Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.036893
Article

A Model for Helmet-Wearing Detection of Non-Motor Drivers Based on YOLOv5s

Hongyu Lin, Feng Jiang*, Yu Jiang, Huiyin Luo, Jian Yao and Jiaxin Liu

College of Computer & Information Engineering, Central South University of Forestry and Technology, Changsha, 410004, China

*Corresponding Author: Feng Jiang. Email: jf09mail@126.com

Received: 15 October 2022; Accepted: 08 February 2023

Abstract: Detecting non-motor drivers' helmets has significant implications for traffic control. Currently, most helmet detection methods are susceptible to the complex background and need more accuracy and better robustness of small object detection, which are unsuitable for practical application scenarios. Therefore, this paper proposes a new helmet-wearing detection algorithm based on the You Only Look Once version 5 (YOLOv5). First, the Dilated convolution In Coordinate Attention (DICA) layer is added to the backbone network. DICA combines the coordinated attention mechanism with atrous convolution to replace the original convolution layer, which can increase the perceptual field of the network to get more contextual information. Also, it can reduce the network's learning of unnecessary features in the background and get attention to small objects. Second, the Rebuild Bidirectional Feature Pyramid Network (Re-BiFPN) is used as a feature extraction network. Re-BiFPN uses cross-scale feature fusion to combine the semantic information features at the high level with the spatial information features at the bottom level, which facilitates the model to learn object features at different scales. Verified on the proposed "Helmet Wearing dataset for Non-motor Drivers (HWND)," the results show that the proposed model is superior to the current detection algorithms, with the mean average precision (mAP) of 94.3% under complex background.

Keywords: Helmet-wearing detection; dilated convolution; feature pyramid network; feature fusion

1 Introduction

Private or shared electric bicycles have become popular, bringing traffic convenience and safety hazards. Some cities have issued helmet safety initiatives, proposing that non-motor drivers wear safety helmets to guarantee travel safety. Currently, the supervision of non-motor drivers wearing helmets mainly relies on traffic policies to monitor the road scene. This method is labor-intensive, tends to omit objects, and disallows effective monitoring during inclement weather. Therefore, automatic

monitoring can effectively solve the shortcomings of manual monitoring. Object detection, a hotspot of computer vision, provides excellent convenience for intelligent surveillance.

Traditional object detection algorithms are suitable for targets with apparent features and simple backgrounds. However, the variable backgrounds and complex targets in the actual scenes prevent excellent detection results. Object detection algorithms based on deep learning have applicated widely in road vehicle monitoring, medical research, mask-wearing detection, video surveillance detection, image classification, and others [1–6]. Deep learning-based object detection algorithms mainly include two-stage algorithms based on region extraction and one-stage algorithms based on regression. The two-stage object detection algorithm first extracts the region of interest from the input image to generate candidate frames. Then it performs regression classification on the candidate frames in the second step. Standard two-stage algorithms include Region CNN (R-CNN) [7], Fast R-CNN [8], Faster R-CNN [9], etc. One-stage object detection algorithm omits the step of generating candidate regions and performs feature extraction, target regression and classification directly in the same CNN. The main algorithms include Single Shot MultiBox Detector (SSD) [10] and You Only Look Once (YOLO) [11–14]. Experiments show that the two-stage object detection algorithm surpasses the one-stage detection algorithm in terms of detection accuracy, but the one-stage algorithm is faster [15].

In the non-motor drivers' helmet-wearing detection tasks, capturing accurately under complex and changing actual scenes is difficult because the monitoring images have many targets, few pixels, rich colors, and similar shapes to ordinary hats. In addition, the detection tasks also have strict requirements for real-time. At present, helmet detection algorithms based on deep learning are developing rapidly. However, there are still problems, such as low accuracy for tiny targets, poor robustness and complex operations. There is also a need for wealthy non-motor drivers' helmet-wearing datasets to evaluate algorithms' performance comprehensively. This paper creates a new dataset HWND, containing different image attributes (e.g., different categories, different angles, different weather) to restore the natural environment of the detection scene as much as possible. In addition, this article uses the improved YOLOv5s network and achieves intelligent detection of helmet-wearing for non-motor drivers. The main contributions in this paper can be summarized as follows:

