicted bounding box during training. K-means and genetic algorithms were employed to update the anchor boxes in the initial YOLOv5 model [28], with the Euclidean distance acting as the metric function. However, when working with samples of different sizes, utilizing Euclidean distance may result in clustering problems. To solve this problem, we suggest a hybrid method that groups anchor boxes using the K-means++ algorithm and the intersection-over-union (IOU) distance metric. This results in previous bounding boxes with a higher *IOU* value, improving object recognition accuracy.

The dimensions of the two boxes are represented by (w_1, h_1) and (w_2, h_2), respectively, as illustrated in Figure 4. The region highlighted in red denotes the intersection of the two boxes, with dimensions (w, h), and is defined as:

$$S_j = w \times h. \tag{1}$$

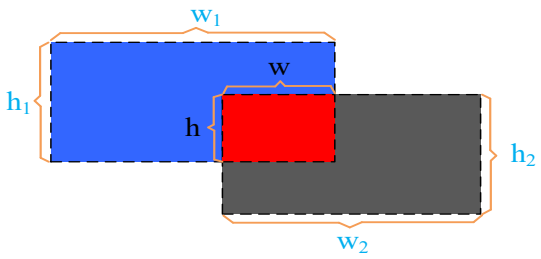

Figure 4. Intersection and comparison schematic diagram.

The combined area of the two frames is represented by the "blue + red + gray" region. This combined area can be calculated using a formula, which is rewritten as:

$$S_b = w1 \times h1 + w2 \times h2 - w \times h, \tag{2}$$

The *IOU* can be obtained based on Equations (1) and (2):

$$IOU = \frac{S_j}{S_b} \tag{3}$$

The overlap between two frames is measured by the *IOU*, a statistic that has a scale from 0 to 1. There is no gap between the two frames when the value is 0, while a value of 1 indicates that the two frames are identical. When the *IOU* value is higher, it indicates that the two previous frames fit better. To ensure that the measurement value and similarity have a negative correlation, when the measurement value is low, the similarity is high, and the value of the *IOU* is subtracted from 1. This gives rise to Equation (4), which calculates the similarity metric between two frames:

$$d_{iou} = 1 - IOU \tag{4}$$

In this paper, we utilized the K-means-based YOLOv5 algorithm in combination with the Euclidean distance measure method to derive 9 prior boxes. These prior boxes corresponded to feature maps of varying scales and had a matching degree of 0.8553. The prior boxes on feature maps of different scales are presented in Table 1.

Table 1. A priori box distribution following K-means clustering.

Feature Map Size	Anchor Frame Size		
80 × 80	(8, 10)	(11, 13)	(15, 18)
40 × 40	(24, 27)	(36, 42)	(54, 64)
20 × 20	(87, 101)	(155, 183)	(267, 339)

The linear growth in the computing complexity and quick convergence time of the K-means clustering method are two of its many benefits. The beginning clustering center must be predetermined for this approach, and various initial clustering centers may provide different clustering outcomes. To address this problem, we leveraged the K-means++ technique to calculate the anchor boxes in our object identification model. The first cluster center is chosen at random by the K-means++ algorithm, ensuring that the mutual distance between the initial cluster centers is as great as is feasible. When the initial n cluster centers (0 n K) have been chosen, the n + 1-th cluster center is chosen by giving sites further from the n cluster centers a greater likelihood. This approach helps to ensure that the anchor boxes are optimized for better accuracy and robustness in object recognition while mitigating the potential effects of initial clustering center selection.

To create previous boxes of feature maps with various scales for this investigation, we combined the IOU measurement method with the K-means++ algorithm. Our approach yielded a matching degree of 0.8689 for the previous frames, which was higher than that achieved by clustering with the K-means algorithm. The resulting prior box distribution is presented in Table 2.

Table 2. Algorithm clustering of the anchor block steps of K-means++.

Feature Map Size	Anchor Frame Size		
80 × 80	(8, 10)	(10, 12)	(15, 18)
40 × 40	(15, 18)	(21, 24)	(29, 33)
20 × 20	(42, 48)	(67, 77)	(120, 143)

2.2. Improvement of Network Structure

The attention mechanism seeks to identify relevant information and disregard irrelevant information, thereby enhancing the efficiency of neural networks. By obtaining detailed information and suppressing unnecessary data, it becomes possible to improve the network's performance [29,30]. In order to do this, we suggest a fusion approach that combines the cross-stage partial (CSP) module built into the convolutional block attention module (CBAM) attention mechanism with the global attentional map (GAM) mechanism. Our method attempts to improve the model's overall performance by strengthening its feature extraction capabilities.

(1) CBAM attention mechanism

A compact and adaptable module for strengthening neural networks is the convolutional block attention module (CBAM) [31]. In this study, the last layer of the cross-stage partial (CSP) modules in the backbone and neck of YOLOv5s includes the CBAM module. This integration enhances the model's ability to extract features while also lowering computational complexity.

The channel attention module and the spatial attention module are two sub-modules that make up the CBAM module. They are used in succession. From the deep network, we first achieve intermediate feature maps. The CBAM modules are then used at each convolutional block to adaptively improve these maps. The attention map is then successively inferred along the channel and space dimensions. To accomplish adaptive feature refinement, the output attention map is multiplied by the input feature map. In Figure 5, we can observe the detailed CBAM attention module of the proposed method.

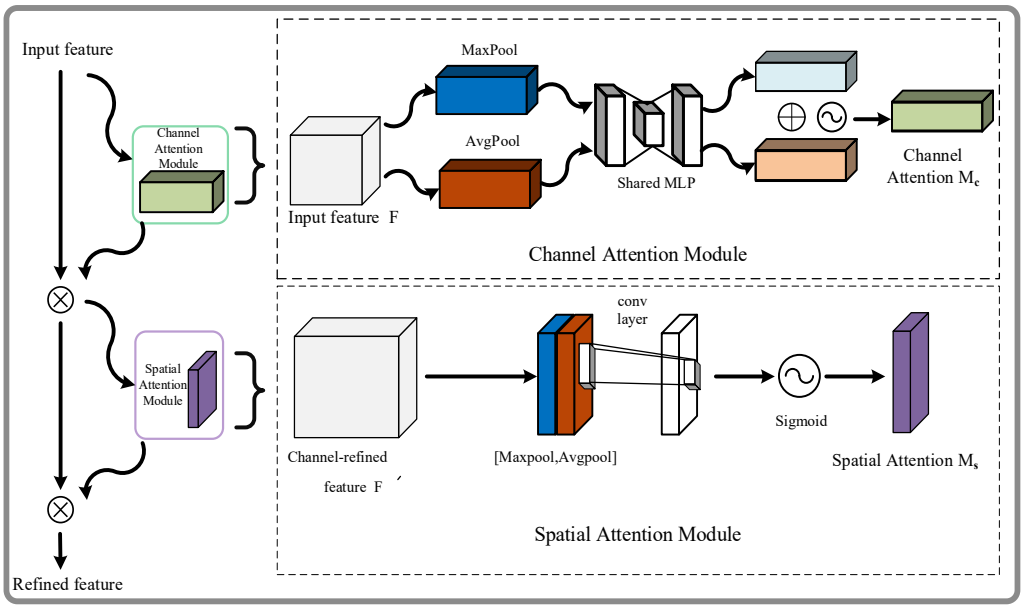

Figure 5. CBAM attention module.

