Computers, Materials & Continua
DOI:10.32604/cmc.2021.017480
Article

Tech Science Press

Algorithm of Helmet Wearing Detection Based on AT-YOLO Deep Mode

Qingyang Zhou[1], Jiaohua Qin[1,*], Xuyu Xiang[1], Yun Tan[1] and Neal N. Xiong[2]

[1]College of Computer Science and Information Technology, Central South University of Forestry & Technology, Changsha, 410004, China
[2]Department of Mathematics and Computer Science, Northeastern State University, Tahlequah, 74464, OK, USA
[*]Corresponding Author: Jiaohua Qin. Email: qinjiaohua@csuft.edu.cn
Received: 31 January 2021; Accepted: 16 March 2021

Abstract: The existing safety helmet detection methods are mainly based on one-stage object detection algorithms with high detection speed to reach the real-time detection requirements, but they can't accurately detect small objects and objects with obstructions. Therefore, we propose a helmet detection algorithm based on the attention mechanism (AT-YOLO). First of all, a channel attention module is added to the YOLOv3 backbone network, which can adaptively calibrate the channel features of the direction to improve the feature utilization, and a spatial attention module is added to the neck of the YOLOv3 network to capture the correlation between any positions in the feature map so that to increase the receptive field of the network. Secondly, we use DIoU (Distance Intersection over Union) bounding box regression loss function, it not only improving the measurement of bounding box regression loss but also increases the normalized distance loss between the prediction boxes and the target boxes, which makes the network more accurate in detecting small objects and faster in convergence. Finally, we explore the training strategy of the network model, which improves network performance without increasing the inference cost. Experiments show that the mAP of the proposed method reaches 96.5%, and the detection speed can reach 27 fps. Compared with other existing methods, it has better performance in detection accuracy and speed.

Keywords: Safety helmet detection; attention mechanism; convolutional neural network; training strategies

1 Introduction

In recent years, intelligent surveillance has played an increasingly important role in our daily life. As a hotspot of computer vision, object detection provides many ideas for intelligent surveillance. Object detection methods mainly use traditional features in early research. Zhu et al. [1] used Histograms of Oriented Gradients (HoG) to extract image features, combined with the cascade-of-rejectors to accelerate the calculation speed and realize pedestrian detection. Zuo et al. [2] used Haar-wavelets transform to model local texture attributes, successfully compensated for the extra cost of two-dimensional texture features, and realized face detection. However, traditional object detection algorithms have many drawbacks, the feature extraction algorithms

have poor generalization ability and low robustness for complex scene images, and the region selection methods using sliding window lacks specific calculation, so its time complexity is too high to solve practical problems.

The development of massively parallel computing technology provides a technical guarantee for deep learning, and deep learning also provides more effective solutions for information hiding [3–7], image classification [8–10], image retrieval [11,12], object detection [13], image inpainting [14], and many other fields. So far, the most state-of-the-art detection algorithms are based on deep learning. R-CNN [15] firstly uses the deep learning model to extract image features and generate region proposals with a sliding window, but it has many repeated calculations, and the amount of calculation is too high. Fast R-CNN [16] integrates the classification and regression of bounding boxes into a network to reduce repeated calculations while using the SPP module to generate fixed-size output. Faster R-CNN [17] inputs the features extracted from images into RPN (Region Proposal Networks). The RPN can input feature maps of any size, output the coordinate information and confidence of the object candidate boxes, and then classify the object candidate boxes. Because region selection and classification should be performed step by step, Faster R-CNN belongs to the two-stage object detection model. With the continuous development of deep learning, two-stage detection methods such as Faster R-CNN are affected by factors such as the complexity of the basic network, the number of candidate boxes, the complexity of classification, and regression sub-networks, the amount of calculation continue to increase. YOLO [18] discards the candidate boxes extraction step in the algorithm and directly implements feature extraction, candidate boxes classification, and regression in an end-to-end deep convolutional network. The detection algorithm similar to YOLO is a one-stage detection algorithm, which further simplifies the implementation steps of object detection, makes the network structure simpler, and the detection speed is faster than that of the two-stage network.

As an essential practical application in object detection, safety helmet wearing detection is closely related to our production and life. Many scholars have done a lot of research about it. Mneymneh et al. [19] extracted the features of the worker and the helmet in the image and set up a cascade according to the feature points to judge whether the worker is wearing a helmet. Li et al. [20] used head positioning, color space transformation, and color feature recognition to achieve the detection of wearing helmets based on the detection results of pedestrians. Wu et al. [21] used the improved YOLO-Densebackbone for helmet detection to improve feature resolution. Long et al. [22] used SSD (Single Shot multi-box Detector) [23] to detect wearing helmets. However, the current detection methods for wearing helmets have disadvantages, such as low detection accuracy for small objects and poor generalization ability for multi-scene detection. Chen et al. [24] introduced the K-means++ clustering algorithm to cluster the size of the helmet in the image and then used the improved Faster-RCNN algorithm to detect the helmet wearing.

However, among the current helmet-wearing detection algorithms, the detector based on one-stage has a faster detection speed, but its detection accuracy for small targets and dense targets is low, and the generalization ability of multiple scenes is poor. The two-stage detector has the disadvantages of large calculation amount and slow detection speed, which is difficult to meet the real-time requirements of helmet detection.