(1) Construct a dataset of non-motor drivers wearing helmets. A portion of the dataset comes from our shots in realistic traffic scenarios containing 1332 high-quality images with multiple targets, ample categories and complex backgrounds. The images cover non-motor drivers in excellent road conditions, angles, lighting, and road congestion conditions. The other part is from the public dataset "Bike Helmets Detection," which contains 764 images.
(2) Propose the helmet-wearing detection model YOLOv5s-Dilated and Rebuild (YOLOv5s-DR) for non-motor drivers. This model combines the dilated convolution with the attention mechanism to increase the perceptual field, and the network acquires more information. Also, it establishes the dependency of target features in the channel and spatial dimensions so that the model pays more attention to small targets such as helmets and heads and enhances the network detection ability for small targets. The weighted Bi-directional Feature Pyramid Network (BiFPN) [16] is used for feature extraction, and a cross-scale feature fusion network is added. The blended feature map contains rich semantic and location information at different levels.

The rest of the paper is organized as follows: It summarizes related literature in Section 2. Dataset HWND is introduced in Section 3. The details of the improved YOLOv5-DR and the underlying methods are described in Section 4. Finally, it analyzes the experimental environment and results in Section 5 and the conclusion in Section 6.

2 Related Work

The helmet-wearing detection of non-motor drivers based on deep learning is still in the early research stage. However, it is similar to the wearing detection of motorcycle helmets and helmets on construction sites. This section classifies the helmet-wearing detection algorithms into two-stage and one-stage to summarize the related literature.

2.1 Two-Stage

Yogameena et al. [17] first segmented the foreground target using the Gaussian model and then used Faster R-CNN to detect motorcycle helmet wear. Chen et al. [18] adopted the Retinex image enhancement technique to improve image quality. K-means++ clustering algorithm was introduced to cluster the helmet sizes in the images, and helmet wearing was detected using the improved Faster-RCNN algorithm. The literature [19] employed multi-scale training and increasing anchors strategies to enhance the robustness of the Faster R-CNN algorithm. Online Hard Example Mining (OHEM) [20] was used to optimize the model and prevent imbalance between positive and negative samples. The issue [21] applied Faster R-CNN and SSD algorithms to detect helmets for motorcycle drivers. A comparative analysis of the detection effects was performed to derive suitable application scenarios for the different algorithms. The results showed that Faster R-CNN captures images slower but performs better in accuracy and can be deployed in places where vehicles move forward slowly. SSD is faster and less accurate, suitable for deployment in fast-flowing traffic scenarios like highways.

2.2 One-Stage

Aiming at the matching problem of small targets and detection head scale, Chen et al. [22] changed the backbone network layer number of YOLOv4, removed the deep feature layer that lacked semantic information, and added a shallow feature layer to enhance the detection effect on small targets. Reference [23] improved based on YOLOv5 by adding a triple attention mechanism to the last layer of the backbone network. In the parallel three-branch structure, two extract the interdimensional dependencies between spatial and channel dimensions, and another extracts the spatial feature dependencies. As a result, the occlusion problem under congestion was solved. Han et al. [24] proposed a cross-layer attention mechanism approach to refine the features of targets. Based on the SSD model, a spatial attention mechanism is used for low-level features, and the high-level features use a channel attention mechanism. The literature [25] used the Sandglass-Residual [26] model for the feature extraction process based on the lightweight algorithm YOLOv3-Tiny to reduce the parameters and computational effort of the network. Paper [27] used the cross-stage hierarchy [28] module to replace a stack of the module made of convolution, BatchNormalization and LeakyReLU (CBL) in the feature pyramid of YOLOv4. This improvement improved the feature richness and reduced the memory and computational effort. Sadiq et al. [29] combined the YOLOv5 with a fuzzy-based data enhancement module to effectively remove the noise in the monitoring system and improve image clarity. The literature [30] used deformable convolutional networks [31] instead of the traditional convolution in the backbone. The Convolutional Block Attention Module Network (CBAM) [32] was also introduced in the neck. This method solved the challenges posed by complex construction environments, dense targets, and irregular shapes of safety helmets. Reference [33] proposed an improved hierarchical matching positive sample strategy. The Intersection over Union (IoU) of Prior_box and Ground Truth (GT) was used as the basis. For $0.1 < \text{IoU} < 0.2$, the feature points at the location of the GT centroid are selected as positive samples, the two adjacent feature points of the GT centroid at the grid position are used as additional positive samples while $0.2 < \text{IoU} < 0$, and if $\text{IoU} > 0.5$, the four feature points

closest to the centroid are selected as positive samples. This hierarchical strategy effectively improved the feature learning ability of the network.