The intermediate feature map $F \in R^{C \times H \times W}$ is the input for the CBAM module. A 1D channel attention map ($M_c \in R^{C \times 1 \times 1}$) and a 2D spatial attention map ($M_s \in R^{1 \times H \times W}$) are then obtained through sequential inference performed by the module. The mathematical representations of the attention process are given as:

$$F' = M_c(F) \otimes F \tag{5}$$

$$F'' = M_s(F') \otimes F' \tag{6}$$

where the symbol $\otimes$ denotes an element-level multiplication in the attention process. The spatial dimension is communicated together with the channel attention levels. F'' represents the refined output.

Notably, the feature map is compressed along the spatial dimension by the channel attention mechanism to produce a one-dimensional vector. The corresponding calculation for the channel attention is expressed as:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1(W_0(F^c_{max}))\right) \end{aligned} \tag{7}$$

The channel attention sub-module uses the shared network's maximum and average pooling outputs to generate an attention map, as shown in Figure 4. Two distinct spatial context descriptors, referred to as F^c_{avg} and F^c_{max}, are produced simultaneously by aggregating the spatial information of the feature maps using average pooling and max pooling. The average and maximum pooled characteristics are represented, respectively, by these two descriptors. The channel attention map $M_c \in R^{C \times 1 \times 1}$ is created by feeding the two feature maps into a common network of multi-layer perceptrons (MLPs). Next, $R^{C/r \times 1 \times 1}$ is chosen as the activation value size, where r is the reduction ratio and is a sigma function. The weights $W_0 \in R^{C/r \times C}$ and $W_1 \in R^{C \times C/r}$ of the MLP are shared with the ReLU activation function that comes after W_0.

The spatial attention mechanism compresses the channel by employing average pooling and maximum pooling in the channel dimension, which is formulated as:

$$\begin{aligned} \mathrm{M_s(F)} &= \sigma\left(f^{7\times7}([AvgPool(\mathrm{F}); MaxPool(\mathrm{F})])\right) \\ &= \sigma\left(f^{7\times7}\left(\left[\mathrm{F^s_{avg}}; \mathrm{F^s_{max}}\right]\right)\right) \end{aligned} \tag{8}$$

To aggregate the feature data of one feature map, two pooling operations—maximum pooling and average pooling—are conducted on the channel dimension, resulting in a dual-channel feature map. Specifically, the number of maximum pooling extractions is $H \times W$, and the number of average pooling extractions is also $H \times W$. Consequently, two 2D feature maps are obtained; the average pooling and maximum pooling characteristics throughout the whole channel are represented by the symbols $\mathrm{F^s}_{avg} \in \mathrm{R}^{1\times H\times W}$ and $\mathrm{F^s}_{max} \in \mathrm{R}^{1\times H\times W}$, respectively. To create a 2D spatial attention map, these two maps are combined and convolved using typical convolutional layers. The convolution operation, denoted as $f^{7\times7}$, employs a filter size of 7×7.

The feature map is compressed along the spatial dimension by the channel attention mechanism to produce a one-dimensional vector. The corresponding calculation for the channel attention is expressed in Equation (7).

(2) Global Attention Mechanism

The goal of the global attention module (GAM) is to improve neural network performance by reducing the loss of useful information and boosting the representation of global interactions. A convolutional spatial attention sub-module with multi-layer perceptions and a three-dimensional channel attention sub-module are introduced to accomplish this. As shown in Figure 5, the GAM uses the channel attention mechanism and the spatial attention mechanism juxtaposition technique, similar to the CBAM approach [32]. An intermediate state $\mathrm{F_2}$ and an output $\mathrm{F_3}$ are defined as follows, given an input feature map $\mathrm{F_1} \in \mathrm{R}^{C\times H\times W}$:

$$\mathrm{F_2} = \mathrm{M}_c(\mathrm{F_1}) \otimes \mathrm{F_1} \tag{9}$$

$$\mathrm{F_3} = \mathrm{M}_s(\mathrm{F_2}) \otimes \mathrm{F_2} \tag{10}$$

The symbols for the channel attention map and the spatial attention map are M_c and M_s, respectively, with element-level multiplication $\otimes$.

To preserve 3D information, the channel attention submodule utilizes 3D permutations. After that, it uses a two-layer multilayer perceptron (MLP) to improve the spatial and cross-dimensional relationships. The MLP is built using a compression ratio of r, and Figure 6 shows the channel attention submodule.

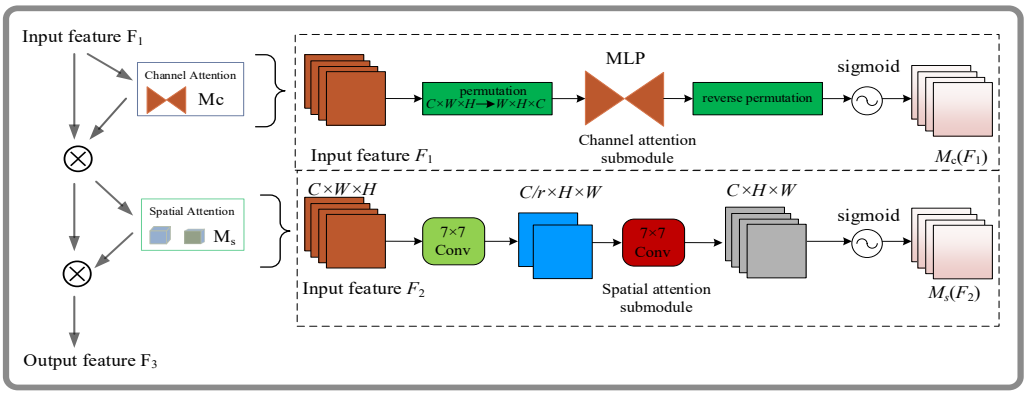

Figure 6. GAM attention module in our proposed method.

To fuse the spatial information, the spatial attention sub-module uses two convolutional layers and maintains the same compression ratio r as the channel attention sub-module.

By reducing feature loss and magnifying the representation of global interactions, the GAM attention mechanism improves the performance of the neural network. Here, we introduce a convolutional spatial attention submodule with multi-layer perceptron and a three-dimensional channel attention submodule. By embedding the CBAM module into the last layer of the CSP of the backbone and neck, the feature map undergoes adaptive refinement for each convolutional block of the deep network through the CBAM module. This process reduces the model's computational complexity and establishes high-dimensional spatial features' correlations, thereby facilitating the extraction of relevant features. The network structure incorporates the GAM and CBAM attention mechanisms, as illustrated in Table 3.

Table 3. Improved YOLOv5s network structure diagram.

Number	From	Params	Module	Arguments
0	−1	3520	Focus	[3, 32, 3]
1	−1	18,560	Conv	[32, 64, 3, 2]
2	−1	18,816	C3	[64, 64, 1]
3	−1	73,984	Conv	[64, 128, 3, 2]
4	−1	156,928	C3	[128, 128, 3]
5	−1	295,424	Conv	[128, 256, 3, 2]
6	−1	625,152	C3	[256, 256, 3]
7	−1	1,639,680	GAM Attention	[256, 256]
8	−1	1,180,672	Conv	[256, 512, 3, 2]
9	−1	656,896	SPP	[512, 512, [5, 9, 13]]
10	−1	1,215,586	CBAMC3	[512, 512, 1, False]
11	−1	131,584	Conv	[512, 256, 1, 1]
12	−1	0	Upsample	[None, 2, 'nearest']
13	[−1, 6]	0	Concat	[1]
14	−1	361,984	C3	[512, 256, 1, False]
15	−1	33,024	Conv	[256, 128, 1, 1]
16	−1	0	Upsample	[None, 2, 'nearest']
17	[−1, 4]	0	Concat	[1]
18	−1	90,880	C3	[256, 128, 1, False]
19	−1	147,712	Conv	[128, 128, 3, 2]
20	[−1, 15]	0	Concat	[1]
21	−1	296,448	C3	[256, 256, 1, False]
22	−1	1,639,680	GAM Attention	[256, 256]
23	−1	590,336	Conv	[256, 256, 3, 2]
24	[−1, 11]	0	Concat	[1]
25	−1	1,215,586	CBAMC3	[512, 512, 1, False]

Table 3 shows the number of input source layers in the "from" column and the number of parameters in the "params" column. The "arguments" column lists information on the number of input and output channels, convolution kernel size, step size, and other relevant specifics. The "module" column lists the name of the module.