In order to solve the above problems, we propose an AT-YOLO network model based on the attention mechanism for helmet wearing detection. In this paper, we model the correlation between the channel and the spatial dimension in the feature to enhance the ability of feature representation. At the same time, this paper optimizes the dataset, loss function, and training

strategy to improve the network detection performance in all directions while maintaining a high detection speed. The main contributions of this paper include:

(1) Construct a helmet dataset with more balanced categories and richer scenarios. Part of the data comes from Safety-Helmet-Wearing-Dataset, and on this basis, site images are collected through web crawler, video capture, and other ways to expand the dataset so as to make the dataset scene richer and categories more balanced.
(2) Propose the AT-YOLO object detection algorithm for helmet wearing detection. The mutual dependence of features is simulated in the dimensions of space and channel, respectively, so that the network can obtain better detection results on small objects and occluded images.
(3) The DIoU bounding box regression loss function is used. Combining the IoU between the prediction box and the ground truth box and the center point distance as the bounding box regression loss function, the loss function measurement is improved. While improving the accuracy of the network's detection of small objects, it also accelerates the speed of the network convergence.
(4) Different training strategies are used to improve network performance in the network training stage. This paper uses several different training strategies to enhance the network's performance without increasing the cost of network inference and providing a valuable reference for other image research.

The rest of this paper is organized as follows. Section 2 reviews the related research on attention mechanisms and target detection algorithm. Section 3 introduces the method proposed in this paper. Section 4 introduces the evaluation experiment of the method in this paper. Section 5 summarizes the work of this paper.

2 Related Work

2.1 You Only Look Once

YOLO integrates the candidate region extraction and classification and regression tasks of object detection into an end-to-end deep convolutional network. That is, the input image is inferred once, and the positions and categories of all objects and the corresponding confidence probabilities can be obtained. The backbone network of the YOLO network is similar to GoogLeNet [25]. The inception structure is removed to make the backbone network simpler. Input the image into the YOLO model to obtain a feature map with a size of 7×7, which divides divide the image into 7×7 regions. Each area has the confidence of the target, the position of the bounding boxes, and the category information. The YOLO network is simple, the detection speed is fast, and the background false detection rate is low, but the detection accuracy is not as good as the R-CNN detection method, and YOLO is not accurate enough in object positioning.

YOLOv2 [26] uses a new backbone network called Darknet-19, which is the same as the VGG16 [27] model design principle. The network mainly adopts the 3×3 convolution and the 2×2 maximum pooling layer. After passing through the maximum pooling layer, the height and width of the feature map are halved, and the number of channels of the feature map is doubled. YOLOv2 still has the advantage of fast speed. However, its backbone network is not deep enough, it is difficult to recognize more abstract image semantic features, and the bounding box predicted by each grid cell is too less, which is not effective in predicting targets with large-scale changes.

YOLOv3 [28] draws on the idea of Resnet [29], introduces the residual structure, and establishes a deeper Darknet-53. And compared to YOLOv2, the downsampling method of the

pooling layer is canceled, but the feature map is downsampled by adjusting the step size of the convolutional layer to obtain more fine-grained features. YOLOv3 uses multiple-scale fusion methods to make predictions. Similar to FPN, YOLOv3 integrates feature maps of three scales and simultaneously detects multiple-scale feature maps. The small size of the feature map is used to detect large-size objects, and the large size of the feature map is used to detect small-size objects. Compared with YOLOv2, YOLOv3 uses multiple scales to predict at the same time to make the bounding box more, cover a richer object size, closer to the real object size, and also strengthen the detection effect of small objects.

However, YOLOv3 still has low utilization of features and poor detection performance for small objects and object-intensive images. In the application of actual helmet wearing detection, small objects and dense objects are very common. Therefore, we need a network with better performance to make the helmet-wearing detection system more robust.

2.2 Attention Mechanism

Both language and vision problems contain information that is closely related to the research task and some irrelevant information. The attention mechanism can help the algorithm focus on analyzing some vital information while ignoring irrelevant information. In recent years, the attention mechanism has been used in various tasks and achieved good results. Bahdanau et al. [30] are the first to use the attention mechanism to solve machine translation problems. Wang et al. [31] draw on the idea of non-local mean filtering to model the correlation of arbitrary non-local features. Non-local operations would not change the size of features and can be embedded in any network. Hu et al. [32] achieved first place in the 2017 ILSVRC classification task through the attention mechanism and adaptive calibration of the channel direction's characteristic response. DAnet [33] can be considered as a specialized example of Nolocal-Net [31], which solves the problem of scene segmentation by capturing rich contextual relevance through a self-attention mechanism.

Different from previous work, this paper adds attention mechanisms to the task of helmet detection. Based on YOLOv3, we have added a spatial attention module and a channel attention module to capture the correlation between features, enrich context information, and improve the detection ability of feature representation in object detection.

3 Our Methods

3.1 AT-YOLO Network Construction

This paper designs the AT-YOLO deep model to optimize the feature expression ability and feature learning ability of the network. The backbone and neck of the YOLOv3 have been added channel attention modules and a spatial attention module, respectively. The network adaptively captures the correlation between the channel and space of the feature and models the global context relationship, improving the feature representation ability of the object detection algorithm. The AT-YOLO helmet detection framework is shown in Fig. 1.

Channel attention block. The Darknet-53 network used residuals linking to merge the features of different layers to alleviate the problem of gradient disappearance. However, the network does not make good use of the dependence between the channel of the feature map. Inspired by [32], we insert CA-block (Channel Attention block) into each residual block in Darknet. Recalibrate the dependencies between channels by learning global features, selectively enhancing high-contribution information, suppressing low-contribution information, and improving the

feature expression ability and feature utilization of the network. The CA-block we designed is shown in Fig. 2.

