electronics

MDPI

Article

DAAM-YOLOV5: A Helmet Detection Algorithm Combined with Dynamic Anchor Box and Attention Mechanism

Weipeng Tai [1,2], Zhenzhen Wang [1], Wei Li [1], Jianfei Cheng [1] and Xudong Hong [1,*]

[1] School of Computer Science and Technology, Anhui University of Technology, Maanshan 243000, China; taiweipeng@ahut.edu.cn (W.T.); wzz_personal@163.com (Z.W.); weiLi_ahut@163.com (W.L.); chengjianfei4505@gmail.com (J.C.)
[2] Research Institute of Information Technology, Anhui University of Technology, Maanshan 243000, China
* Correspondence: xdhong@ahut.edu.cn

Abstract: Helmet recognition algorithms based on deep learning aim to enable unmanned full-time detection and record violations such as failure to wear a helmet. However, in actual scenarios, weather and human factors can be complicated, which poses challenges for safety helmet detection. Camera shaking and head occlusion are common issues that can lead to inaccurate results and low availability. To address these practical problems, this paper proposes a novel helmet detection algorithm called DAAM-YOLOv5. The DAAM-YOLOv5 algorithm enriches the diversity of datasets under different weather conditions to improve the mAP of the model in corresponding scenarios by using Mosaic-9 data enhancement. Additionally, this paper introduces a novel dynamic anchor box mechanism, K-DAFS, into this algorithm and enhances the generation speed of the blocked target anchor boxes by using bidirectional feature fusion (BFF). Furthermore, by using an attention mechanism, this paper redistributes the weight of objects in a picture and appropriately reduces the model's sensitivity to the edge information of occluded objects through pooling. This approach improves the model's generalization ability, which aligns with practical application requirements. To evaluate the proposed algorithm, this paper adopts the region of interest (ROI) detection strategy and carries out experiments on specific, real datasets. Compared with traditional deep learning algorithms on the same datasets, our method effectively distinguishes helmet-wearing conditions even when head information is occluded and improves the detection speed of the model. Moreover, compared with the YOLOv5s algorithm, the proposed algorithm increases the mAP and FPS by 4.32% and 9 frames/s, respectively.

Keywords: YOLOv5; attention mechanism; dynamic anchor box; helmet detection; occlusion detection

Citation: Tai, W.; Wang, Z.; Li, W.; Cheng, J.; Hong, X. DAAM-YOLOV5: A Helmet Detection Algorithm Combined with Dynamic Anchor Box and Attention Mechanism. *Electronics* **2023**, *12*, 2094. https://doi.org/10.3390/electronics12092094

Academic Editor: Hüseyin Kusetogullari

Received: 22 March 2023
Revised: 21 April 2023
Accepted: 24 April 2023
Published: 4 May 2023

1. Introduction

Every year, the failure of workers to wear safety helmets results in significant losses for construction sites and society. In terms of monitoring the wearing of safety helmets, it is more common to use manual supervision. However, this method is unable to achieve real-time and all-weather monitoring. Solutions based on computer supervision offer advantages, such as fast response, full-time detection, and low monitoring cost, when compared with traditional monitoring methods. However, existing computer supervision technologies often fall short in terms of detection speed and accuracy, particularly in complex and open scenarios. In actual scenarios, areas close to the water, open fields, and strong winds can shake the camera during object detection, affecting the recognition rate and detection speed of the model. Changes in weather and lighting can also lead to difficulties in target recognition. Furthermore, the size of objects in images can differ significantly at varying distances, making it difficult to use fixed prior bounding boxes, resulting in inaccurate prior information and other issues. Additionally, during rush hours, large crowds and mutual occlusion due to people holding umbrellas on rainy

days can increase missed-recognition rates. This paper aims to address the challenges of target occlusion and video frame shaking by proposing a helmet detection algorithm named Dynamic Anchor Box and Attention Mechanism YOLOv5 (DAAM-YOLOv5) that incorporates a dynamic anchor box and attention mechanisms. Our experimental results demonstrate that the proposed DAAM-YOLOv5 algorithm outperforms YOLOv5s in terms of both mAP (by 4.32%) and FPS (by 9 frames per second), indicating its superiority in real-time and all-weather helmet detection. This proves its effectiveness in improving worker safety at construction sites.

The main contributions of this paper are summarized as follows:

1. The improved Mosaic data enhancement method was used to randomly crop, scale, and arrange of up to nine (instead of four) pictures and make appropriate adjustments to light and angles to generate datasets with various brightness of light, angles of rotation, and target sizes. This method effectively increases the richness of the datasets and enhances the recognition ability of the algorithm under complex weather, light, shaking, and other conditions.
2. A new dynamic anchor box mechanism called Dynamic Anchor Feature Selection with K-means++ (K-DAFS) is proposed. Based on the dynamic anchor box [1], this mechanism adopts the K-means++ [2] clustering algorithm to automatically adjust the size of the training anchor boxes, accelerate the convergence of the target anchor boxes, solve the problem of mismatch between prior information and truth, and achieve faster detecting speed.
3. A new YOLOv5 model with attention mechanism is proposed. The attention mechanism was introduced to enhance the weight of the header information and weaken the background information. The ROI was pooled in the spatial attention module to properly adjust the sensitivity of the model to the object's marginal information to improve the recognition rate of the model for the occlusion target.
4. It improves the GIoU in the original model structure, to CIoU, and solves the problem of slow convergence or non-convergence caused by GIoU_Loss degenerating to IoU when the prediction box and ground truth box are included.

2. Related Works

Currently, there are two main methods for detecting safety helmets: traditional image recognition methods and deep learning-based image recognition methods. The former employs a sliding window approach that performs convolution operations on each receptive field in the image, using multiple convolution kernels of varying sizes to achieve object detection and localization. However, convoluting the entire image can cause the model to focus too much on non-important regions, posing challenges for the accuracy and generalization performance of the detection model. To address this issue, researchers around the world have explored attention-based object detection methods, such as class activation mapping [3] (CAM), stereo attention module [4] (SAM), convolutional block attention module [5] (CBAM), and squeeze-and-excitation [6] (SE).

Computer vision algorithms based on deep learning can be divided into two detection types: one-stage and two-stage object.

The former includes the region-based convolutional neural network (R-CNN) series (including R-CNN [7], fast R-CNN [8], faster R-CNN [9], and mask-RCNN [10]), region-based fully convolutional network [11] (R-FCN), and so on. The latter includes you only look once (YOLO) series (from YOLOv1 to YOLOv5) [12–15], single-shot multiBox detector [16] (SSD) series, such as rainbow SSD [17] (R-SSD), deconvolutional SSD [18] (DSSD) and feature-fusion SSD [19] (FSSD), and RetinaNet [20], etc. At present, YOLOv5 has the better performance among the one-stage object detection algorithms. However, it still has the problem of it being difficult to distinguish objects of different size, which is determined by the one-stage property. Researchers have tried to approach the problem in various ways. Kun Han et al. [21] adopted multiscale detection for detecting helmets to predict a large number of small targets by adding a fourth detection dimension, which