As shown in Figure 7, we used Grad-CAM to display the model's heat map characteristics. The visualization demonstrates the requirement for the original YOLOv5s model feature extraction to be more coherent and suitable for small targets. However, after incorporating the combined attention mechanism, the model focuses on extracting critical information, reduces attention to irrelevant details, and remarkably enhances the feature extraction of small objects.

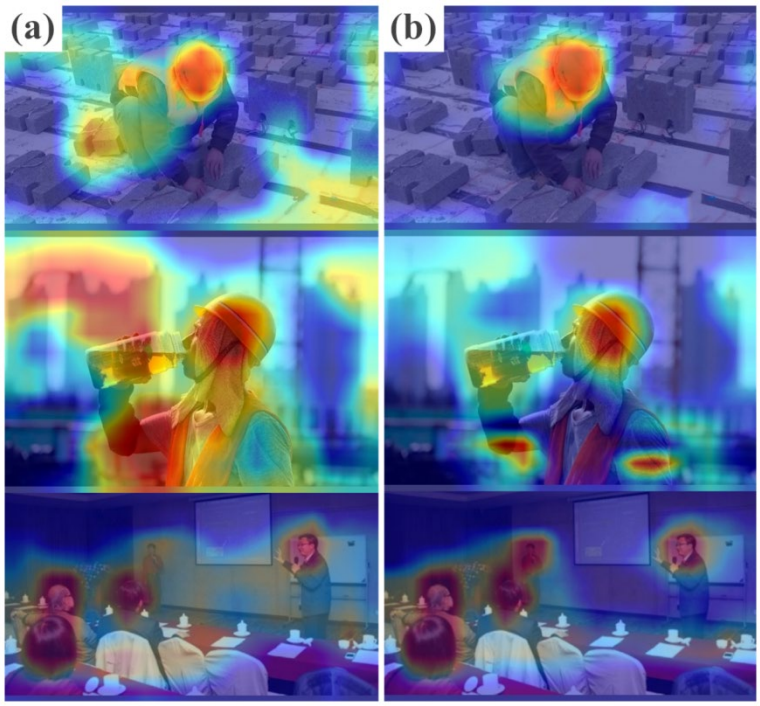

Figure 7. Comparison of model thermal map feature visualization output before and after adding attention mechanism. (**a**) Original YOLOv5s model. (**b**) Model after adding the combined attention mechanism.

2.3. Bounding Box Loss Function

Object detection accuracy and effectiveness are heavily reliant on the loss function employed. Traditional object detection loss functions are based on aggregating bounding box regression metrics. However, the distance between the expected target box and the predicted box, the overlapping area, and aspect ratio are a few of the characteristics that greatly affect aggregation accuracy. Some examples include the fact that YOLOv5's GIoU, CIoU, etc. do not take into consideration the direction discrepancy between the desired target box and the forecast box, resulting in a slower convergence speed and poorer model performance [33]. On the other hand, the SCYLLA-IoU LOSS (SIoU) [34] considers the vector' angle between the regressed boxes and the orientation discrepancy between the anticipated box and the required item box, resulting in increased detection precision.

Conventional object detection loss algorithms are considerably improved by the SIoU loss function, as it not only considers the angle and distance between the regressed boxes, but also addresses the orientation mismatch between the predicted and desired object boxes. This improves training effectiveness and ultimately enhances target box regression's stability, resulting in a more accurate model. The angle cost, distance cost, shape cost, and IoU cost make up the SIoU loss function.

(1) Angle cost

An extra term, LF, in the SioU loss function integrates an adaptive angle adjustment function and greatly lowers the number of variables linked to distance. As seen in Figure 8, the model first lines up the predicted box with either the X or Y axis (whichever is closest), and then it optimizes the distance along the pertinent axis.

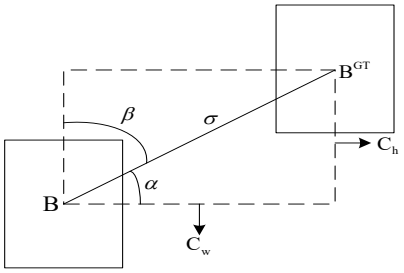

Figure 8. Effect of the angular factors on the loss function.

When $\alpha \leq \Pi/4$, minimize α, and when $\alpha > \Pi/4$, minimize β. The definition of LF is obtained, which can be constructed as:

$$\Lambda = 1 - 2 * sin^2\left(arcsin(x) - \frac{\pi}{4}\right), \tag{11}$$

where

$$x = \frac{c_h}{\sigma} = sin(\alpha), \tag{12}$$

$$\sigma = \sqrt{\left(b_{c_x}^{gt} - b_{c_x}\right)^2 + \left(b_{c_y}^{gt} - b_{c_y}\right)^2}, \tag{13}$$

$$c_h = max\left(b_{c_y}^{gt}, b_{c_y}\right) - min\left(b_{c_y}^{gt}, b_{c_y}\right). \tag{14}$$

(2) Distance cost

Based on the redefined angle cost, SIoU defines the distance cost:

$$\Delta = \sum_{t=x,y} \left(1 - e^{-\gamma \rho_t}\right), \tag{15}$$

where

$$\rho_x = \left(\frac{b_{c_x}^{gt} - b_{c_x}}{c_w}\right)^2, \ \rho_y = \left(\frac{b_{c_y}^{gt} - b_{c_y}}{c_h}\right)^2, \gamma = 2 - \Lambda, \tag{16}$$

Equations (12) to (16) show that the effect of distance cost on the output decreases noticeably as the value of α approaches 0. Conversely, as α approaches $\Pi/4$, the impact of the distance cost on the output becomes more significant.

(3) Shape cost

A definition of the shape cost function is:

$$\Omega = \sum_{t=w,h} \left(1 - e^{-\omega_t}\right)^{\theta}, \tag{17}$$

where

$$\omega_w = \frac{\left|w - w^{gt}\right|}{max(w, w^{gt})}, \tag{18}$$

$$\omega_h = \frac{\left|h - h^{gt}\right|}{max(h, h^{gt})} \tag{19}$$

The value of θ can have a variety of effects on the shape cost depending on the shape of each dataset. To ascertain the relative significance of the form cost, a certain value of θ is determined. A genetic algorithm is used during training to determine the ideal value of θ for each dataset.

(4) IoU cost

The IoU cost is described as:

$$IoU = \frac{|B \cap B^{GT}|}{|B \cup B^{GT}|} \tag{20}$$

The L_{box} regression loss function is formulated as:

$$L_{box} = 1 - IoU + \frac{\Delta + \Omega}{2} \tag{21}$$

The total loss function is constructed as:

$$L = W_{box} L_{box} + W_{cls} L_{cls} \tag{22}$$

To calculate the loss function, we used a genetic algorithm to determine the values of W_{box}, W_{cls}, and θ. L_{cls} represents the focal loss, while W_{box} and W_{cls} are the weights for the prediction box and classification loss, respectively. Moreover, we chose a small subset from the training set and computed these values iteratively until the number of iterations was either below a threshold or the maximum number was achieved, at which time the iterations were terminated.