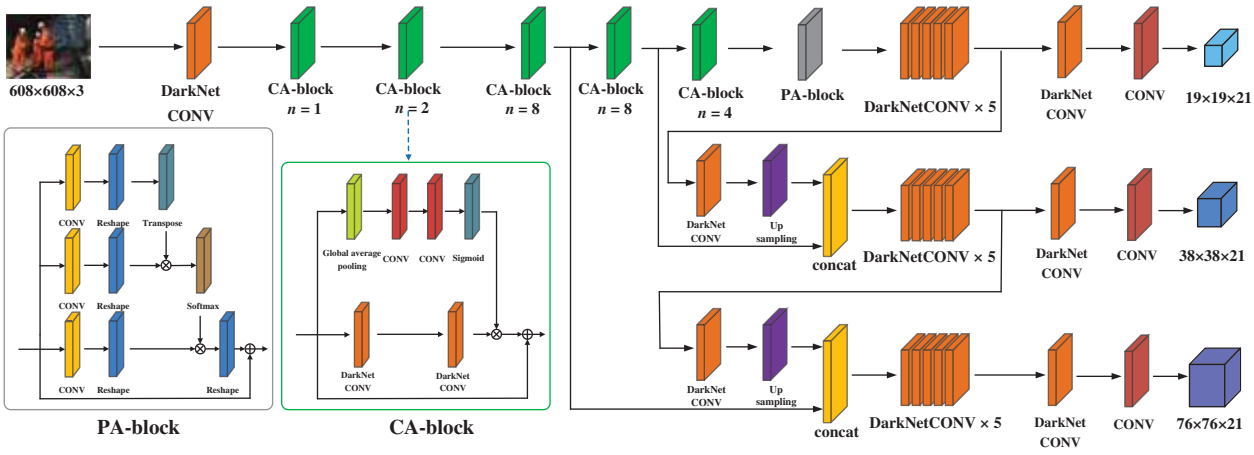

Figure 1: AT-YOLO helmet wearing detection model frame

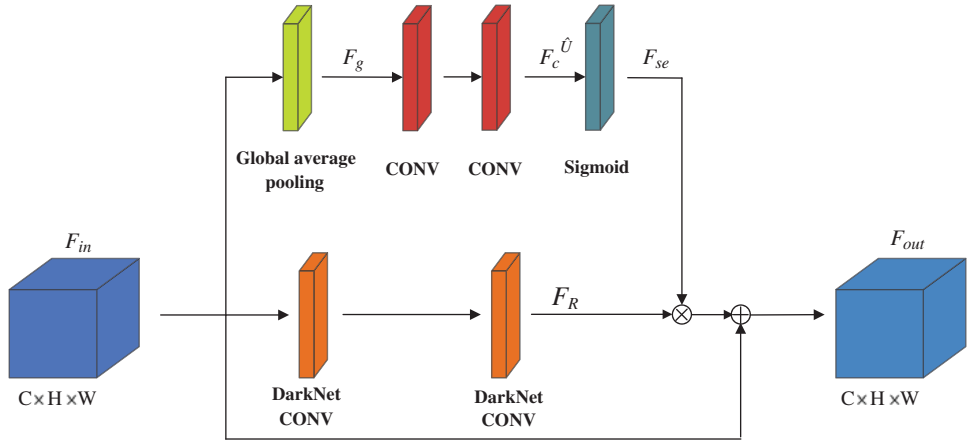

Figure 2: Channel attention block

Suppose the input feature map F_{in}, first obtain the feature map $F_R \in R^{C\times H\times W}$ through two DarkNet convolutions.

$$F_R = Res\,(F_{in}) \tag{1}$$

The feature map F_{in} obtains a global feature map $F_g \in R^{C\times 1\times 1}$ after the global average pooling operation. Subsequently, two convolution kernels with a size of 1 × 1 are performed to obtain the information of the global receptive field, and simultaneously, the feature dependence in the channel is learned. The feature vector $F'_C \in R^{1\times 1\times C}$ was obtained at this time. Then activate the feature vector F'_C to get F_{se}. F_{se} is used to describe the learned channel weight. Finally, the

learned channel weights are used to weight the feature map F_R, and then the feature maps of the residual link are combined to obtain the final output.

$$F_{out} = F_{in} + F_{se} \times F_R \tag{2}$$

Position attention block: In object detection, there are many detection objects dense or object occlusion, which leads to object false detection. Therefore, to obtain richer contextual information and enhance the expressive ability of feature maps, this paper adds a spatial attention module PA-block (Position Attention block), as shown in Fig. 3.