Essays [17–19] have achieved excellent detection accuracy, but real-time detection is still a problem. Papers [22–25,27,29,30,33] take into account the interference factors existing in the factual background and make a series of measures to improve the ability of the network to extract features, which effectively improves the detection accuracy of helmets. In real scenarios, a helmet occupies a small number of pixels in an image. However, the features learned by the network are limited, which will inevitably cause the loss of small target information. Therefore, this work focuses on improving the feature expression of small objects, making the network learn complete and richer features about small objects.

3 HWND Dataset

Dataset is a primary condition for implementing experiments and evaluating algorithms' performance in helmet detection tasks. However, a large dataset of non-motorized drivers wearing helmets needs to be improved in research. In this section, the paper proposes a high-quality helmet-wearing detection dataset for non-motor drivers named "Helmet Wearing dataset for Non-motor Drivers (HWND)," which mainly covers electric bicycles in separate streets. The HWND has 2096 images in JPG format, which consists of two parts: the publicly available dataset "Bike Helmets Detection [34]" on the Kaggle website and our photographic collection. Next describes the process of creating the benchmark dataset and the information about the composition of the dataset.

3.1 Image Screening and Annotation

A part of the dataset is the images we took. They are first filtered to remove the samples with similar content and single targets. High-quality images with many targets, rich categories, and complex backgrounds are retained. The preserved dataset contains 1332 images of non-motor drivers in different periods, roads, angles, and lighting conditions, with a resolution is 4032×3024 pixels. This dataset uses the LableImg software to manually label the images by category and calibrate the annotated coordinates. All the images are annotated into three categories: "non-motor," "helmet," and "head". Among them, the head area without a helmet, sun hat, baseball cap, and other standard hats are all defined as "head". Each image generates an annotation file in an EXtensible Markup Language (XML) format, which contains the file name of the image, the name of the target category, and the coordinates of the target annotation box.

The other part is the public dataset "Bike Helmets Detection" from the Kaggle website. This dataset contains 764 images of bicyclists wearing helmets and their annotated files in XML format, including two categories, "With helmet" and "Without helmet". The experiment modifies them to "helmet" and "head" to unify the label names.

3.2 Dataset Description

1757 images are selected as the training set, including 1070 from our shooting pictures and 687 pieces from "Bike Helmets Detection". In order to increase the diversity of the training set, photos are taken in multiple scenarios. For capturing non-motor vehicles in different traffic flow states, this dataset obtains images of non-motor drivers in various road conditions, including intersections, non-motor lanes and sidewalks. Acquired pictures are taken from multiple angles, including front, rear and side, which enrich the state types of non-motor vehicle riders. Images of multiple driver situations are also included in the dataset, including a single rider, two people riding together, and

three people riding together. Get images of multiple e-bike types to increase the diversity of non-motor vehicles, including shared bikes, standard household e-bikes and take-out electro mobiles. To restore the weather conditions in the genuine scene as much as possible, get pictures of various light intensities, including sunny days, cloudy days and tree shadows. These images of various wearing types are gained to improve the richness of small objects in the dataset, including helmets and ordinary hats. Get images with complex backgrounds, including scenes with high traffic density, similar target color and background color, and many occlusions. The training set samples are shown in Fig. 1.

Figure 1: Some examples of the training set in HWND

A validation set of 339 images is selected, including 262 self-photographed images and 77 images from "Bike Helmets Detection." The images selected for the validating set are helmets of various colors and shapes. There are also a variety of distractions, such as baseball caps, sun visors, and helmets placed in the basket. A rich and diverse validation set helps to evaluate the model comprehensively, and samples in the validation set for the experiment are shown in Fig. 2.

Ultimately, the integrated HWND contains a total of 2096 images with three category labels: "non-motor," "helmet," and "head." The number of each label in the HWND is counted, and the comparison of labels in the training and validation set is shown in Fig. 3.