2.4. Knowledge Distillation

Knowledge distillation is a technique utilized to extract the knowledge of a large teacher model and condense it into a small student model. It can be understood as a large teacher neural network teaching his knowledge to a small student network [35–37].

The process is transferred from the teacher network to the student network. The teacher network is generally bloated; therefore, the teacher network provides knowledge to the student network. The student network is a relatively small network and can thus obtain a lightweight network model. Knowledge distillation adopts the teacher–student mode. In this mode, the teacher is the output party of "knowledge", and the student is the receiver of "knowledge" [38].

The teacher has a strong learning ability and can transfer the learned knowledge to the student model with a lower learning ability, so as to improve the generalization ability of the student model. The complicated and cumbersome but easy-to-use teacher model has no upper limit; it is purely a tutor, and in reality, a simple and flexible student model is deployed. The knowledge distillation process is shown in Figure 9 below.

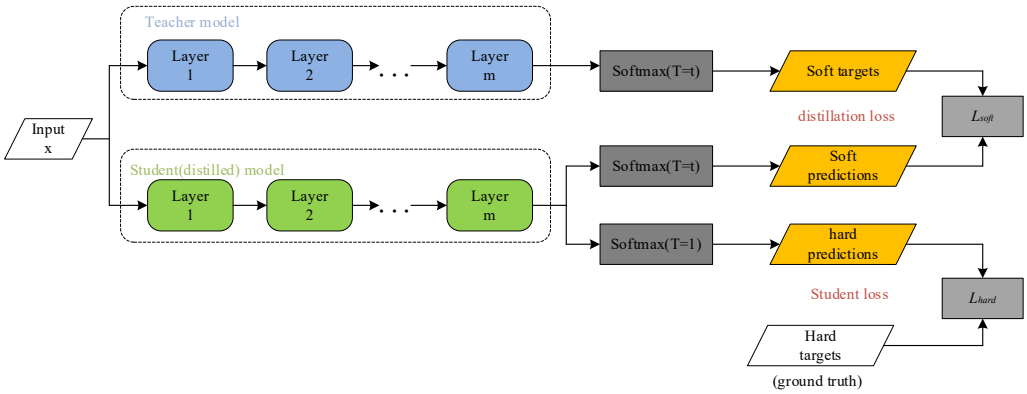

Figure 9. Schematic diagram of knowledge distillation process.

First, distill a deeper teacher network with a better extraction ability to obtain a logit, and distill it at T temperature. Then, use the classification prediction probability distribution in the Softmax layer to obtain soft targets. At the same temperature T, the

logits in the student network are distilled, and then the category prediction probability distribution in Softmax is used to obtain the loss function L_{soft}. Its expression is:

$$L_{soft} = -\sum_{j}^{N} p_j^T \mathrm{lb}\left(q_j^T\right) \tag{23}$$

where C_j is the true label value of the j-th class.

Finally, L_{hard} and L_{soft} are weighted and summed to obtain the final loss function L. This loss function can prevent the wrong information from the teacher network from being transmitted to the student network by comparing it with the real label. In this study, the improved YOLOv5s model was used as the teacher network, and the YOLOv5s model with the large target detection layer removed by structural pruning was used as the student model for knowledge distillation to obtain the final model and reduce the amount of calculation and parameters of the improved network model.

3. Performance Analysis

3.1. Dataset and Experimental Environment

The dataset used in this study consists of 14,966 images extracted from video streams, comprising 7000 images from the public Safety Helmet Wearing and Head Detection (SHWD) dataset and 7966 images of extracted video stream frames. The images are divided into two categories: person and hat. The training set comprises 11,973 images, and the validation set comprises 2993 images, with an 8:2 ratio of training to validation data. Using two NVIDIA RTX 3060 graphics cards and the Linux operating system, the tests were carried out. Using the CUDA 11.1 computing architecture and the Pytorch deep learning framework, we built, trained, and validated our models. The batch size was 32, the workers were 8, and the image resolution was 640 × 640. The model was trained for 300 epochs with a learning rate of 0.001. The results achieved using these settings are displayed in Table 4.

Table 4. Experimental parameters.

Parameter	Value
Lr0	0.01
Lrf	0.2
Warmup_epochs	3
Batchsize	32

3.2. Evaluation Criteria

As assessment measures for our model in this article, we use precision (P), recall (R), mean average precision (mAP), and detection speed (FPS). Precision and recall are defined in Equations (24) and (25).

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{24}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{25}$$

True positives, or TP, in this context refers to the total number of accurately identified items. False positives, or FP, on the other hand, are the quantity of items that were mistakenly identified. Last but not least, FN stands for false negatives and denotes the quantity of items the model failed to detect. These assessment metrics offer insightful information about the model's functionality and precision in object detection.

According to Equation (26), the average accuracy (AP) denotes the average accuracy rate under various recall rates.

$$\mathrm{AP} = \int_0^1 \mathrm{P}(\mathrm{R})\mathrm{d}R \tag{26}$$

According to Equation (27), this yields the mean average precision (mAP).

$$\text{mAP} = \frac{\sum_{i=}^{N} \text{AP}_i}{\text{N}} \tag{27}$$

Here, N stands for the total number of categories, and n stands for the category.

While the IOU threshold is set to 0.5, the average AP is represented by the mAP@0.5. The average value of mAP while the IOU threshold varies from 0.5 to 0.95 in steps of 0.05 is represented by the mAP@0.5:0.95 value.

The F1-score is used to comprehensively evaluate the recall and accuracy indicators, as shown in Equation (28):

$$\text{F1} - \text{score} = \frac{2\text{TP}}{(\text{Total number of samples} + \text{TP} - \text{TN})} \tag{28}$$

The number of pictures detected per second is indicated by the detection speed (FPS).

3.3. Ablation Experiments

To examine how various loss functions affect the YOLOv5s algorithm, we conducted experiments using commonly used loss functions, such as GIoU, CioU, DioU [39], EioU [40], and SioU. The training accuracy obtained after replacing the original GioU loss of the YOLOv5s algorithm with different loss functions is shown in Table 5.

Table 5. Comparison of experimental results with different loss functions.

Loss Function	Precision (P)/%	Recall (R)/%	mAP0.5/%	mAP0.5:0.95/%
GIoU	0.903	0.862	0.898	0.571
CIoU	0.906	0.865	0.908	0.58
DIoU	0.891	0.866	0.904	0.575
EIoU	0.891	0.868	0.91	0.582
SIoU	0.906	0.876	0.913	0.586

In Table 5, we list the training accuracies of various loss functions, including GIoU, CIoU, DIoU, EIoU, and SIoU, used in the YOLOv5s algorithm. The findings demonstrate that, in comparison to other loss functions, SIoU has significantly increased the precision rate (P), recall rate (R), and mAP. Conventional object detection loss functions primarily depend on combining bounding boxes regression variables, such as the separation between the anticipated object box and the predicted box, the area that overlaps with the predicted box, and the aspect ratio. However, the use of GIoU, CIoU, etc. by YOLOv5 ignores the mismatch between the desired target frame and the prediction frame, resulting in sluggish convergence and causing the prediction frame to fluctuate throughout training. Ultimately, a poor model is produced. Using the SIoU, the vector's angle between the regression boxes and the mismatch direction between the predicted and expected target boxes is considered, thereby changing the calculation method.