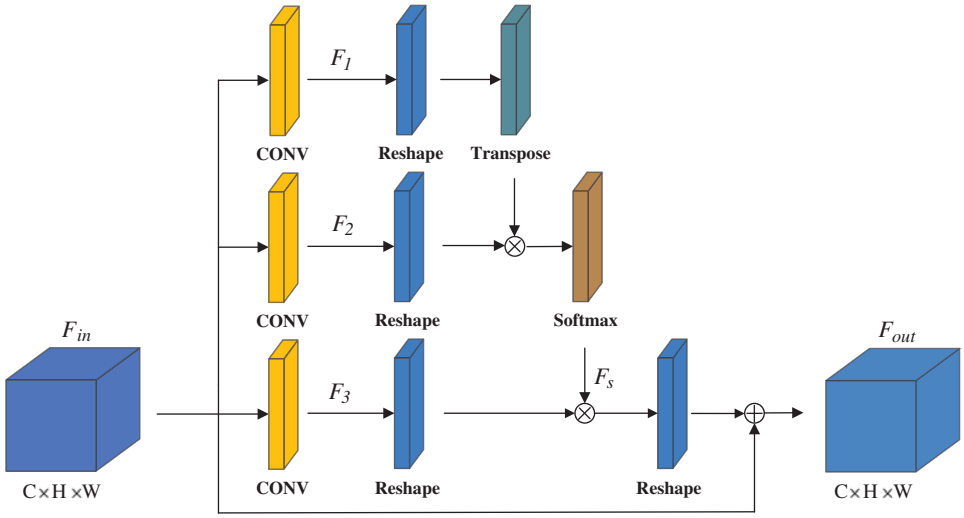

Figure 3: Position attention block

Suppose the input local feature map is $F_{\text{in}} \in R^{C\times H\times W}$. First, two new feature maps F_1 and $F_2$are obtained through a convolutional layer $\{F_1, F_2\} \in R^{C\times H\times W}$. Then reshape them to obtain feature maps with the scale of $R^{C\times N}$ where $N = H \times W$. Then perform matrix multiplication on the F_1 and F_2 transpose, and the spatial attention feature map$F_s \in R^{N\times N}$ is obtained after normalization processing with *softmax* activation function:

$$F_{s_{ji}} = \frac{\exp\left(F_{1_i} \cdot F_{2_j}\right)}{\sum_{i=1}^{N} \exp\left(F_{1_i} \cdot F_{2_j}\right)} \tag{3}$$

where $F_{s_{ji}}$ represents the degree of influence of the i-th position on the j-th position. The greater the connection between the two locations, the more similar their feature representations.

Simultaneously, we convolve the feature map F_{in} to obtain a feature map $F_3 \in R^{C\times H\times W}$ and then reshape it to obtain a feature map of size $R^{C\times N}$. Then, matrix multiplication is performed on the deformations of F_3 and F_s, and the result is reshaped to obtain a feature map of size $R^{C\times H\times W}$. Finally, multiply it with the scale parameter α and perform element addition with the feature map F_{in} to obtain the final output $F_{out} \in R^{C\times H\times W}$, which is calculated as follows:

$$F_{out_j} = \alpha \sum_{i=1}^{N} \left(F_{s_{ji}} F_{3_i}\right) + F_{1_j} \tag{4}$$

where the scale parameter α is initialized to 0, and it is learned to give greater weight to different location features during the training process. It can be seen from the above formula that F_{out} each point is the weighted sum of the original feature F_{in} and features across all locations. Therefore, the F_{out} has a global receptive field and selectively aggregates the context information in the location attention feature map.

3.2 DIoU Bounding Box Regression Loss Function

In the YOLOv3, the MSE (mean square error) loss function is used for bounding box regression, but the mean square error has problems in evaluating the prediction results. That is, the mean square error loss function is quite sensitive to the object scale. In order to solve the above problems, we modified the bounding box regression loss function part, using DIoU [34] loss function.

Besides, using the IoU of the real boxes and the predicted boxes can better reflect the detection effect of the predicted boxes. When the prediction boxes are completely contained in the real boxes, the relative position between the prediction boxes and the real boxes is ambiguous, and the gradient descent is easy to occur slowly.

Therefore, when training the model, this paper uses DIoU as the bounding box regression loss function, which is defined as follows:

$$L_{DIoU} = 1 - IoU + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} \tag{5}$$

IoU means the intersection ratio of the predicted boxes and the real boxes, b, b^{gt} respectively represent the center points of the predicted boxes and the real boxes, ρ represents the calculation of the Euclidean distance between the two central points, and c represents the diagonal distance of the minimum closure distance area that can contain both the predicted box and the real box, as shown in Fig. 4.

The DIoU loss function increases the normalized distance between the predicted box and the real box based on IoU. When the predicted box is completely contained in the real box, it can still provide the moving direction for the bounding box to make the network convergence faster and more accurate.

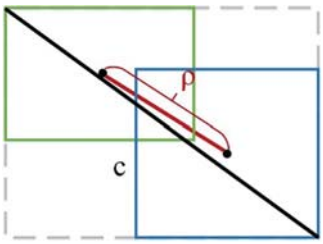

Figure 4: Center point distance normalization

3.3 Training Strategy

Without changing the network structure, different training strategies get different results. Thanks to Zhang et al. [35], we have explored different training strategies to enhance the performance of the model without adding additional calculations to model inference.

Label smoothing. In order to prevent the network from over-confidence about the category probability prediction and causing overfitting, we use a label smoothing strategy. In the classification model, we use a one-hot form of encoding to predict the probability that the sample belongs to each label. However, there would be such a label encoding that the model may be too confident in the prediction to ignore the small number of sample labels in the actual process. Therefore, this paper adopts the label smoothing strategy for learning. The formula for the real probability of label smoothing is:

$$p = (1 - \varepsilon)\, y + \frac{\varepsilon}{K} \tag{6}$$

where y is the label probability generated by one-hot encoding, and K is the number of categories. There are two categories of hat and person, so $K = 2$. We set $\varepsilon = 0.01$, the probability value $p = 0.995$ after label smoothing when the true label $y = 1$ of a certain sample. Label smoothing can avoid the absolutization of the predicted value of the network, which can improve the overall effect.

Multi-scale input. Due to the different resolutions in the captured natural images and the memory limitations of different operating devices, we need to improve the adaptability of the network to different scale image input, so we use multi-scale input for model training.

During training, each iteration of the network randomly selects pictures of different sizes for training. Since the network uses a total of 32 times downsampling, the size of our image input is a multiple of 32, for example: {320, 352,..., 608}. Therefore, the smallest input picture size is 320 $\times$ 320, and the largest is 608 $\times$ 608. There are a total of ten different sizes of the input pictures.

Using this method can force the network to learn predictions of various input scales while preventing the network from overfitting. The speed of the network is faster under a smaller input size, and the smaller the memory of the running device is. In the case of high-resolution image input, it has higher accuracy.