Considering the effect of realistic complex road conditions, the training set is increased to 14056 using horizontal flip, random rotation, increasing noise, and changing image brightness and contrast. The expanded non-motor labels are 10720, helmets are 14472, and heads are 8360.

Figure 2: (a) Shows the helmets in different colors and styles; (b) displays situations in the absence of a helmet and the wearing of a regular hat

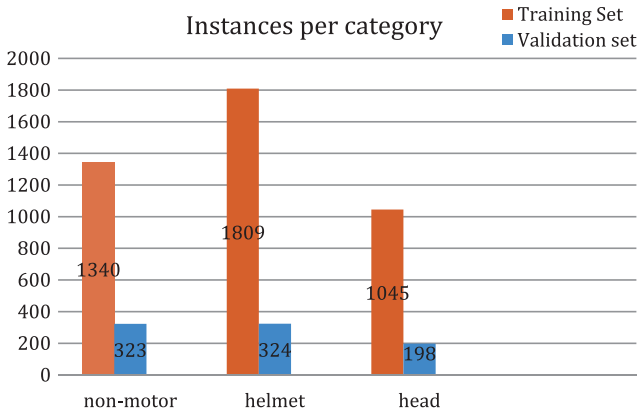


Figure 3: Comparison of the number of categories in HWND

4 Methodology

4.1 Dilated Convolution in Coordinate Attention (DICA)

In computer vision tasks, there is always information closely related to the study and some irrelevant information. However, the attention mechanism can help the network focus on analyzing vital information and ignore insignificant information. In recent years, attention mechanisms have been widely used for tasks such as semantic segmentation, image classification, and target detection [31,32,35–37], which have achieved remarkable works. However, most attention mechanisms have limited perceptual fields and do not easily capture contextual information at different scales. Therefore, this paper proposes the DICA mechanism, which combines the dilated convolution with the Coordinate Attention (CA) mechanism [38]. DICA block uses a multi-branch dilated convolution structure

with different dilation rates to obtain multi-scale features. The structure of the DICA mechanism is shown in Fig. 4.

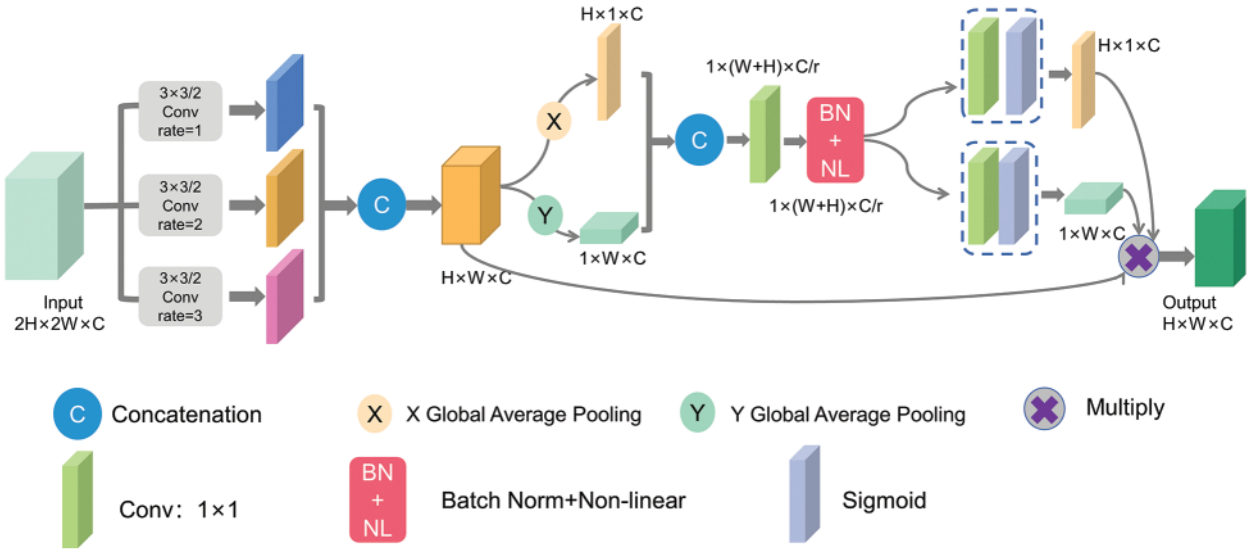


Figure 4: DICA block

First, the input feature map obtains a larger perceptual field by dilated convolution. In this paper, the dilated convolution structure is set as follows: convolution kernel size is 3×3, step size of 2, expansion rates are 1, 2 and 3, and the convolved perceptual fields are 3×3, 5×5, and 7×7. The feature map obtained after the convolution of the three branches has the same number of channels as the input feature map, but the size is reduced by 1/2.