effectively improved the recognition accuracy. Li M et al. [22] used the method of combining cascade parallel multi-scale convolution residual block and dual-channel fusion to solve the problem of different target sizes when analyzing facial visual expressions. Unlike them, others tried to improve the recognition of small targets by combining YOLOv5 with an attention mechanism. For example, Yan J et al. [23] proposed a feature pyramid network based on an attention mechanism that improves the detection performance of small targets by generating target features of different sizes in the feature pyramid network. Tan S et al. [24] not only added a functional detection scale based on YOLOv5 but also replaced non-maximum suppression (NMS) with DIOU-NMS, which made the bounding box of suppression predictions more accurate and improved the accuracy of recognition for small targets. In a real situation, there is also the phenomenon that the detected objects occlude each other. To solve this problem, some researchers, such as Li J et al. [25], have used the head information detection algorithm to further improve detection accuracy by extracting the histogram of oriented gradient features of the human head, and then they classified different labels with a support vector machine (SVM). Finally, they realized helmet detection with color feature recognition. Other researchers, such as Tian Qing et al. [26], employed an attention mechanism to mitigate the influence of object occlusion on detection. By designing and adding a variability convolutional network [27] and an attention mechanism to the feature network, they improved the feature extraction ability of the model for the occlusion phenomenon. By introducing the spatial pyramid pooling module and the squeeze-and-excitation channel attention mechanism [6] before the YOLO input layer, to strengthen the feature fusion of pedestrians of different scales, and then pruning the network, Xiang N et al. [28] finally improved the detection accuracy and reduced the missed detection rate in the case of small pedestrians. Aiming at the problem of strong interference factors, such as different light intensities and weather conditions, Bin Dai et al. [29] used the method of multilayer fusion, taking into account shallow-level semantic information and deep-level semantic information, and improved the safety helmet recognition rate in conditions of insufficient light.

Based on the above research, the structure of YOLOv5 is further improved in this paper to improve recognition accuracy and rate in the case of target occlusion, video shaking, insufficient light, and other problems.

3. DAAM-YOLOv5

3.1. Network Structure

The network structure of DAAM-YOLOv5 is shown in Figure 1. It is also divided into four parts, similar to YOLOv5: input, backbone, neck, and prediction.

Where the "*" is multiplication sign, taking 320*320*32 in Figure 1 as an example, it means a photo with a length of 320 pixels, a width of 320 pixels, and a number of channels of 32.

Compared with YOLOv5, the main changes are as follows:

1. Mosaic-4 is modified to Mosaic-9 data enhancement in the Input;
2. A CBAM [5] attention mechanism is added to the Backbone;
3. A K-DAFS dynamic anchor box is added during the training phase and recognition phase;
4. In the Prediction part, CIoU is used to replace the original GIoU to compensate for the inherent defects of GIoU.

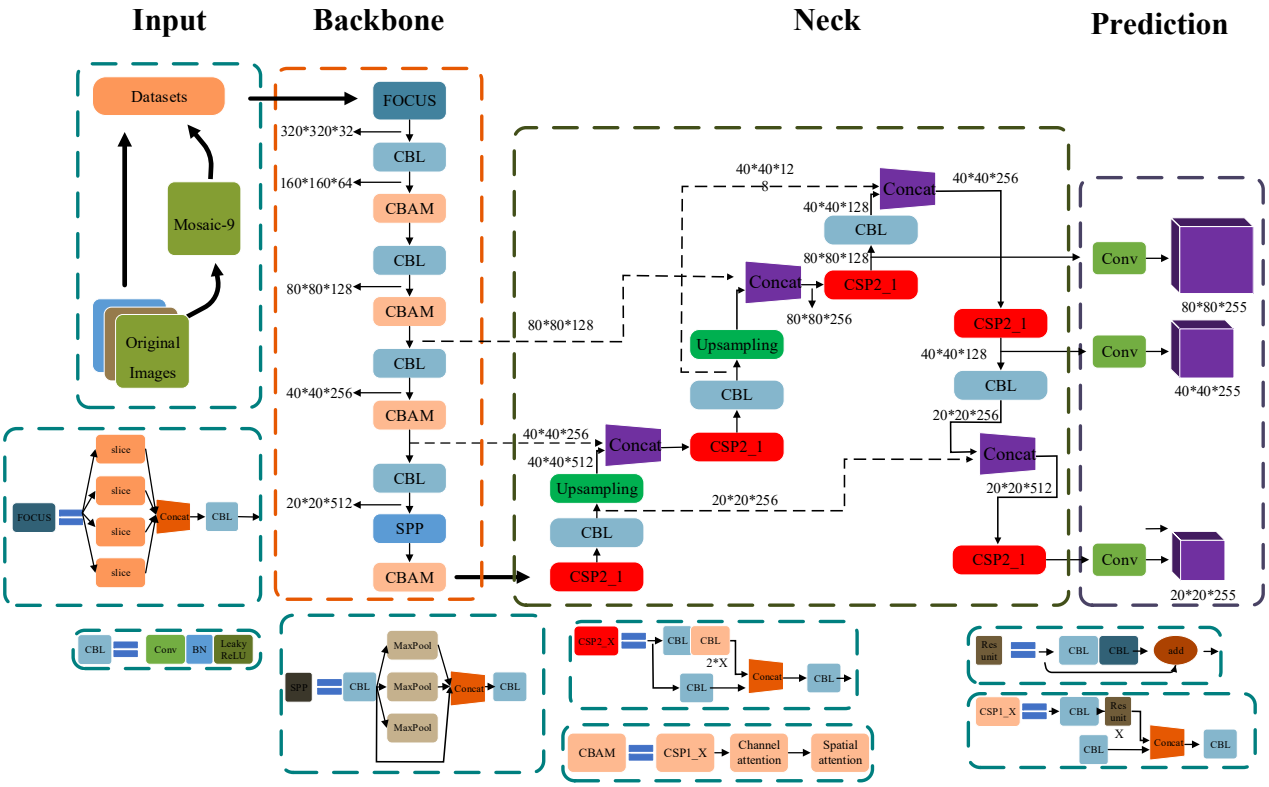

Figure 1. Schematic diagram of the DAAM-YOLOv5 network structure model.

In this algorithm, the main functions of the input are Mosaic-9 data enhancement and the adaptive size of the anchor box. The original images generate new images through Mosaic-9, and then the newly generated images are sent to Backbone together with the original images. The backbone mainly contains the FOCUS structure, CBAM attention mechanism, and Cross Stage Partial_1 (CSP1_X) structure. The main function of the FOCUS structure is slicing the images. It splits the high-resolution feature maps into multiple low-resolution feature maps. The schematic diagram is shown in Figure 2. While reducing the number of parameters and network layers, FOCUS minimizes information loss and improves the convolution speed as a whole.

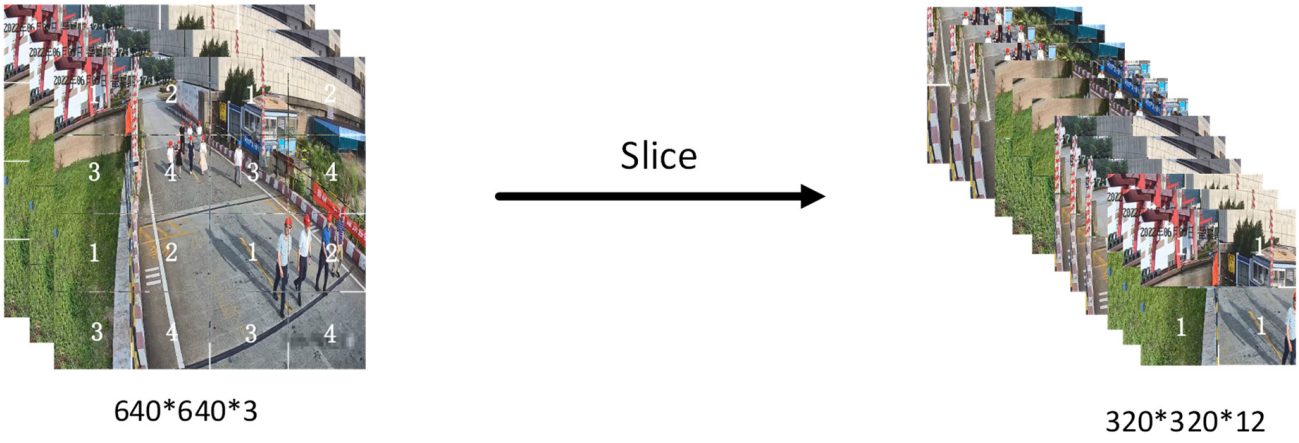

Figure 2. Slicing of Focus.