4 Experimental Results and Analysis

4.1 Datasets

Dataset acquisition. Part of the dataset in this paper comes from Safety-Helmet-Wearing-Dataset, which contains 7581 pictures and the label information about whether a person wears a helmet or not. The label contains two detection categories, *person* and *hat*, where the *hat* represents the head of the person wearing a safety helmet, the *person* represents the person's head without a safety helmet. This dataset contains a large number of crowded scene pictures, but there is a phenomenon of unbalanced categories, as shown in Fig. 5. The number of hat and person instances is 9031 and 111514, respectively, and the category ratio is close to 1:11. Simultaneously, this dataset has fewer detection scenarios, which easily causes the network to overfit, low generalization ability, and poor multi-scene detection ability. We used web crawlers to grab pictures of wearing helmets on the Internet. At the same time, we selected video clips containing hard hat scenes and cropped them into images to expand the dataset, alleviating the imbalance of dataset categories and enhancing the generalization ability of the model. Particularly, we have selected pictures of wearing ordinary hats so that the network can better distinguish the difference between safety helmets and ordinary hats.

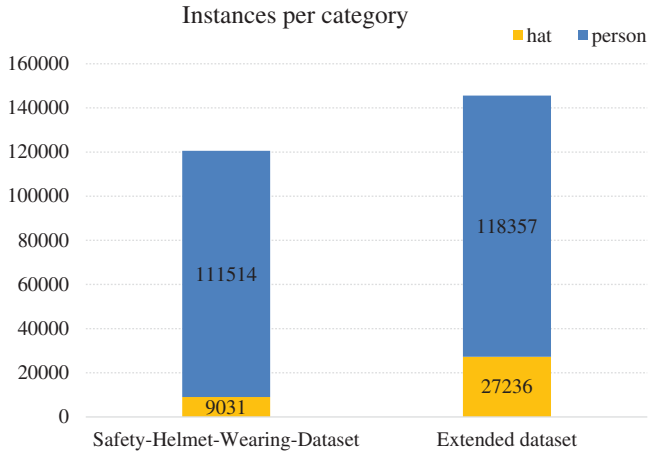

Figure 5: Comparison of the number of extended categories in the dataset

Data cleaning. A large number of similar pictures in the dataset would affect the model performance. Therefore, we use the deep learning model DenseNet [36], which is pre-trained on ImageNet [37] to extract the features of each image in the dataset and then calculate the Euclidean distance of the image features of each two pictures in the dataset to determine the similarity of the two images. The manually annotate the collected image, following the VOC labeling format, marking the coordinates of the upper left and lower right vertices of the bounding box of the object and the bounding box category, and stored in the XML file. Finally, we obtained 13620 pictures, the number of hat and person instances is 27236 and 118357, respectively, and the category ratio is close to 1:4 (see Fig. 5), which greatly alleviates the category imbalance of the dataset.

Data augmentation. To avoid possible over-fitting of the network, we use six methods for data augment: random horizontal flipping of images, image cropping, image filling, color dithering, brightness augment, and mixup [38].

4.2 Implementation Details

Due to the uneven distribution of the dataset categories, we use random sampling to randomly select 1000 images on the dataset as the test set, 500 images as the validation set, and 12260 images as the training set.

All experiments are completed in Intel (R) Core (TM) i7-7800X CPU@3.50 GHz, 64.00 GB RAM and Nvidia GeForce GTX 1080Ti GPU, the Tensorflow framework is adopted.

SGD is used as training optimizer, the initial learning rate is set to 0.0001, and it is changed every iteration. The momentum coefficient is 0.9, the L2 weight decay coefficient is 0.0005, the batch normalized decay coefficient is 0.99, and the batch size is set to 4. Iterate for 400 epochs and save the model weight with the least loss value of the validation set.

4.3 Ablation Experiments for Training Strategy

To compare the impact of different training strategies on the network, we explored the impact of label smoothing, multi-scale input, and data augmentation on network detection performance, as shown in Tab. 1, where *Baseline* represents the baseline model of YOLOV3, *LS*

represents label smoothing, *Multi-Input* represents random multi-scale input, and *DA* represents data augmentation.

Table 1: Network detection performance comparison under different training strategies

Method	LS	Multi-input	DA	mAP (%)
Baseline				87.16
Baseline	√			90.31
Baseline	√	√		91.93
Baseline	√	√	√	93.90

Compared with YOLOv3, the label smoothing increases mAP by 3.15%, which proved that the label smoothing strategy could effectively prevent network overfitting. The multi-scale increases mAP by 1.62%, which can make the network adapt to more input size images, thus adapting to the memory limitations of different code running devices and preventing network overfitting. The data augmentation increases mAP by 1.97%, the color dithering helps the network recognize different color scenes, and the brightness augmentation helps the network recognize images in dark or bright environments.

4.4 Ablation Experiments for Attention Module

In order to explore the influence of the attention mechanism on the network model, we set up the experiment in Tab. 2. We insert CA-block into darknet-53 to make the network focus on channel features with a more significant contribution. After the last module of the backbone network, the spatial attention module PA-block is used to obtain richer context information. To verify the improvement of the model performance by this design, the network designed in this paper is compared with the object detection network YOLOv3 when the input image size is 608 × 608, and the training strategy is the same. The detection accuracy improvement effect is shown in Tab. 2.