$$T_i = DConv\,(P) \tag{1}$$

In Eq. (1), T_i represents the output feature map of the *ith* ($i = 1, 2, 3$) branch, $DConv()$ represents the dilated convolution operation, and P is the input feature map.

The feature map U is obtained by concating the feature maps obtained from the three branches, and then the number of channels is recovered by 1×1 convolution. As shown in Eq. (2), where $[\cdot, \cdot, \cdot]$ delegates the concat operation of the three feature maps T_1, T_2, and T_3 in the channel dimension, F_1 represents the 1×1 convolutional transform function.

$$U = F_1\,([T_1, T_2, T_3]) \tag{2}$$

After that, the feature map U is divided into horizontal and vertical directions. Each channel is encoded using pooling layers with kernel sizes $(H, 1)$ and $(1, W)$, respectively, to obtain the output of both directions in each channel. The calculation processes are shown in Eqs. (3) and (4).

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} u_c\,(h, i) \tag{3}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} u_c\,(j, w) \tag{4}$$

where H, W are the height and width of the feature map U, respectively. $z_c^h(h)$ denotes the output of the cth channel with height h, and $z_c^w(w)$ is the output of the cth channel with width w.

Then the feature maps z^h and z^w are concated, the dimensionality is reduced using a shared 1×1 convolution, batch normalized, and finally fed into the activation function to obtain the feature map f. In Eq. (5), δ represents the nonlinear activation function, f represents the intermediate feature map obtained by encoding spatial information in the horizontal and vertical directions, the range of f is set to $f \in R^{C/r\times(H+W)}$, and r is the reduction rate for controlling module size, which is set to 32 in the experiment.

$$f = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right) \tag{5}$$

Next, the feature map f is divided into two separate tensors $f^h \in R^{C/r\times H}$ and $f^w \in R^{C/r\times W}$. The number of channels is recovered from being consistent with the input feature map U using two 1×1 convolutional transform functions F_h and F_w, respectively. The attention weights g^h and g^w of the feature map in height and width are obtained after the Sigmoid activation function. Detailed expressions are shown in Eqs. (6) and (7), where σ denotes the Sigmoid activation function.

$$g^h = \sigma\left(F_h\left(f^h\right)\right) \tag{6}$$

$$g^w = \sigma\left(F_w\left(f^w\right)\right) \tag{7}$$

Finally, the output feature map Y of the DICA block is obtained by multiplicative weighting calculation on the original feature map, and the formula is shown below.

$$y_c(i,j) = u_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{8}$$

In this paper, the DICA module is added to the backbone and placed in front of the SPP module to replace the original CBL module. The DICA mechanism sets different dilation rates for multi-branch dilated convolution networks to expand the perceptual field from different scales. The convolved feature map acquires more high-resolution information in the shallow feature maps. After that, the channel information and the position information of the feature map are encoded simultaneously to obtain the attentional feature maps in both horizontal and vertical directions. Each element of the feature map visually reflects whether the target is present in the corresponding row and column.

4.2 Rebuild-Bidirectional Feature Pyramid Network (Re-BiFPN)

The dataset proposed in this paper contains numerous helmet and head labels, and these small targets have few pixel values and inconspicuous feature information. As the network's layers increases, too much convolution will reduce or even disappear small target features. Therefore, to improve the feature representation of small targets, the Re-BiFPN structure is proposed in this paper to fuse feature maps of different scales. Based on BiFPN [16], cross-scale connections are added to nodes with two inputs to ensure that each node contains at least three inputs, effectively reducing the loss of features. Fig. 5 shows the structure of our Re-BiFPN.