The CSP1_X structure can increase the gradient value of the backpropagation between layers to avoid the gradient disappearance caused by gradient descent, and so extract more fine-grained features without worrying about network degeneration [30]. The CBAM attention mechanism redistributes the recognition weight of the targets in the image and

appropriately reduces the edge information sensitivity. The main function of prediction is to calculate and adjust the loss of generalized intersection over union (GIoU_Loss) and NMS and present the inference results. CIoU_Loss is used instead of GIoU_Loss to solve the problem that GIoU_Loss degenerates into IoU when the prediction box or the ground truth box is completely contained by the other boxes, which further relieves the matter that the convergence of IoU_Loss is slow. A diagram of the CIoU is shown in Figure 3, and the calculations of CIoU_Loss are defined in Equations (1)–(4).

$$CIoU = IoU - \frac{d}{C^2} - \alpha\nu, \tag{1}$$

$$CIoU_{LOSS} = 1 - CIoU, \tag{2}$$

$$\alpha = \begin{cases} 0, & IoU < 0.5 \\ \frac{\nu}{(1-\mathrm{IoU})+\nu}, & IoU \geq 0.5' \end{cases} \tag{3}$$

$$\nu = \frac{4}{\pi^2}\left(\tan^{-1}\frac{\omega^{gt}}{h^{gt}} - \tan^{-1}\frac{\omega^{pred}}{h^{pred}}\right)^2, \tag{4}$$

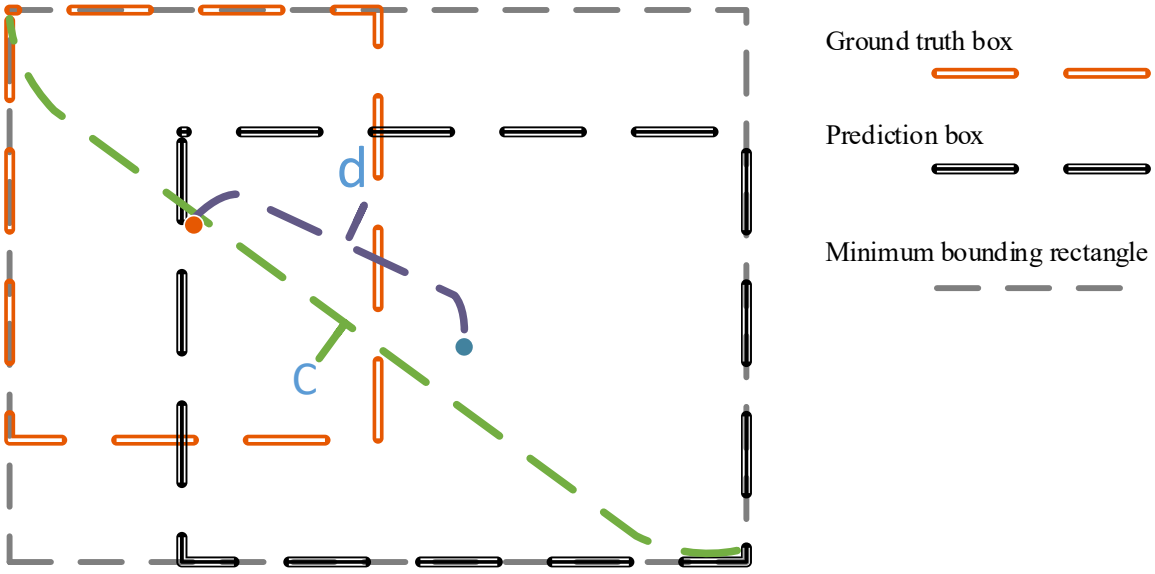

Figure 3. Diagram of CIoU.

In Figure 3, where d is the Euclidean distance between the center point of the prediction box and the ground truth box, c is the diagonal distance of the smallest circumscribed rectangle of the two boxes. In Equations (1)–(4), α is the weight coefficient and ν is the similarity of the aspect ratio between two boxes; the higher the similarity is, the better the prediction effect.

3.2. Improvements over the Original

3.2.1. Mosaic-9 Data Enhancement

As shown in the representation of Mosaic-9 data enhancement workflow in Figure 4, compared to Mosaic-4 in the original YOLOv5, this algorithm has made significant improvements in the random number and adjustment methods of images. First, in the Mosaic-9 process, a maximum of nine pictures are selected for cropping, zooming, arranging, and changing the brightness and then reintegrated and tiled to form a 640*640 picture; second, each of these pictures is rotated by plus or minus 10 degrees to form 18 rotated pictures,

and then these 18 pictures are tiled and resized to 640*640. In the end, a total of 19 new dataset pictures were formed, which greatly enriched the type and quantity of the datasets.

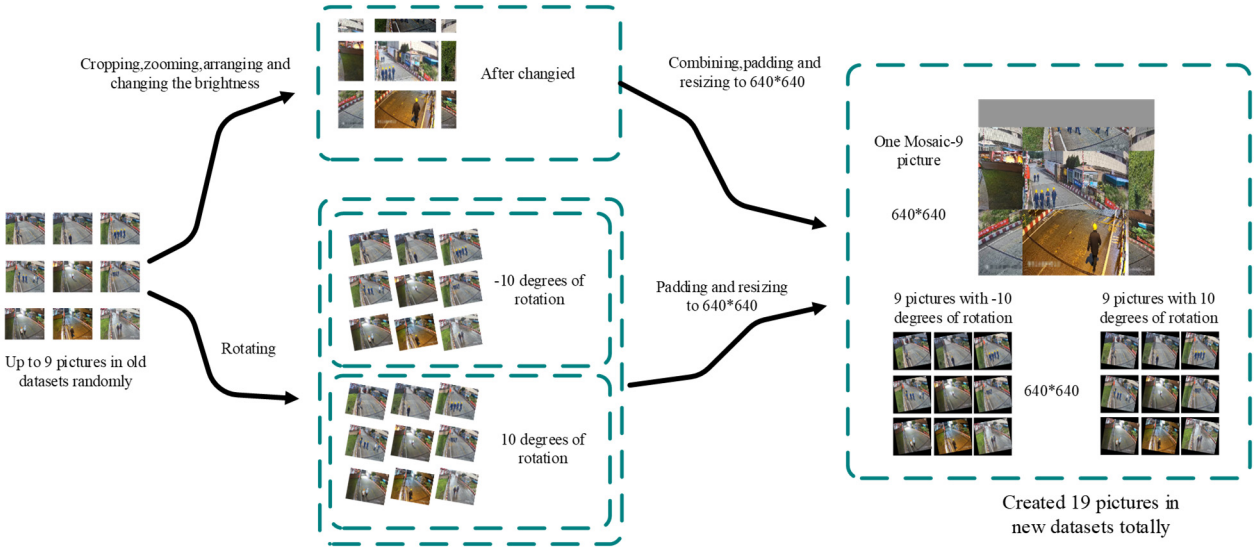

Figure 4. Mosaic-9 data enhancement workflow.