Table 2: Attention module ablation contrast experiment

Method	mAP (%)	FPS	Params
YOLOv3	93.90	30.3	61.5 M
YOLOv3 + CA	95.62	27.4	62.3 M
AT-YOLO(YOLOv3 + CA + PA)	96.30	25.7	74.2 M
AT-YOLO + DIOU	**96.50**	**25.7**	**74.2 M**

It can be seen from Tab. 2 that compared with YOLOv3, after adding CA-block, the accuracy is increased by 1.72%, reaching an accuracy of 95.62%, and the detection speed is 27.4 FPS. After adding PA-block, the accuracy is increased by 0.68%, reaching an accuracy of 96.30%, while the detection speed is maintained at 25.7 FPS. We can find that although the computational cost increases after the dual attention module are added, the frame rate of the general video does not

exceed 25 FPS, which can meet the requirements of real-time video detection. Moreover, we also use the DIoU bounding box loss function to train the model, which improves the convergence speed of the network, makes the bounding box prediction more accurate, improves the detection accuracy of the network, and makes the model testing accuracy reach 96.50%.

Considering, our detection model is used for video surveillance of construction sites. Since the location of the surveillance camera is generally far away from the surveillance scene, and the pixel size of the worker in the video is small, we designed a comparative experiment on the detection accuracy of the model for small objects (pixel area $< 32^2$).

It can be seen from Tab. 3 that the proposed network can also achieve better detection accuracy when detecting small objects. That's because small objects can be inferred through the contextual information aggregated by the attention mechanism, and feature representation can be highlighted, thereby improving the detection results.

Table 3: Accuracy comparison under the small object

Method	mAP_{small} (%)
YOLOv3	30.7
YOLOv3 + CA	43
AT-YOLO (YOLOv3 + CA + PA)	45.1
AT-YOLO + DIoU	**45.7**

We also compared the speed and accuracy of AT-YOLO with an input size of 320 × 320 (AT-YOLO 320), AT-YOLO with an input size of 512 × 512 (AT-YOLO 512), AT-YOLO with an input size of 608 × 608 (AT-YOLO 608), and YOLOv3, then drawn a speed-accuracy curve. As shown in Fig. 6, although AT-YOLO increases the attention mechanism, it is slower than YOLOv3 under the same input size, but under the premise of the same inference speed, AT-YOLO has higher accuracy than YOLOv3. AT-YOLO is more efficient and accurate.

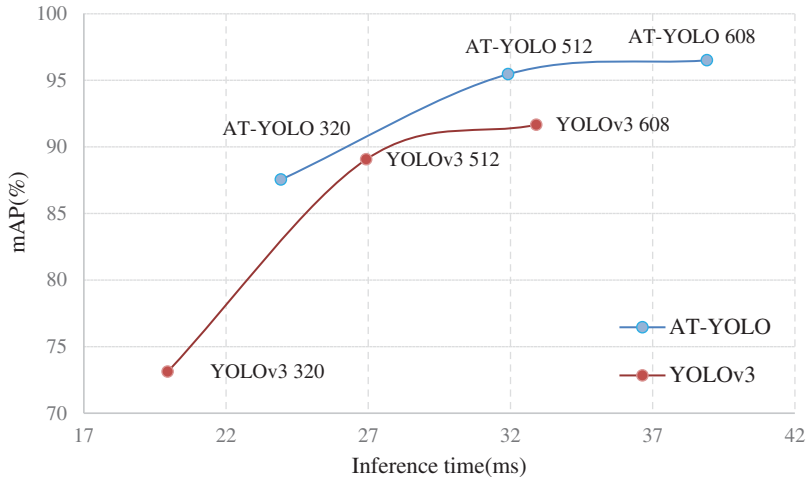

Figure 6: Speed (ms) *vs.* accuracy (mAP)