Based on the original YOLOv5s model, this study conducted ablation experiments to verify each improvement's impact on model training. Table 6 displays the trial outcomes. Precision, recall, mAP, and mAP@0.5:0.95 of 90.3%, 86.2%, 89.8%, and 57.1%, respectively, were attained by the original YOLOv5s model. YOLOv5s-K increased recall by 0.7% and mAP by 0.7% in comparison to the original YOLOv5s algorithm, and mAP@0.5:0.95 by 1.4%. Using the K-means++ algorithm measurement method to adjust the prior frame improved the matching degree of the set target box with the preceding frame and data. From the beginning, YOLOv5s-KS significantly improved recall by 0.8% and mAP by 1.8%, compared to the YOLOv5s method. This increase is attributable to YOLOv5s-KS's large improvement in precision, which was made possible by taking into account the vector's angle between the regression boxes and the mismatch between the target and prediction

frames while utilizing SIoU as the bounding box loss function. Comparing the upgraded model to the original YOLOv5s model, the better model saw gains in precision, recall, mAP, and mAP@0.5:0.95 of 1%, 1.1%, 2.6%, 2.1%, and 0.95, respectively. The performance of the deep neural network was enhanced by the addition of the GAM attention mechanism and the combined attention mechanism of the CSP module incorporated in the CBAM attention mechanism in the backbone and neck. This improvement was made possible by lowering the feature loss, enhancing the representation of global interactions, and adding a multi-layer three-dimensional arrangement of the channel attention sub-module and the convolutional space attention sub-module, which enhanced the efficiency of object feature extraction.

Table 6. Comparison of the ablation experiment results.

Model	K-Means++	SIoU	GAM	CBAM	Precision/%	Recall/%	mAP0.5/%	mAP0.5:0.95/%
YOLOv5s	×	×	×	×	90.3	86.2	89.8	57.1
YOLOv5s-K	√	×	×	×	89.7	86.9	90.5	58.5
YOLOv5s-KS	√	√	×	×	90.6	87.0	91.6	58.5
YOLOv5s-KSG	√	√	√	×	89.9	87.5	92.0	58.9
YOLOv5s-KSGC	√	√	√	√	91.3	87.3	92.4	59.2

The improved YOLOv5s-KSGC model was leveraged as the teacher network, and the YOLOv5s model with the large target detection layer removed by structural pruning was utilized as the student model for knowledge distillation to obtain the final model. The experimental effect comparison between the improved YOLOv5s-Improved model and the original YOLOv5s is shown in Table 7 below. It can be seen from Table 7 that the improved model not only reduces the number of parameters and model size, but also effectively improves other indicators. Among them, mAP0.5 increased by 2.6%, mAP@0.5: 0.95 increased by 2.1%, and FPS increased by 9.33.

Table 7. Comparison of the experimental effects before and after the improvement.

Model	Image Size	Params/MB	Model Size/MB	mAP0.5/%	mAP@0.5:0.95/%	FPS
YOLOv5s	640 × 640	7.06	14.4	89.8	57.1	133.33
YOLOv5s-Improved	640 × 640	5.06	12.5	92.4	59.2	142.66

Figure 10a–c shows the training loss of the original YOLOv5s model and the improved YOLOv5s-Improved model. Figure 10a is the Box_Loss obtained from training. It can be seen from Figure 10a that the Box_Loss of the improved model is much lower than the loss of the original YOLOv5s model training. It can be seen from Figure 10b that the Cls_Loss of the original YOLOv5s model fluctuates greatly, and the improved model significantly improves the fluctuation of Cls_Loss and reduces the loss value. It can be seen from Figure 10c that the Obj_Loss of the improved YOLOv5s-Improved model is also lower than the original YOLOv5s model at the beginning, and finally tends to be equal. The experiments prove that the improved model is reliable and stable and has higher robustness.

As can be seen from Figure 11, although the YOLOv5s-Improved and YOLOv5 training effect is good, both demonstrate the overfitting and underfitting phenomena. However, the modified model greatly increased the average accuracy in comparison to the old model, proving the viability of the revised technique.

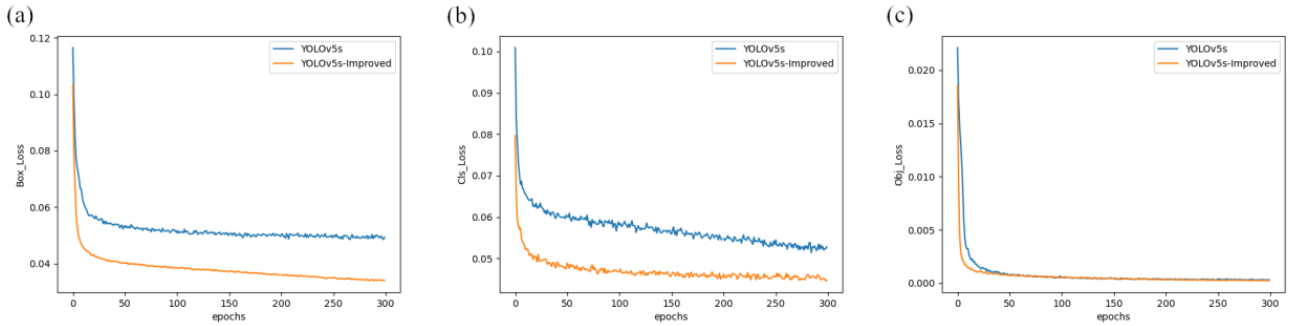

Figure 10. Loss comparison between this paper's methodology with YOLOv5s. (**a**) Box_Loss comparison diagram of training results. (**b**) Cls_Loss comparison diagram of training results. (**c**) Obj_Loss comparison diagram of training results.

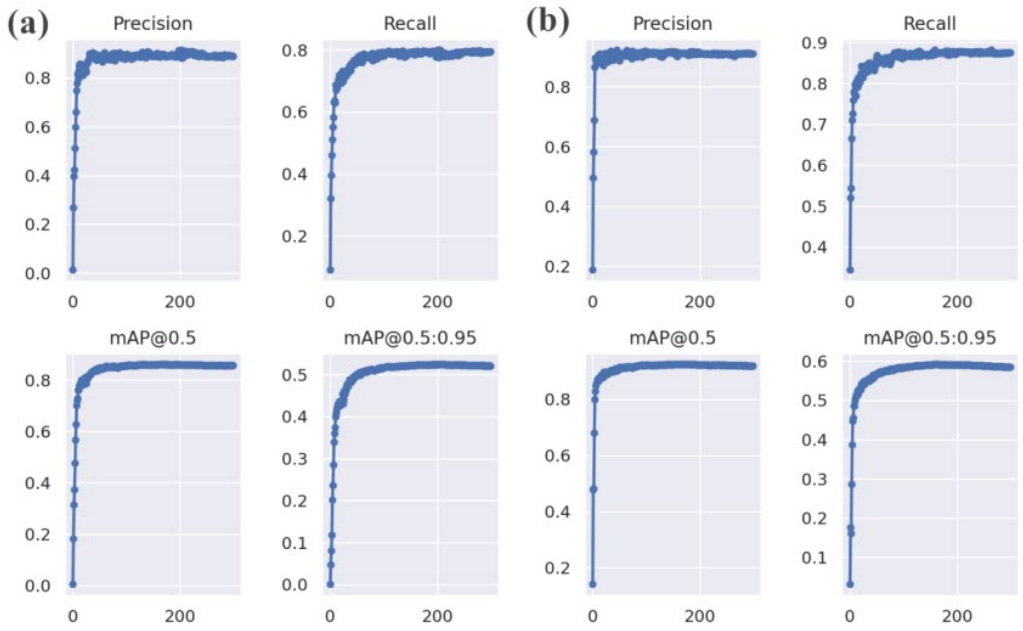

Figure 11. Training results for the different methods. (**a**) YOLOv5s model. (**b**) YOLOv5s-Improved model.