3.2.2. K-DAFS Dynamic Anchor Box

In the one-stage target detection to which YOLO belongs, an anchor box is generally used instead of the target selection stage in two-stage object detection. Generally, there are three methods for selecting an anchor box: generating an anchor box through K-means [31] clustering, artificial regulation, and training. However, these three methods all have certain shortcomings. The K-means method is easily affected by outliers and initial values due to its limitations. Although manually setting anchor boxes is relatively flexible, it requires a lot of effort to find suitable anchor boxes, which increases the difficulty of training [32]. Training to generate anchor boxes is a commonly used method, but traditional training methods are from the top down, cannot be dynamically adjusted, and easily produce a prior information mismatch truth. To address these issues, this paper proposes a novel method for a dynamic anchor box, called K-DAFS, that combines dynamic anchor feature selection (DAFS) [1] with the K-means++ algorithm [2]. The module is shown in Figure 5. This module is constructed based on the anchor refinement module (ARM) [33]. For any nonterminal feature map, the relevant data information about the anchor box comes from both the upper and lower layers simultaneously.

DAAM-YOLOv5 uses manually specified anchor boxes in the initial training, uses the K-means++ clustering algorithm to cluster the marked target anchor boxes multiple times, and generates multiple prior bounding boxes of different sizes. It filters the anchor boxes, which have an IoU of less than 0.5 with the ground truth box, through ARM. However, ARM will result in a mismatch between the receptive field of the point and the anchor boxes for some feature point, which may weaken the detection ability of the model [1]. Therefore, this paper fuses top-down and bottom-up bidirectional paths using bidirectional feature fusion (BFF) combined with K-means++ to connect feature maps of different scales and receive information from the upper and lower layers to dynamically adjust the anchor boxes. Finally, the dynamic refining anchor box is obtained as the prior bounding box for model training.

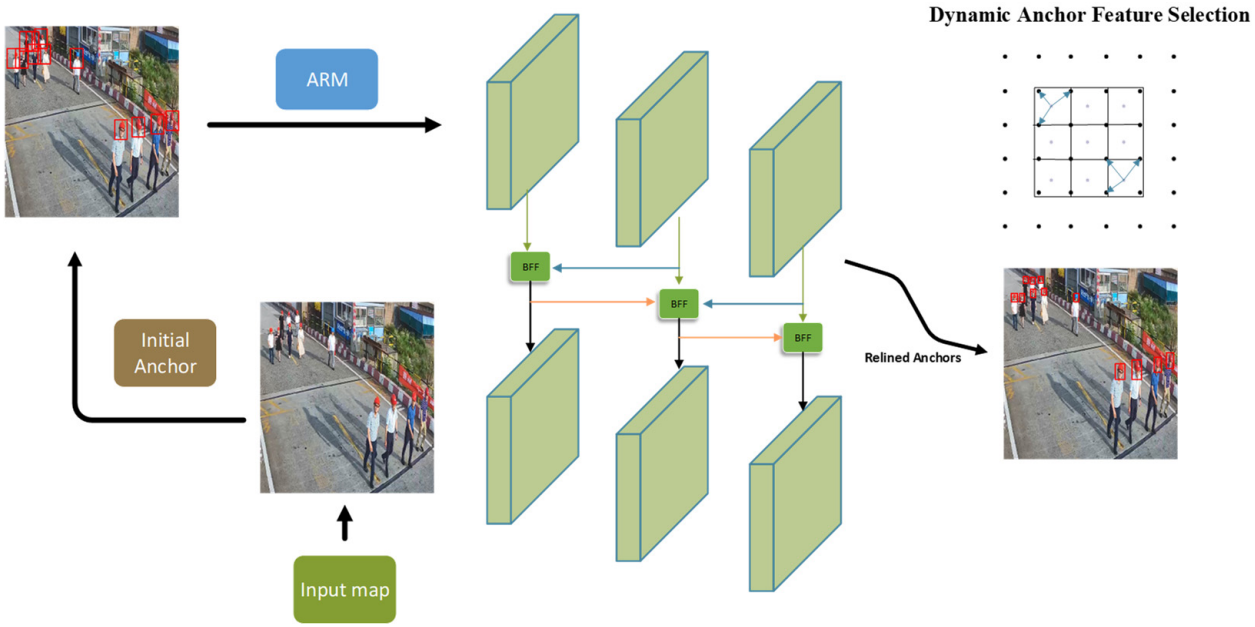

Figure 5. DAFS model structure.

In detail, the workflow of K-means++ in BFF is shown in Figure 6, and is as follows:

1. Among the prediction boxes, the set B is obtained through the feedback from the upper and lower levels of BFF, the center point of a prediction box is randomly selected as the initial center point u_0.
2. The Euclidean distance $d_i(x_i, u_0)$ between u_0 and the four anchor points x_i of the ground truth box is calculated, and the farthest point x_t is obtained as the new center point u_1.

$$d_i(x_i, u_0) = ||x_i - u_0||_2^2, \ \mathrm{i}\epsilon\{1,2,3,4\}, \tag{5}$$

$$d_t(x_t, u_0) = \mathrm{Max}\ d_i(x_i, u_0), \ \mathrm{t}\epsilon\{1,2,3,4\}, \tag{6}$$

3. The minimum value E of the Euclidean distance between u_1 and the anchor point x_i in all prediction boxes in B is calculated, then the anchor box j corresponding to E is the optimal anchor box.

$$d_j = \sum_{i=1}^{4} ||x_i - u_1||_2^2, \ \mathrm{j}\epsilon\mathrm{B}, \tag{7}$$