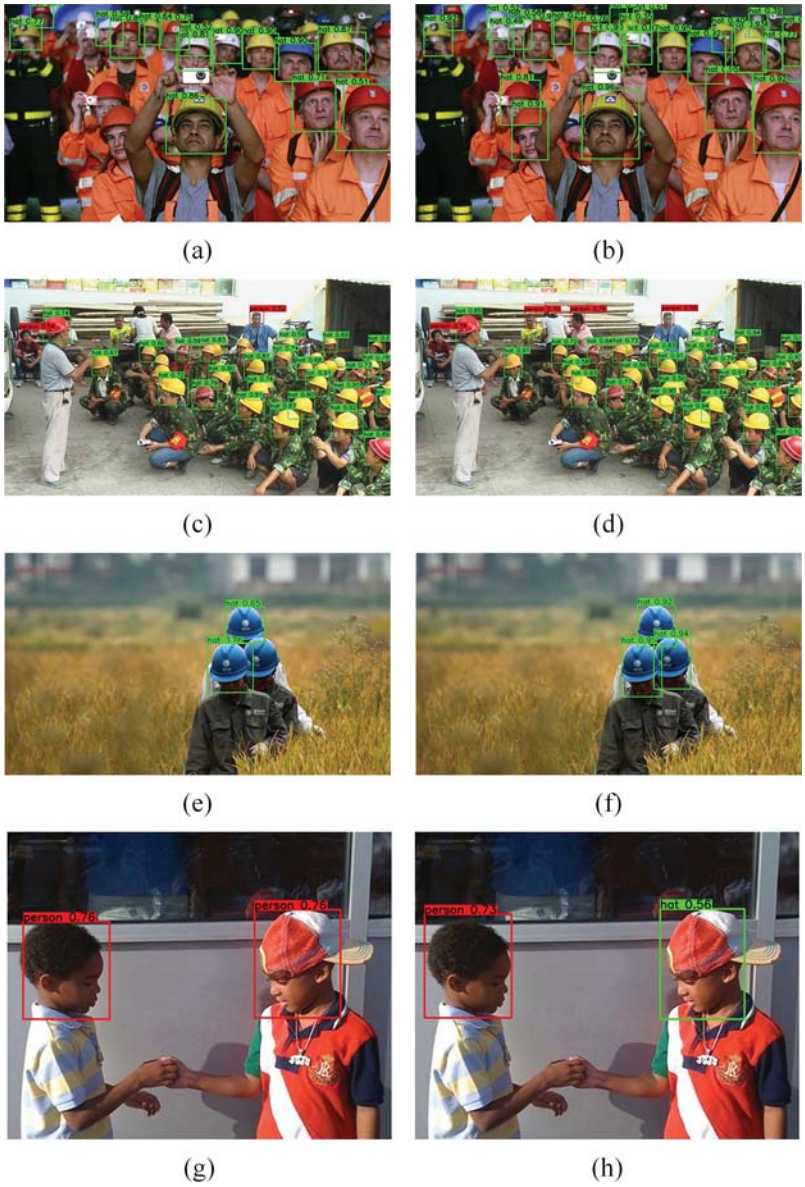

Figure 7: (a), (c), (e) are the detection results of YOLOv3, (b), (d), (f) are the detection results of AT-YOLO, (g) is the detection result without adding the ordinary hat image in the dataset, (h) is the detection result with adding the ordinary hat image in the dataset

The visualization test results are shown in Fig. 7, where (a), (c), (e) are the detection results of YOLOv3, (b), (d), (f) are the detection results of AT-YOLO, (g) is the detection result without adding the ordinary hat image in the dataset, (h) is the detection result with adding the ordinary hat image in the dataset. For (a), (b) with dense and more occluded objects, and (e), (f) with occluded objects, some occluded objects can't be detected by YOLOv3 but can be detected by AT-YOLO. Since AT-YOLO extracts the surrounding information of the object through the spatial attention module, making it easier to infer the occluded object. For (c), (d) with low resolution. AT-YOLO analyzes the interdependence between the channels through the channel attention

mechanism, extracts more effective features, and detects low-pixel images excellent. For (g), (h) containing ordinary hats, it is easier for the model to distinguish ordinary hats from safety helmets after adding ordinary hats into the dataset.

4.5 Performance Comparison of Different Models

We tested our dataset on different networks, as shown in Tab. 4. The mAP of our method is 9.34% higher than that of YOLOv3 without increase too much inference cost. Compared with the two-stage method (Faster-RCNN with FPN [39]), our method is better in accuracy and efficiency.

Table 4: Comparison of the performance of different models

Models	Input size	Param	FPS	mAP (%)
YOLOv3	608 × 608	61.5 M	30.3	87.16
Faster-RCNN w FPN	800 × 800	–	11.62	91.15
AT-YOLO (ours)	608 × 608	74.2 M	25.7	**96.50**

5 Conclusion

This paper proposes an AT-YOLO helmet wearing detection model. By introducing the attention mechanism into the YOLOv3, the modeling ability of the network on the dependencies between different positions in the image is enhanced to extract more accurate features and effectively improve the model's feature representation ability. At the same time, we combine the optimized training strategy to improve the network performance without increasing the inference cost. To verify the performance of these proposed methods, we produced a dataset and conducted some evaluation tests on it. The experimental results show that the methods proposed in this paper effectively improve the performance of the AT-YOLO network, thereby providing an excellent solution for the helmet-wearing detection system in actual scenarios.

In the next work, we will combine the correlation of time series to optimize the video's object detection to improve the accuracy and speed of detection and make it more suitable for the safety helmet wearing detection system.

Acknowledgement: The author would like to thank the support of Central South University of Forestry & Technology and the support of National Natural Science Fund of China.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China under Grant 61772561, author J. Q, http://www.nsfc.gov.cn/; in part by the Degree & Postgraduate Education Reform Project of Hunan Province under Grant 2019JGYB154, author J. Q, http://xwb.gov.hnedu.cn/; in part by the Postgraduate Excellent teaching team Project of Hunan Province under Grant [2019]370-133, author J. Q, http://xwb.gov.hnedu.cn/; in part by the Science Research Projects of Hunan Provincial Education Department under Grant 18A174, author X. X, http://kxjsc.gov.hnedu.cn/; in part by the Science Research Projects of Hunan Provincial Education Department under Grant 19B584, author Y. T, http://kxjsc.gov.hnedu.cn/; in part by the Natural Science Foundation of Hunan Province (No.2020JJ4140), author Y. T, http://kjt.hunan.gov.cn/; and in part by the Natural Science Foundation of Hunan Province (No. 2020JJ4141), author X. X, http://kjt.hunan.gov.cn/; in part by the Key Research and Development

Plan of Hunan Province under Grant 2019SK2022, author Y. T, http://kjt.hunan.gov.cn/; in part by the Graduate Science and Technology Innovation Fund Project of Central South University of Forestry and Technology under Grant CX2020107, author Q. Z, https://jwc.csuft.edu.cn/.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Q. Zhu, M. C. Yeh, K. T. Cheng and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, New York, USA, vol. 2, pp. 1491–1498, 2006.

[2] F. Zuo and P. H. N. de With, "Fast facial feature extraction using a deformable shape model with Haar-wavelet based local texture attributes," in *2004 Int. Conf. on Image Processing*, Singapore, vol. 3, pp. 1425–1428, 2004.

[3] Y. Luo, J. Qin, X. Xiang and Y. Tan, "Coverless image steganography based on multi-object recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. https://doi.org/10.1109/TCSVT.2020.3033945.

[4] J. Qin, J. Wang, Y. Tan, H. Huang, Z. He *et al.,* "Coverless image steganography based on generative adversarial network," *Mathematics*, vol. 8, no. 9, pp. 1394–1404, 2020.