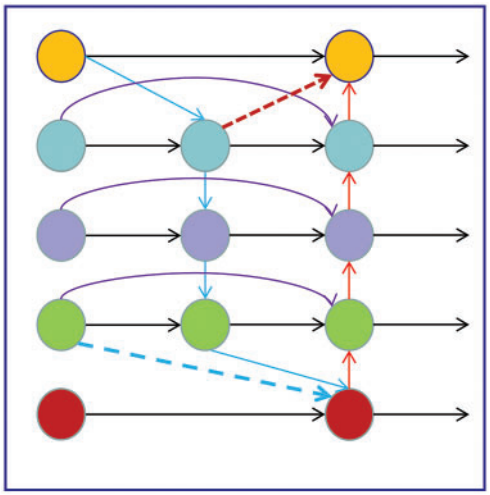


Figure 5: The structure of Re-BiFPN

In Fig. 5, the blue arrows represent top-down pathways, red arrows represent bottom-up pathways, purple arrows represent jump-connected pathways of nodes at the same level, and the same color nodes have the same size.

The Re-BiFPN proposed in this paper adds jump connections of high-level input nodes to the bottom nodes, as shown by the blue dashed arrows in Fig. 5, incorporating the rich semantic information of the high-level feature maps. The connection of lower-level nodes is added to the higher-level nodes, as shown by the red dashed arrows in Fig. 5. The rich contour and edge information of the low-level feature map are fused with the high-level feature map by taking advantage of the high resolution of the low-level feature map.

In the Bi-FPN, the original large target detection layer P_3^{out} is obtained by fusing the feature map P_3^{in} of 80 × 80 size from the same layer, and P_4^{td} obtained by convolving the feature map of 40 × 40 size from the lower layer. The output expression of P_4^{td} is shown in Eq. (9). The original small target detection layer P_5^{out} is obtained by fusing the feature map P_5^{in} of 20 × 20 size from the same layer, and the output feature map P_4^{out} of 40 × 40 size from the upper layer. Re-BiFPN adds a channel to the input P_4^{in} and P_4^{td}, respectively, as shown by the blue dashed line and the red dashed line in Fig. 6. The final outputs of P_3^{out} and P_5^{out} are shown in Eqs. (10) and (11).

$$P_4^{td} = Conv\left(\frac{w_1 \cdot P_4^{in} + w_2 \cdot Resize\left(P_5^{in}\right)}{w_1 + w_2 + \epsilon}\right) \tag{9}$$

$$P_3^{out} = Conv\left(\frac{w_1' \cdot P_3^{in} + w_2' \cdot Resize\left(P_4^{td}\right) + w_3' \cdot Resize\left(P_4^{in}\right)}{w_1' + w_2' + w_3' + \epsilon}\right) \tag{10}$$

$$P_5^{out} = Conv\left(\frac{w_1'' \cdot P_5^{in} + w_2'' \cdot Resize\left(P_4^{td}\right) + w_3'' \cdot Resize\left(P_4^{out}\right)}{w_1'' + w_2'' + w_3'' + \epsilon}\right) \tag{11}$$

In the formula, w_i represents the weights obtained from network training, and the ReLu activation function is used after each w_i to ensure that the weights $w_i \geq 0$, the values of output weights are controlled between 0 and 1 by regularization, and the learning rate ϵ is set to 0.0001 to avoid unstable values, P_4^{td} represents the intermediate feature layer of the fourth layer in the top-down pathway, *Conv*() represents the convolution operation, and Resize stands for up-sampling or down-sampling operation.