$$\mathrm{E} = \mathrm{Min}\ d_j, \tag{8}$$

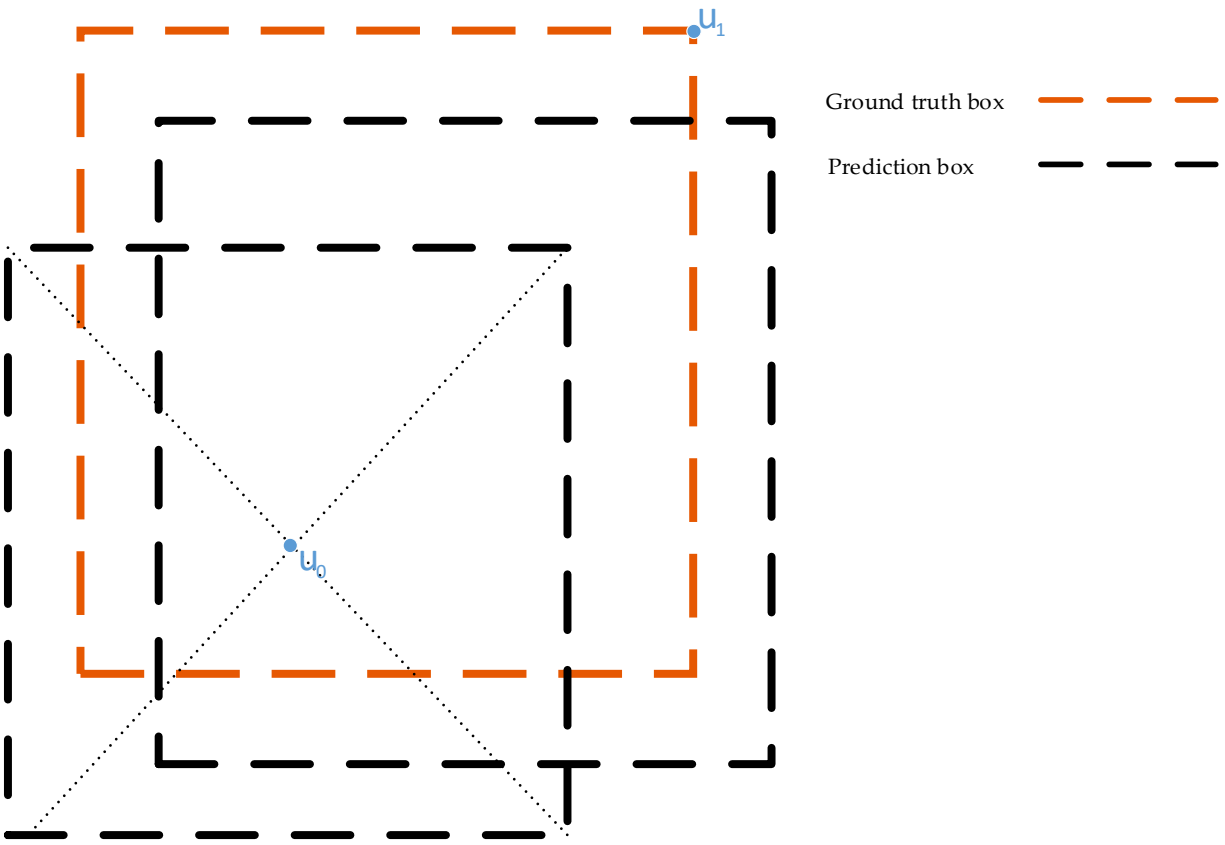

Figure 6. Calculation diagram of K-means++ in BFF.

3.2.3. Attention Mechanism

To improve the recognition accuracy of the algorithm in a specific area, a currently more advanced attention mechanism, CBAM [5], is introduced based on YOLOv5. The main method is to redistribute the weights of information and filter appropriate information to fit the current scene. Compared with the attention mechanism of SENet [6], which only focuses on channel information, CBAM can take into account the reception of spatial information while paying attention to channel information. Without obviously increasing the number of convolution operation layers, it is more effective for solving the problems of small edge areas, blurred edges, and partial information loss. Experiments have proven that [5], in the process of training and learning, channel attention is first used in sequence, and then spatial attention is used to achieve the best results. CBAM has a better performance in lightweight models, but it increases the complexity of the model to a certain extent and may reduce the detection speed by a small amount. The structural model of the CBAM attention mechanism is shown in Figure 7:

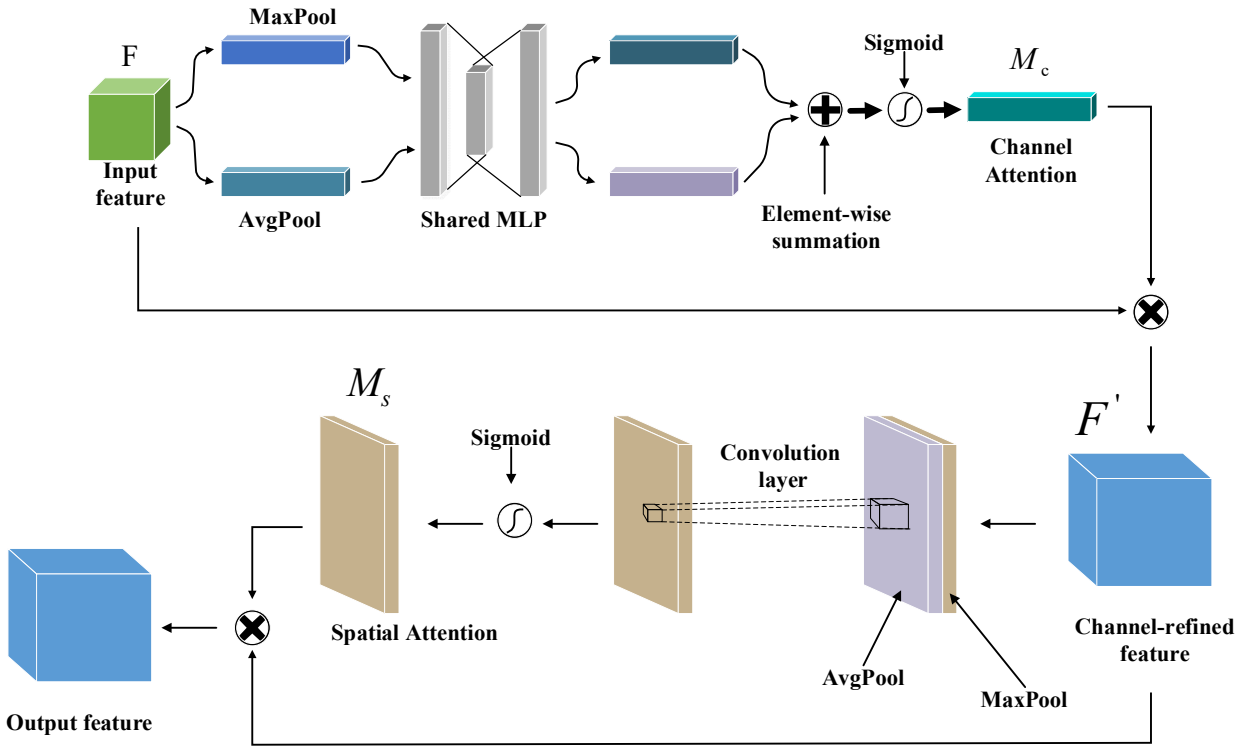

Figure 7. The structural model of the CBAM attention mechanism.

As shown in Figure 7, CBAM mainly includes two modules, namely, the channel attention module and the spatial attention module. The two focus on different objects. The former mainly focuses on the category of the current convolution object, and the latter focuses on the position of the current object. The former half of CBAM is the channel attention module, which first performs maximum pooling and average pooling on the input feature map and then roughly perceives the object category using the shared multilayer perceptron. Finally, sum up the data and then feed it into the activation function for processing after elementwise operation. The calculation process can be briefly described as Equations (9) and (10):

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))), \tag{9}$$

$$= \sigma\left(W_1\left(W_0\left(F^C_{avg}\right)\right) + W_1\left(W_0\left(F^C_{max}\right)\right)\right), \tag{10}$$

where σ is the sigmoid function, C is the number of channels, and W_0 and W_1 are the weights of the first layer and the second layer of the perceptron, respectively.

In CBAM, the position information of the target is supplemented by introducing the spatial attention module. The pooling operation along the channel axis can effectively highlight the information area [34]. First, the output feature map in the attention channel is weighted with the original input feature map. The obtained image is pooled with the maximum and the average. Second, a 7*7 convolution kernel is used to convolve the two feature layers at the same time, and the two feature layers are aggregated into one feature layer through pooling. Finally, the feature after the layer is processed by the sigmoid function, M_s, that contains the emphasized or suppressed position information that has been obtained. The calculation process for spatial attention can be briefly described as Equations (11) and (12):

$$M_s(F) = \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)]\right), \tag{11}$$

$$= \sigma\left(f^{7\times7}\left(\left[F^C_{Avg}; F^C_{Max}\right]\right)\right), \tag{12}$$

where σ is a sigmoid function, $f^{7\times7}$ is a convolution operation with a size of 7×7, and F^C_{Avg} and F^C_{Max} are the feature layers after average pooling and maximum pooling, respectively.