[5] Q. Liu, X. Xiang, J. Qin, Y. Tan, Y. Luo *et al.,* "Coverless steganography based on image retrieval of DenseNet features and DWT sequence mapping," *Knowledge-Based Systems*, vol. 192, no. 2, pp. 105375–105389, 2020.

[6] Z. Yang, S. Zhang, Y. Hu, Z. Hu and Y. Huang, "VAE-Stega: Linguistic steganography based on variational auto-encoder," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 880–895, 2021.

[7] L. Y. Xiang, S. H. Yang, Y. H. Liu, Q. Li and C. Z. Zhu, "Novel linguistic steganography based on character-level text generation," *Mathematics*, vol. 8, no. 9, pp. 1558, 2020.

[8] J. Qin, W. Pan, X. Xiang, Y. Tan and G. Hou, "A biological image classification method based on improved CNN," *Ecological Informatics*, vol. 58, no. 4, pp. 1–8, 2020.

[9] J. Wang, J. H. Qin, X. Y. Xiang, Y. Tan and N. Pan, "CAPTCHA recognition based on deep convolutional neural network," *Mathematical Biosciences and Engineering*, vol. 16, no. 5, pp. 5851–5861, 2019.

[10] T. Q. Zhou, B. Xiao, Z. P. Cai and M. Xu, "A utility model for photo selection in mobile crowdsensing," *IEEE Transactions on Mobile Computing*, vol. 20, no. 1, pp. 48–62, 2021.

[11] Z. Wang, J. Qin, X. Xiang and Y. Tan, "A privacy-preserving and traitor tracking content-based image retrieval scheme in cloud computing," *Multimedia Systems*, 2021. https://doi.org/10.1007/s00530-020-00734-w.

[12] W. Ma, J. Qin, X. Xiang, Y. Tan and Z. He, "Searchable encrypted image retrieval based on multi-feature adaptive late-fusion," *Mathematics*, vol. 8, no. 6, pp. 1–15, 2020.

[13] Y. Tan, L. Tan, X. Xiang, H. Tang and J. Qin, "Automatic detection of aortic dissection based on morphology and deep learning," *Computers, Materials & Continua*, vol. 62, no. 3, pp. 1201–1215, 2020.

[14] Y. Chen, L. Liu, J. Tao, R. Xia, X. Chen *et al.,* "The improved image inpainting algorithm via encoder and similarity constraint," *Visual Computer*, vol. 28, no. 3, pp. 1–15, 2020.

[15] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, Ohio, USA, pp. 580–587, 2014.

[16] R. Girshick, "Fast R-CNN," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 1440–1448, 2015.

[17] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[18] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, pp. 779–788, 2016.

[19] B. E. Mneymneh, M. Abbas and H. Khoury, "Automated hardhat detection for construction safety applications," *Procedia Engineering*, vol. 196, pp. 895–902, 2017.

[20] K. Li, X. Zhao, J. Bian and M. Tan, "Automatic safety helmet wearing detection," arXiv preprint, 2018. https://arxiv.org/abs/1802.00264.

[21] F. Wu, G. Jin, M. Gao, Z. He and Y. Yang, "Helmet detection based on improved YOLO V3 deep model," in *2019 IEEE 16th Int. Conf. on Networking, Sensing and Control*, Banff, Canada, pp. 363–368, 2019.

[22] X. Long, W. Cui and Z. Zheng, "Safety helmet wearing detection based on deep learning," in *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conf.*, Beijing, China, pp. 2495–2499, 2019.

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, A. C. Berg *et al.,* "SSD: Single shot multibox detector," in *European Conf. on Computer Vision*, Cham: Springer, pp. 21–37, 2016.

[24] S. Chen, W. Tang, T. Ji, H. Zhu, W. Wang *et al.,* "Detection of safety helmet wearing based on improved Faster R-CNN," in *2020 Int. Joint Conf. on Neural Networks*, Glasgow, UK, pp. 1–7, 2020.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, A. Rabinovich *et al.,* "Going deeper with convolutions," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, Massachusetts, USA, pp. 1–9, 2015.

[26] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, pp. 7263–7271, 2017.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint, 2014. https://arxiv.org/abs/1409.1556.

[28] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint, 2018. https://arxiv.org/abs/1804.02767.

[29] K. He, X. Zhang, S. Re and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, pp. 770–778, 2016.

[30] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint, 2014. https://arxiv.org/abs/1409.0473.

[31] X. Wang, R. Girshick, A. Gupta and K. He, "Non-local neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, pp. 7794–7803, 2018.

[32] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, pp. 7132–7141, 2018.

[33] J. Fu, J. Liu, H. Tian, Y. Li, H. Lu *et al.,* "Dual attention network for scene segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Long Beach, California, USA, pp. 3146–3154, 2019.

[34] Z. Zheng, P. Wang, W. Liu, J. Li, D. Ren *et al.,* "Distance-IoU loss: Faster and better learning for bounding box regression," *AAAI Conf. on Artificial Intelligence*, vol. 34, no. 7, pp. 12993–13000, 2020.

[35] Z. Zhang, T. He, H. Zhang, Z. Zhang, M. Li *et al.,* "Bag of freebies for training object detection neural networks," arXiv preprint, 2019. https://arxiv.org/abs/1902.04103.

[36] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, pp. 4700–4708, 2017.

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, A. C. Berg *et al.,* "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[38] H. Zhang, M. Cisse, Y. N. Dauphin and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," arXiv preprint, 2017. https://arxiv.org/abs/1710.09412.

[39] T. Y. Lin, P. Dollár, R. Girshick, K. He, S. Belongie *et al.,* "Feature pyramid networks for object detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, pp. 2117–2125, 2017.