3.4. Comparison of Different Methods

The YOLOv5s-Improved model was trained on the dataset to assess the performance of the suggested approach, and the outcomes were compared with those of other cutting-edge object identification models, such as SSD, Faster-RCNN, YOLOv3, ML-YOLOv3 [41], YOLOv4, YOLOv5s, YOLOv5u, YOLOv5s-DM [42], YOLOv6s, and YOLOv7-w6 [43]. Table 8 displays the experimental findings.

Table 8. Comparison table of training results of different algorithms.

Model	Image Size	Params/MB	Model Size/MB	P/%	R/%	F1-Score/%	mAP0.5/%	mAP@0.5:0.95/%	FPS
SSD	512 × 512	41.18	200	83.5	78.9	41.8	77.5	/	48.5
Faster-RCNN	1000 × 600	60.17	420	87.5	82.3	44.5	82.6	/	11.6
YOLOv3	640 × 640	61.95	123.5	86.5	84.5	46.8	92.3	60.7	37.59
ML-YOLOv3	640 × 640	18.15	37	89.5	85.8	47.5	90.2	58.2	88.88
YOLOv4	640 × 640	52.92	112.6	87.2	85.1	47.4	91.5	58.2	45.3
YOLOv5s	640 × 640	7.06	14.4	90.3	86.2	48.2	89.8	57.1	133.33
YOLOv5u	640 × 640	6.52	11.4	89.8	86.2	47.5	88.5	56.2	136.5
YOLOv5s-DM	640 × 640	7.06	14.4	90.5	86.1	48.8	90.2	56.5	133.33
YOLOv6s	640 × 640	/	38.16	89.5	85.8	48.3	90.9	57.9	79
YOLOv7-w6	640 × 640	69.83	140.1	90.1	87.2	48.5	91.7	58.8	55.4
YOLOv5s-Improved	640 × 640	5.06	12.5	91.2	87.1	51.5	92.4	59.2	142.66

It can be seen from Table 8 that the improved YOLOv5s model has significantly improved mAP and FPS compared with the SSD model, Fast-RCNN model, and Faster-RCNN model under the premise of maintaining a light weight. Although the model mAP proposed in this study is similar to the YOLOv3 model and YOLOv4 model, their parameter quantity and model size are much larger than the model proposed in this paper, and the detection speed is much lower than the model in this paper. When comparing YOLOv6s, which is also a lightweight model, the parameter quantity and model size of the model in this paper are lower than YOLOv6s, and the detection accuracy speed is also higher than that of the YOLOv6s model. Due to the large number of parameters of the YOLOv7 model, this paper chose the YOLOv7-w6 model with fewer parameters for experimental comparison. It is proved by the experiments that the parameter quantity of this method is lower than that of the YOLOv7-w6 model, the detection accuracy is slightly higher than that of the YOLOv7-w6 model, and the detection speed is much higher than that of the YOLOv7-w6 model. Additionally, the method in this paper scores higher than other methods in precision, recall, and the F1-score. Compared with the ML-YOLOv3 model, the method in this paper not only has obvious advantages in precision, recall, the F1-score, and the map, but the model size and parameters are also lower than the ML-YOLOv3 model, and the detection speed is also higher than the ML-YOLOv3 model. Compared with the anchor-free YOLOv5u model, it is not as effective as the original YOLOv5s model on the helmet dataset. The advantages of the method in this paper are also superior to the YOLOv5s-DM model in precision, recall, F1-score, map, model size, and parameter quantity. Compared with the original YOLOv5s model, the results of the method in this paper show that the four indicators of precision (P), recall (R), mAP, and detection speed FPS are better than the original YOLOv5s, and have higher accuracy and detection speed. This reflects the excellent performance of the method in this paper.

4. Case Analysis

In this study, the proposed method was practically applied to detect helmets in various indoor and outdoor scenes at different distances. Furthermore, the detection results were compared with those obtained from the SSD, Faster R-CNN, YOLOv5s, YOLOv6s, YOLOv7-w6, and YOLOv5s-Improved models.

The results of the six approaches' detection in a scenario with sunshine are shown in Figure 12. The detection results show that the suggested approach can identify two types of small target items in a bright outdoor setting. Figure 12a in particular illustrates the SSD model's detection impact, with detected confidence values for the hat and person being 0.74 and 0.80, respectively. Similarly, Figure 12a,b shows the Faster R-CNN model's detection impact, with detected confidence values of 0.75 and 0.82 for the hat and person, respectively. Figure 12c depicts the YOLOv5s model's detection impact, with detected confidence values for the hat and person of 0.76 and 0.85, respectively. Figure 12d shows the YOLOv6s model's detection effect, with detected confidence values of 0.77 and 0.86 for the hat and person, respectively. Figure 12e illustrates the YOLOv7-w6 model's detection impact, with detected confidence values for the hat and person of 0.78 and 0.88, respectively. The detection impact of the suggested model is finally shown in Figure 12f, where the detected confidence values for the hat and person are 0.81 and 0.92, respectively. The detection findings indicate that the suggested approach, when used in an outdoor setting sunshine, provides much greater detection accuracy than previous target detection methods.

Figure 13 shows the detection results of the six methods in outdoor shaded scenes. The results show that two classes of small target objects are detected in the outdoor shadow environment. Figure 13a depicts the SSD model's detection effect, and the detection's confidence scores are 0.73 for the hat and 0.78 for the person. Figure 13b illustrates the Faster R-CNN model's detection impact, and the detection's confidence scores are 0.74 for hats and 0.79 for people. Similarly, Figure 13b,c displays the YOLOv5s model's detection impact, with the detected confidence values of 0.75 for hats and 0.80 for people. Additionally, Figure 13d demonstrates the YOLOv6s model's detection impact. The detected confidence

is 0.77 for the hat and 0.82 for the person. In contrast, 13e displays the YOLOv7-w6 model's detection effect, and the detection's confidence scores are 0.79 for hats and 0.84 for people. The confidence gained by the detection is 0.83 for the hat and 0.85 for the person, and Figure 13f demonstrates the detection impact of the suggested approach. One may draw the conclusion that the suggested technique outperforms previous target identification algorithms in the outdoor shadow environment, leading to increased detection accuracy.