4. Experimental Results and Analysis

4.1. Experimental Datasets

The dataset of 32,000 images used in this paper was collected from specific scenes or generated by Mosaic-9 according to real images and includes 12,000 actually collected images and 20,000 automatically generated images. Considering the influence of the natural environment on the recognition rate, the images include sunny, cloudy, rainy, day, night, and other conditions. There were 123,688 positive samples (helmet), 75,289 negative samples (head), 256,832 human contour samples (people), and 57,855 unlabeled samples. The unlabeled samples had insufficient clarity, serious occlusion, or poor contour information, so they were not used for training. Finally, the mixed datasets were divided into training datasets, validation datasets and testing datasets, at a ratio of 7:2:1, for the experiments. The partial images generated by Mosaic-9 are shown in Figure 8, and some low-brightness images have been automatically generated to compensate for the lack of nighttime training datasets.

Figure 8. Some of the images generated by Mosaic 9.

4.2. Experimental Environments

Both the training and testing of this experiment were carried out using the Windows 10 operating system. The specific software and hardware environments are listed in Table 1.

Table 1. Training Environmental Configuration.

Configuration Name	Version Parameters
Operating system	Windows10
Graphics (GPU)	Tesla P4*4 (8G*4)
Processor (CPU)	16 core Intel(R) Xeon(R) Gold 5120 CPU @ 2.2GHz
Framework	Pytorch (1.10.2)
GPU acceleration environment	CUDA10.2

4.3. Experimental Hyperparation and Evaluation Criteria

To ensure the effectiveness of the experimental data, the same hyperparameters were used for training and testing different algorithm models. Stochastic gradient descent [35] backpropagation was used to fine-tune and optimize the network parameters, and the parameter settings are shown in Table 2. The initial learning rate was set to 0.01, weight decay was set to 0.00002, batch size was set to 32, IoU was set to 0.35, the momentum factor was set to 0.937, to avoid the model training falling into local optimum, and the number of epochs was set to 300.

Table 2. Environmental Hyperparameter Configuration.

Parameter	Value
Learning rate	0.01
Weight decay	0.00002
Batch size	32
IoU	0.35
Momentum factor	0.937
Epoch	300

Model evaluation is an important task in deep learning. The evaluation of indicators usually includes accuracy, precision, recall, mAP, parameters, and giga-floating-point operations per second (GFLOPs).

These are described by the following equations:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}}, \tag{13}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{14}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{TN}}, \tag{15}$$

where TP indicates the number of samples in which positive class samples are correctly predicted as positive, TN indicates the number of samples in which negative class samples are correctly predicted as negative, FP indicates the number of negative samples incorrectly predicted as positive samples and FN represents the number of positive samples that are wrongly predicted as negative samples.

The more widely used criterion in model evaluation is mAP, and the mAP calculation is as Equation (16):

$$mAP = \frac{1}{K}\sum_{i=1}^{N} AP_i;\ AP = \sum_{i=1}^{n-1} (r_{i+1} - r_i) P_{inter}(r_i + 1), \tag{16}$$

where r is the recall value corresponding to each interpolation segment of Precision sorted in ascending order.

4.4. Training Results

Figure 9 shows the comparison of mAP under the same configurations of the diverse algorithm models involved in this experiment for approximately 300 epochs. The results show that the mAP of each algorithm rapidly rises in the top 20 epochs, among which DAAM-YOLOv5 and YOLOv3 are the fastest rising; this paper's proposed algorithm reached a mAP of 0.91 and YOLOv3s reached 0.83. The proposed algorithm converges stably at epoch = 52, which is later than the other algorithms. All algorithms tend to be stable at approximately epoch = 150, and the subsequent training makes the mAP slightly decrease, showing signs of overfitting. There is no significant difference between the CBAM and SE attention mechanisms in the first 40 epochs, and the mAP of CBAM is slightly higher than that of the SE attention mechanism after 40 epochs. Finally, the proposed algorithm's optimal mAP is 0.9636, which is higher than the 0.9204 of the YOLOv5 model and increased by approximately 4.32%, verifying the feasibility of this improved algorithm.

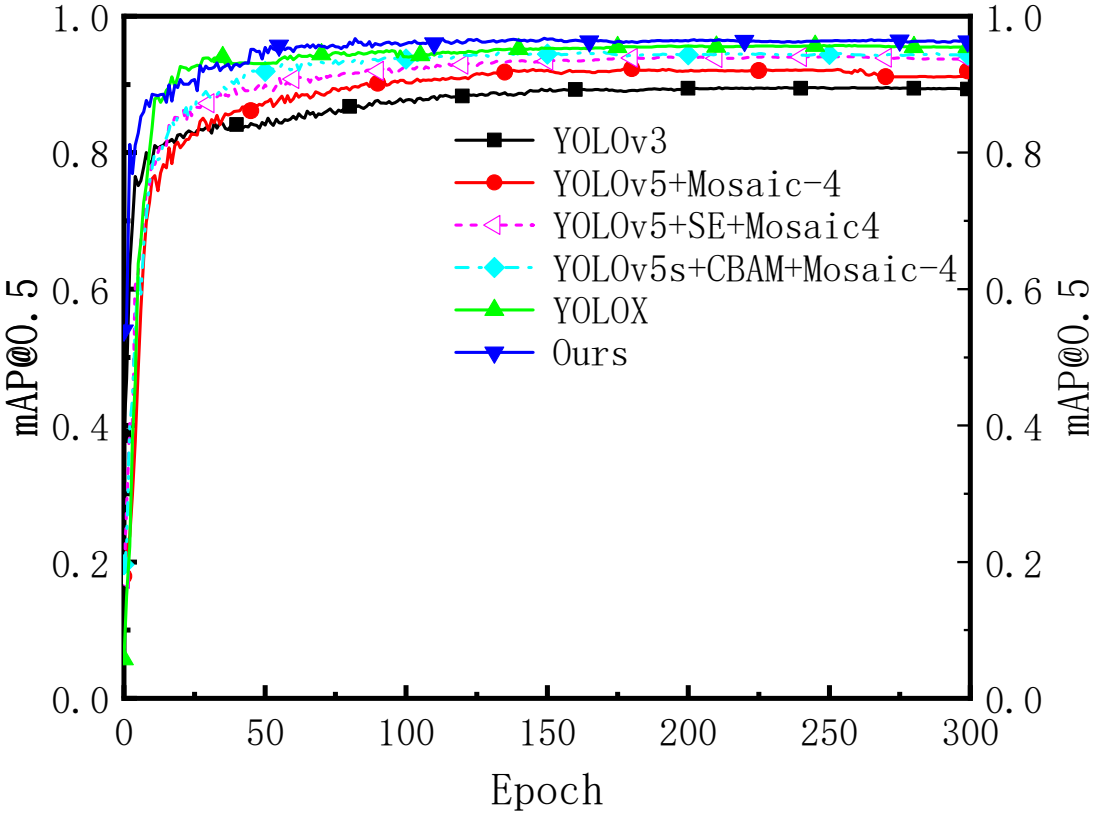

Figure 9. The mAP of related algorithms in training.

The key data points for these algorithms are shown in Table 3.

Table 3. The mAP comparison of different algorithms in different epochs.

Algorithms	mAP@0.5						
	Epoch = 10	**Epoch = 20**	**Epoch = 50**	**Epoch = 150**	**Epoch = 234**	**Epoch = 248**	**Epoch = 299**
YOLOv3	0.78805	0.82702	0.84067	0.89059	0.89519	0.89512	0.89322
YOLOv5	0.76171	0.80696	0.87373	0.92099	0.92049	0.92041	0.91988
YOLOX	0.83723	0.92661	0.93067	0.95289	0.95618	0.95675	0.95464
Ours	0.88596	0.90738	0.95428	0.96609	0.96234	0.96363	0.96344

4.5. Detection Results of DAAM-YOLOv5 in Various Cases

Figure 10 shows the partial detection results of DAAM-YOLOv5 in multiple cases in the testing datasets. Among them, (a)–(c) show the situations in which a tester rides an electric bicycle through the detection area at a speed of 40 km/h. The algorithm can still correctly identify the targets in the case of high speed. Figure 10d shows the detection of ROI, the people are detected within the ROI, but not outside it. It can be seen in (e) that a safety helmet will not be detected in the case of no human wearing a safety helmet, because the identification of the safety helmet is not set when no one is wearing it, and the algorithm does not appear to be misidentified. In (f), the violator intentionally obscured the head information in an attempt to interfere with the detection results, but the algorithm still detected it correctly, indicating that the generalization of this algorithm is in good shape. For irregular safety helmets that do not meet the requirements of production, worn by the personnel in (g,h), the algorithm correctly detects the violation information. In (i), there is the case of insufficient light at night, and the detection effect is good.

Figure 10. Detection results of DAAM-YOLOv5 in multiple cases.

4.6. Comparison of Detection Effects

4.6.1. Comparison between the YOLOv5 Algorithm and DAAM-YOLOv5 Algorithm

Figure 11 compares the detection effects of YOLOv5 and DAAM-YOLOv5 under the same configuration. It is obvious that the improved DAAM-YOLOv5 has good performance in the case of occluded objects after adding the attention mechanism. It effectively reduces the false recognition and misidentification rates of the targets.

(**a**) YOLOv5 (**b**) DAAM-YOLOv5

Figure 11. Comparison between YOLOv5 algorithm and DAAM-YOLOv5 algorithm.

4.6.2. Ablation Experiment Comparison

Figure 12 shows a detailed mAP of the algorithms in Figure 9 with epochs between 200 and 290. An ablation experiment was also performed to verify the impact of each module on the model performance. The results are shown in Table 4. The first row in the table is the YOLOv5 algorithm without improvement. In the original algorithm, Mosaic-4 data enhancement is used by default. Mosaic-4 data enhancement is used by default in the absence of Mosaic-9 data enhancement. The second and third rows show the comparison of the YOLOv5+CBAM and YOLOv5+SE attention mechanisms, respectively. The fourth row is the DAAM-YOLOv5 algorithm in this paper.

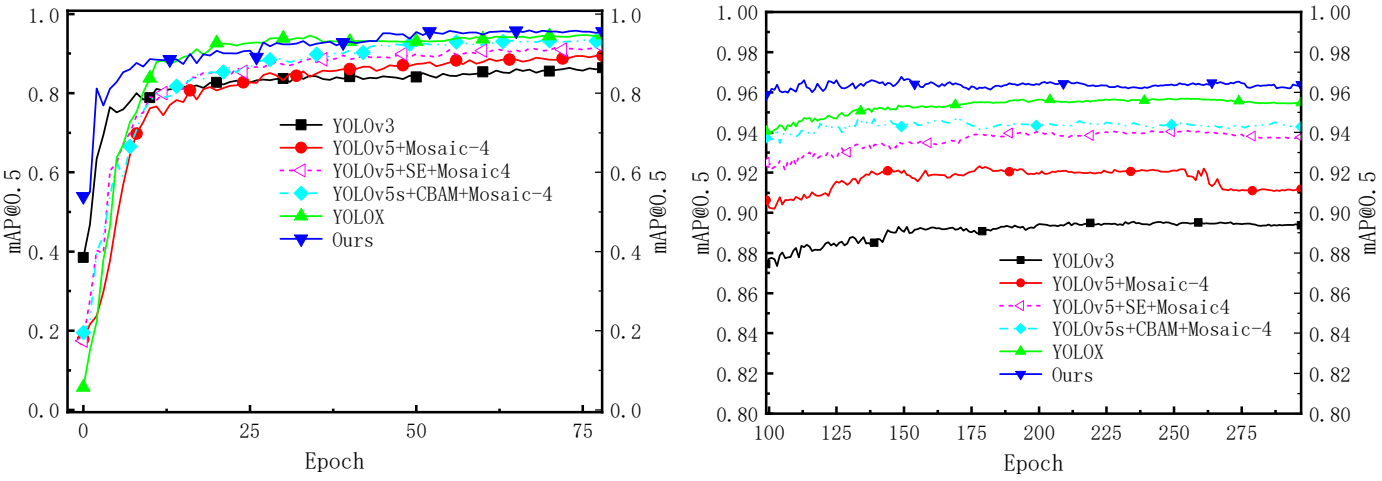

Figure 12. The mAP details of related algorithms with epochs from 0 to 75 and 100 to 299.

Table 4. Ablation experiment.

Algorithms	Accuracy	Precision	Recall	Parameters (M)	mAP@0.5	GFLOPs	Recognition Speed (f/s)
YOLOv5	0.8549	0.8465	0.7536	7.2	0.9204 (+0)	16.1 (+0)	135 (+0)
YOLOv5+CBAM	0.8841	0.8631	0.7255	8.0	0.9440 (+0.0236)	14.9 (−1.2)	118 (−17)
YOLOv5+SE	0.8658	0.8699	0.7320	8.1	0.9399 (+0.0195)	15.5 (−0.6)	127 (−8)
YOLOv5+Mosaic-9	0.8984	0.8223	0.8428	7.5	0.9366 (+0.0162)	16.4 (+0.3)	124 (−11)
YOLOv5+K-DAFS	0.8837	0.8037	0.8633	8.5	0.9108 (−0.0096)	16.8 (+0.7)	158 (+23)
Ours	0.8708	0.8471	0.8221	10.3	0.9636 (+0.0432)	16.5 (+0.4)	144 (+9)

The results showed the following:

1. After the addition of CBAM alone, the mAP of the model has been improved by approximately 2.36%. However, compared with the YOLOv5 algorithm, the detection

speed and recall have decreased to a certain extent, with a reduction of 17 frames/s and 2.81%, respectively. This indicates that the CBAM module can increase the mAP of the model, but it will slightly reduce the model detection speed and increase the missed-recognition rate.

2. Compared with the second and fourth lines, it can be seen that the precision of the SE attention mechanism is slightly lower than that of the CBAM attention mechanism, by 0.41%, but the detection speed is faster than that of CBAM attention mechanism by 9 frames/s.
3. Compared with the first and fourth lines, it can be seen that, after improving the Mosaic-9 data enhancement, the algorithm's mAP has increased by 1.62%, while the speed rate has decreased by 11 frames/s. In addition, this module has shown significant improvements in both accuracy and recall, with gains of 4.35% and 8.68%, respectively.
4. According to the fifth line, it can be seen that there is a slight decrease in mAP, by approximately 0.96%, after adding the K-DAFS dynamic anchor box. However, the obvious improvements in the detection speed and recall of the algorithm are 23 frames/s and 10.97%, respectively. This indicates that the K-DAFS module can significantly increase the detection speed of the model while reducing the model's missed-recognition rate.
5. After integrating the CBAM attention mechanism, the Mosaic-9 data augmentation, and the K-DAFS, the model's detection speed was only slower than the K-DAFS model but it had the highest mAP. Compared with the original YOLOv5, the mAP, detection speed, and recall were improved by 4.32%, 9 frames/s, and 6.85%, respectively.