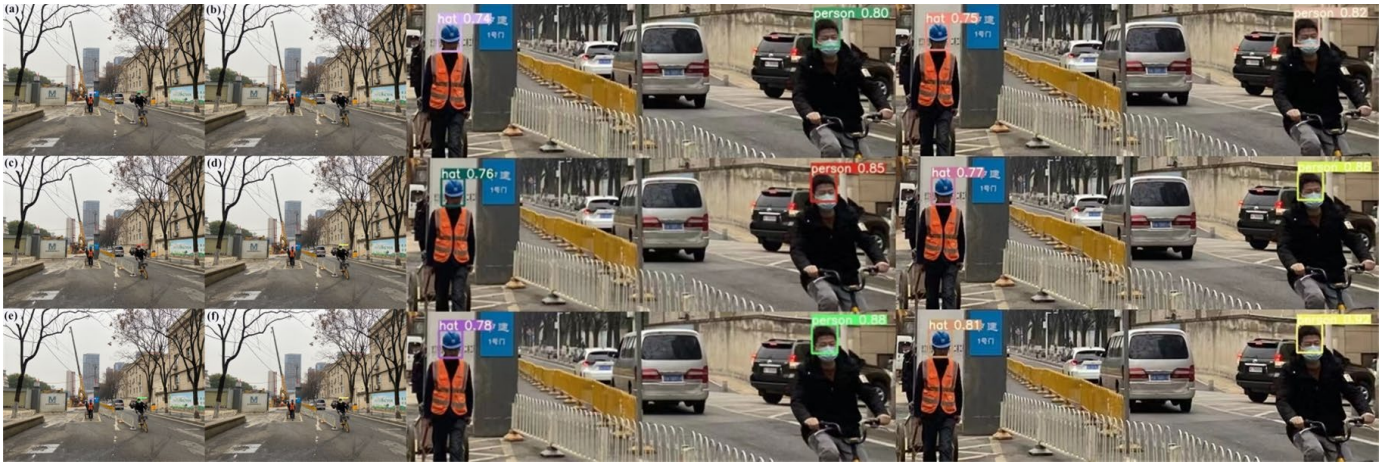

Figure 12. Schematic comparison of actual detection effect of different detection algorithms in outdoor environment. (**a**) The detection effect diagram of the SSD model. (**b**) The detection effect diagram of the Faster R-CNN model. (**c**) The detection effect diagram of the YOLOv5s model. (**d**) The detection effect diagram of the YOLOv6s model. (**e**) The YOLOv7-w6 model detection effect diagram. (**f**) The YOLOv5s-Improved model detection effect diagram.

Figure 13. Comparison of different detection algorithms' real detection results in darkened situations. (**a**) The detection effect diagram of the SSD model. (**b**) The detection effect diagram of the Faster R-CNN model. (**c**) The detection effect diagram of the YOLOv5s model. (**d**) The detection effect diagram of the YOLOv6s model. (**e**) The YOLOv7-w6 model detection effect diagram. (**f**) The YOLOv5s-Improved model detection effect diagram.

Figure 14 shows the detection results of the six methods in indoor scenarios. It can be seen from the detection results in Figure 14 that in the indoor environment, two types of small target objects are detected. Figure 14a is the detection effect of the SSD model, and the confidence obtained by the detection is 0.89 for hats and 0.70 for people. Figure 14b is the detection effect of the Faster R-CNN model, and the confidence obtained by the detection is 0.92 for the hat and 0.71 for the person. In Figure 14c, we show the detection effect of the YOLOv5s model, and the detected confidence for hats is 0.92 and 0.72 for people.

Figure 14d shows the detection effect of the YOLOv6s model, and the detected confidence is 0.93 for hats and 0.73 for people. Figure 14e is the detection effect of the YOLOv7-w6 model, and the confidence obtained by the detection is 0.93 for the hat and 0.74 for the person. Figure 14f is the detection effect of the model in this paper, and the confidence obtained by the detection is 0.95 for hats and 0.80 for people. It is concluded that the detection accuracy of this method is higher than that of other target detection algorithms in different indoor and outdoor environments, which proves the feasibility and effectiveness of the improvement. Due to the improvement of the original anchor box mechanism of YOLOv5, the matching degree between the preselected box and the target box has been increased. The attention mechanism has been added to increase the extraction of effective target information features. The loss function has been improved to effectively increase the speed of prediction box regression and precision. The experiments have proved that the method in this paper can be applied to helmet detection in various scenarios, and the detection accuracy has reached more than 90%. The higher the detection accuracy, the higher the detection efficiency in actual deployment. Finally, knowledge distillation is used to reduce the number of parameters and the model size and increase the detection speed, and is more conducive to the deployment of the model.

Figure 14. Comparison of the actual detection effect of different detection algorithms in indoor environment. (**a**) The detection effect diagram of the SSD model. (**b**) The detection effect diagram of the Faster R-CNN model. (**c**) The detection effect diagram of the YOLOv5s model. (**d**) The detection effect diagram of the YOLOv6s model. (**e**) The YOLOv7-w6 model detection effect diagram. (**f**) The YOLOv5s-Improved model detection effect diagram.

5. Conclusions

In this paper, we proposed a modified YOLOv5 network to adaptively adjust the anchor box to increase the matching degree between the anchor box and the target box, which can extract discriminative image features from small targets. In the proposed method, the GAM attention mechanism is combined with the CPS module of the CBAM attention mechanism. It is added to the backbone network (Backbone) and neck network (Neck) of the original YOLOv5s network to improve the performance of the neural network by reducing the loss of feature information and amplifying the global interaction. This article introduces a three-dimensionally arranged channel attention and convolutional

spatial attention sub-module with a multi-layer perceptron, and the feature map adaptively refines each convolutional block of the network structure through a combination module, which is conducive to the establishment of high dimensional spatial feature correlation and the extraction of effective features of the target. While introducing the attention mechanism, the latest SIoU LOSS is used as the bounding box loss function at the output end, which effectively improves the speed and accuracy of the prediction box regression. The experiments prove that the improved network structure has higher performance. Finally, knowledge distillation is used to realize a lightweight network to obtain the final model, which reduces the amount of computation and parameters of the improved network model and improves the detection speed FPS, which is more conducive to the deployment of the model.

According to the experimental findings, the suggested strategy enhances the accuracy indicators and mean average precision (mAP) acquired from training on the hard hat dataset. Further evidence that our technique may greatly increase the detection accuracy of small targets while meeting real-time detection requirements is shown by the large improvement in the confidence level attained by actual detection.

Author Contributions: Conceptualization, Q.A. and Y.X.; methodology, Y.X.; software, J.Y. and M.T.; validation, T.L.; formal analysis, F.X.; investigation, F.X.; resources, J.Y.; data curation, Y.X. and J.Y.; writing—original draft preparation, Q.A.; writing—review and editing, Y.X.; funding acquisition, Q.A. and Y.X. 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 No. 62277041, and in part by the technology project of the Hubei Province Safety Production special fund (Program SN: SJZX 20211006) and the Opening Foundation of State Key Laboratory of Cognitive Intelligence, iFLYTEK (CIOS-2022SC07).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

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

References

1. Liu, P.; Li, Q.; Bian, J.; Song, L.; Xiahou, X. Using Interpretative Structural Modeling to Identify Critical Success Factors for Safety Management in Subway Construction: A China Study. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1359. [CrossRef] [PubMed]
2. Jia, W.; Xu, S.; Liang, Z.; Zhao, Y.; Min, H.; Li, S.; Yu, Y. Real-time automatic helmet detection of motorcyclists in urban traffic using improved YOLOv5 detector. *IET Image Process.* **2021**, *15*, 3623–3637. [CrossRef]
3. Kartik, B.; Manimaran, P. IOT based Smart Helmet for Hazard Detection in mining industry. *arXiv* **2023**, arXiv:2304.10156.
4. Zhang, C.; Liu, H.; Deng, Y.; Xie, B.; Li, Y. TokenHPE: Learning Orientation Tokens for Efficient Head Pose Es-timation via Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; Volume 32, pp. 8897–8906.
5. Liu, H.; Zhang, C.; Deng, Y.; Xie, B.; Liu, T.; Zhang, Z.; Li, Y.-F. TransIFC: Invariant Cues-aware Feature Concentration Learning for Efficient Fine-grained Bird Image Classification. *IEEE Trans. Multimed.* **2023**, 1–14. [CrossRef]
6. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. *IEEE CVPR* **2001**, *15*, 66–82.
7. Viola, P. Robust real-time object detection. In Proceedings of the International Workshop on Statistical and Computational Theories of Vision—Modeling, Learning, Computing, and Sampling, Vancouver, BC, Canada, 13 July 2001.
8. Mahum, R.; Rehman, S.U.; Meraj, T.; Rauf, H.T.; Irtaza, A.; El-Sherbeeny, A.M.; El-Meligy, M.A. A Novel Hybrid Approach Based on Deep CNN Features to Detect Knee Osteoar-thritis. *Sensors* **2021**, *21*, 6189. [CrossRef]
9. An, Q.; Chen, X.; Zhang, J.; Shi, R.; Yang, Y.; Huang, W. A Robust Fire Detection Model via Convolution Neural Networks for Intelligent Robot Vision Sensing. *Sensors* **2022**, *22*, 2929. [CrossRef]
10. Liu, T.; Wang, J.; Yang, B.; Wang, X. NGDNet: Nonuniform Gaussian-label distribution learning for infrared head pose es-timation and on-task behavior understanding in the classroom. *Neurocomputing* **2021**, *436*, 210–220. [CrossRef]
11. Liu, H.; Nie, H.; Zhang, Z.; Li, Y.-F. Anisotropic angle distribution learning for head pose estimation and attention under-standing in human-computer interaction. *Neurocomputing* **2021**, *433*, 310–322. [CrossRef]
12. Liu, H.; Fang, S.; Zhang, Z.; Li, D.; Lin, K.; Wang, J. MFDNet: Collaborative Poses Perception and Matrix Fisher Distribution for Head Pose Estimation. *IEEE Trans. Multimed.* **2021**, *24*, 2449–2460. [CrossRef]

13. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-Convolutional Neural Network Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 580–587.
14. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1440–1448.
15. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
16. Liu, H.; Wang, X.; Zhang, W.; Zhang, Z.; Li, Y.-F. Infrared head pose estimation with multi-scales feature fusion on the IRHP database for human attention recognition. *Neurocomputing* **2020**, *411*, 510–520. [CrossRef]
17. Liu, T.; Liu, H.; Li, Y.-F.; Chen, Z.; Zhang, Z.; Liu, S. Flexible FTIR Spectral Imaging Enhancement for Industrial Robot Infrared Vision Sensing. *IEEE Trans. Ind. Inform.* **2019**, *16*, 544–554. [CrossRef]
18. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 779–788.
19. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the 2016 European Conference on Computer Vision, LNCS 9905, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.
20. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA; pp. 6517–6525.
21. Ge, P.; Guo, L.; He, D.; Huang, L. Light-weighted vehicle detection network based on improved YOLOv3-tiny. *Int. J. Distrib. Sens. Netw.* **2022**, *18*, 15501329221080665. [CrossRef]
22. Park, M.; Ko, B.C. Two-Step Real-Time Night-Time Fire Detection in an Urban Environment Using Static ELASTIC-YOLOv3 and Temporal Fire-Tube. *Sensors* **2020**, *20*, 2202. [CrossRef] [PubMed]
23. Wang, K.; Liu, M. Toward Structural Learning and Enhanced YOLOv4 Network for Object Detection in Optical Remote Sensing Images. *Adv. Theory Simul.* **2022**, *5*, 2200002. [CrossRef]
24. Lin, B.-H.; Chen, J.-C.; Lien, J.-J.J. Defect Inspection Using Modified YoloV4 on a Stitched Image of a Spinning Tool. *Sensors* **2023**, *23*, 4476. [CrossRef]
25. Mekhalfi, M.-L.; Nicolo, C.; Bazi, Y.; Al Rahhal, M.M.; Alsharif, N.A.; Al Maghayreh, E. Contrasting YOLOv5, Transformer, and EfficientDet Detectors for Crop Circle Detection in Desert. *IEEE Geosci. Remote. Sens. Lett.* **2022**, *19*, 288–302. [CrossRef]
26. Wang, R.; Zhang, Z.-F.; Yang, B.; Xi, H.-Q.; Zhai, Y.-S.; Zhang, R.-L.; Geng, L.-J.; Chen, Z.-Y.; Yang, K. Detection and Classifi-cation of Cotton Foreign Fibers Based on Polarization Imaging and Improved YOLOv5. *Sensors* **2023**, *23*, 4415. [CrossRef]
27. Lin, F.-C.; Ngo, H.-H.; Dow, C.-R.; Lam, K.-H.; Le, H.L. Student Behavior Recognition System for the Classroom Environment Based on Skeleton Pose Estimation and Person Detection. *Sensors* **2021**, *21*, 5314. [CrossRef]
28. Xu, D.; Wu, Y. Improved YOLO-V3 with DenseNet for Multi-Scale Remote Sensing Target Detection. *Sensors* **2020**, *20*, 4276. [CrossRef] [PubMed]
29. Bao, H.-L.; Wan, M.; Liu, Z.-X. Real-Time Semantic Segmentation Network Based on Regional Self-Attention. *Laser Optoelectron. Prog.* **2021**, *58*, 0810018.
30. Chen, Y.H.; Wu, H.B.; Pei, H. Image Super-Resolution Reconstruction Method Based on Self-Attention Deep Network. *Laser Optoelectron. Prog.* **2021**, *58*, 0410013. [CrossRef]
31. 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.
32. Liu, Y.-C.; Shao, Z.-R.; Hoffmann, N. Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions. *arXiv* **2021**, arXiv:2112.05561.
33. Wu, S.; Du, C.; Chen, H.; Jing, N. Coarse-to-Fine UAV Image Geo-Localization Using Multi-stage Lucas-Kanade Networks. In Proceedings of the 2021 2nd Information Communication Technologies Conference (ICTC), Nanjing, China, 7–9 May 2021; Volume 81, pp. 564–577.
34. Gevorgyan, Z. SIoU Loss: More Powerful Learning for Bounding Box Regression. *arXiv* **2022**, arXiv:2205.12740.
35. Liu, H.; Liu, T.; Zhang, Z.; Sangaiah, A.K.; Yang, B.; Li, Y. ARHPE: Asymmetric Relation-Aware Representation Learning for Head Pose Estimation in Industrial Human–Computer Interaction. *IEEE Trans. Ind. Inform.* **2022**, *18*, 7107–7117. [CrossRef]
36. Li, Z.; Liu, H.; Zhang, Z.; Liu, T.; Xiong, N.N. Learning Knowledge Graph Embedding with Heterogeneous Relation Attention Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 3961–3973. [CrossRef]
37. Liu, H.; Zheng, C.; Li, D.; Shen, X.; Lin, K.; Wang, J.; Zhang, Z.; Zhang, Z.; Xiong, N. EDMF: Efficient Deep Matrix Factorization with Review Feature Learning for Industrial Recommender System. *IEEE Trans. Ind. Inf.* **2022**, *18*, 4361–4371. [CrossRef]
38. Gou, J.; Yu, B.; Maybank, S.-J.; Tao, D. Knowledge Distillation: A Survey. *Int. J. Comput. Vis.* **2021**, *129*, 1789–1819. [CrossRef]
39. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 65, pp. 34–48. Available online: https://arxiv.org/abs/1911.08287 (accessed on 1 June 2022).
40. Zhang, Y.-F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [CrossRef]

41. Deng, L.; Li, H.; Liu, H.; Gu, J. A lightweight YOLOv3 algorithm used for safety helmet detection. *Sci. Rep.* **2022**, *12*, 534–556. [CrossRef] [PubMed]
42. Tan, S.; Gonglin, L.; Ziqiang, J.; Li, H. 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), Tokoname, Japan, 4–6 March 2021; Volume 78, pp. 330–333, 837–850.
43. 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, 18–22 June 2023; Volume 25, pp. 157–179.