5. Conclusions

The impact of safety production on businesses and society is significant. However, existing safety helmet detection algorithms have difficulty meeting the needs of complex open scenarios. This paper proposes a safety helmet detection algorithm, DAAM-YOLOv5, which combines a dynamic anchor box with an attention mechanism and demonstrates superior performance in addressing significant target variations, target occlusions, and video frame jitter. The algorithm improves the model's accuracy when dealing with significant variations in targets by adjusting the Mosaic data enhancement in the input layer of YOLOv5. In addition, a new dynamic anchor box is added, and BFF is used to dynamically adjust anchor box sizes to improve the model's detection speed. By incorporating an attention mechanism to adjust the weighting of the background and target recognition, as well as the sensitivity of the algorithm to edge information, the issue of low recognition accuracy in occluded scenarios is resolved. Finally, the algorithm improves the original model structure by replacing GIoU with CIoU, addressing the problem of slow or failed convergence when the prediction box contains the ground truth box.

The results of experimental analysis show that the DAAM-YOLOv5 algorithm has improved mAP by 4.32% and detection speed by 9 FPS. In comparison to other object detection algorithms, we achieved higher accuracy and speed. Therefore, the improved algorithm is better suited for scenarios such as ports, smart construction sites, and danger warning zone. However, the algorithm's recognition performance on flooded road surfaces is not significant at present. In the future, we will increase our research in this field to improve the recognition rate in this scenario.

Author Contributions: Conceptualization, W.T.; Methodology, W.T. and Z.W.; Software, X.H.; Validation, Z.W.; Formal analysis, X.H.; Investigation, J.C. and X.H.; Resources, Z.W. and J.C.; Data curation, W.L.; Writing—original draft, Z.W.; Visualization, W.T. and W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This Paper was funded by Anhui Provincial Key Research and Development Project (Grant No. 202004a07020028); and Natural Science Foundation of Anhui Province, China (Grant No. 2008085QF305).

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

YOLO	You Only Look Once
BFF	Bidirectional Feature Fusion
ROI	Region of Interest
mAP	Mean Average Precision
DAFS	Dynamic Anchor Feature Selection
IoU	Intersection of Union
CIoU	Complete Intersection over Union
GIoU	Generalized Intersection over Union
SAM	Stereo Attention Module
CAM	Class Activation Mapping
CBAM	Convolutional Block Attention Module
SE	Squeeze-and-Excitation
R-CNN	Region based Convolutional Neural Network
R-FCN	Region-based Fully Convolutional Network
SSD	Single-shot MultiBox Detector
R-SSD	Rainbow Single-shot MultiBox Detector
DSSD	Deconvolutional Single-shot MultiBox Detector
FSSD	Feature Fusion Single-shot MultiBox Detector
NMS	Non-maximum Suppression
CSP1_X	Cross Stage Partial_1
ARM	Anchor Refinement Module
SGD	Stochastic Gradient Descent
GFLOPs	Giga-floating-point Operations per Second

References

1. Li, S.; Yang, L.; Huang, J.; Hua, X.S.; Zhang, L. Dynamic anchor feature selection for single-shot object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6609–6618.
2. Arthur, D.; Vassilvitskii, S. *k-Means++: The Advantages of Careful Seeding*; Stanford University: Stanford, CA, USA, 2006.
3. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
4. Ying, X.; Wang, Y.; Wang, L.; Sheng, W.; An, W.; Guo, Y. A stereo attention module for stereo image super-resolution. *IEEE Signal Process. Lett.* **2020**, *27*, 496–500. [CrossRef]
5. 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.
6. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
7. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
8. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, 28. [CrossRef] [PubMed]
10. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
11. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 379–387.
12. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
13. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.

14. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
15. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
16. 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 Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14. Springer International Publishing: Amsterdam, The Netherlands, 2016; pp. 21–37.
17. Jeong, J.; Park, H.; Kwak, N. Enhancement of SSD by concatenating feature maps for object detection. *arXiv* **2017**, arXiv:1705.09587.
18. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
19. Li, Z.; Zhou, F. FSSD: Feature fusion single shot multibox detector. *arXiv* **2017**, arXiv:1712.00960.
20. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
21. Han, K.; Zeng, X. Deep Learning-Based Workers Safety Helmet Wearing Detection on Construction Sites Using Multi-Scale Features. *IEEE Access* **2022**, *10*, 718–729. [CrossRef]
22. Li, M.; Zhang, W.; Hu, B.; Kang, J.; Wang, Y.; Lu, S. Automatic Assessment of Depression and Anxiety through Encoding Pupil-wave from HCI in VR Scenes. *ACM Trans. Multimid. Comput. Commun. Appl.* **2022**. [CrossRef]
23. Yan, J.; Zhao, L.; Diao, W.; Wang, H.; Sun, X. AF-EMS Detector: Improve the Multi-Scale Detection Performance of the Anchor-Free Detector. *Remote Sens.* **2021**, *13*, 160. [CrossRef]
24. 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.
25. Li, J.; Liu, H.; Wang, T.; Jiang, M.; Wang, S.; Li, K.; Zhao, X. Safety helmet wearing detection based on image processing and machine learning. In Proceedings of the 2017 Ninth International Conference on Advanced Computational Intelligence (ICACI), Doha, Qatar, 4–6 February 2017; pp. 201–205.
26. Qing, T.; Yuan, H.; Fei, D.; Yao, N. Research on Head and Shoulders Detection Algorithm in Complex Scene Based on YOLOv5. In Proceedings of the 2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), Dalian, China, 14–16 April 2022; pp. 368–371.
27. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
28. Nan, X.; Lu, W.; Chongliu, J.; Yuemou, J.; Xiaoxia, M. Simulation of Occluded Pedestrian Detection Based on Improved YOLO [J/OL]. *J. Syst. Simul.* **2023**, *35*, 286. [CrossRef]
29. Dai, B.; Nie, Y.; Cui, W.; Liu, R.; Zheng, Z. Real-time Safety Helmet Detection System based on Improved SSD. In Proceedings of the 2nd International Conference on Artificial Intelligence and Advanced Manufacture, Manchester, UK, 15–17 October 2020; pp. 95–99.
30. Orhan, A.E.; Pitkow, X. Skip connections eliminate singularities. *arXiv* **2017**, arXiv:1701.09175.
31. Tibshirani, R.; Walther, G.; Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2001**, *63*, 411–423. [CrossRef]
32. Zhou, X.; Lan, X.; Zhang, H.; Tian, Z.; Zhang, Y.; Zheng, N. Fully convolutional grasp detection network with oriented anchor box. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 7223–7230.
33. Sloss, A.; Symes, D.; Wright, C. *ARM System Developer's Guide: Designing and Optimizing System Software*; Elsevier: Amsterdam, The Netherlands, 2004.
34. Zagoruyko, S.; Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv* **2016**, arXiv:1612.03928.
35. Song, S.; Chaudhuri, K.; Sarwate, A.D. Stochastic gradient descent with differentially private updates. In Proceedings of the 2013 IEEE Global Conference on Signal and Information Processing, Austin, TX, USA, 3–5 December 2013; pp. 245–248.