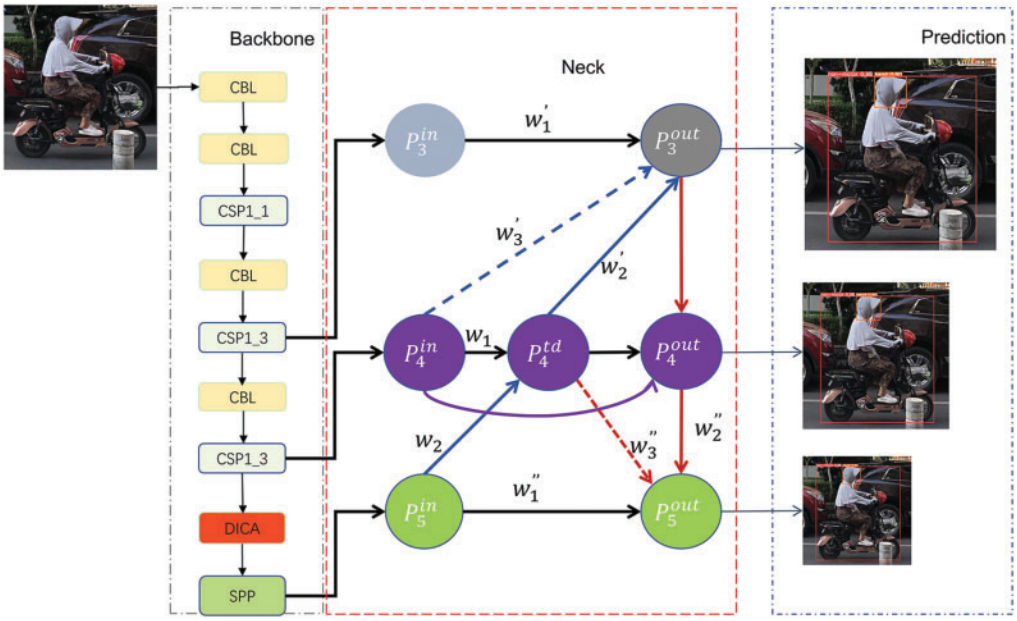


Figure 6: YOLOv5s-DR network structure

4.3 YOLOv5s-DR

To address the problems of poor robustness and low accuracy of small targets detection in the current helmet detection algorithms, we optimize the YOLOv5s network and propose the YOLOv5s-DR model with the structure in Fig. 6. This paper adds the DICA block to replace the original CBL module in front of the SPP module, and the DICA step size is set to 2. DICA combines the dilated convolution with the CA [38] mechanism to obtain more dense information by expanding the perceptual field. The attention mechanism ignores irrelevant information in the background and improves the network's focus on small targets. Re-BiFPN increases the cross-scale feature extraction layer and improves the feature fusion capability of the model.

5 Experimental Results and Analysis

5.1 Experimental Setup and Model Training

This experiment sets the momentum to 0.937, the weight decay rate to 0.005, and uses SGD with the initial learning rate of 0.001 to optimize the model. The experiment adopts the K-means algorithm to recalculate the anchors' values of the model before training. The optimal anchors' values obtained are (7,9), (13,18), (21,30), (28,39), (38,57), (52,86), (80,122), (138,197) and (230,309). Since the original YOLOv5s network structure is improved in this experiment, the official pre-trained weights are not fit. Therefore, this experiment uses the modified YOLOv5s-DR model to retrain the HWDN. Batchsize

is set to 32, and iterations are 200 epochs. Training platform parameters are detailed in Table 1, and the changes in training loss are shown in Fig. 7.

Table 1: Hardware and software platforms

CPU	Intel(R) Core(TM) i5-10600KF CPU @4.10 GHz
GPU	NVIDIA GeForce RTX 3080
Deep learning framework	PyTorch
Development of language	Python 3.8

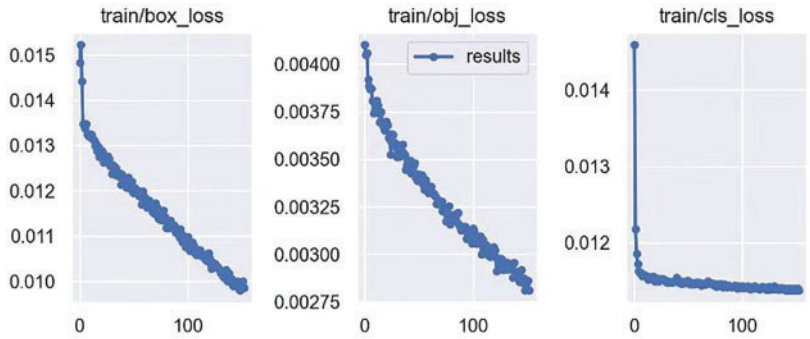


Figure 7: Training loss change curve

5.2 Ablation Experiments for YOLOv5s-DR

To explore the effects of different improvements on the model, the DICA module and Re-biFPN are used sequentially on the original YOLOv5s model to evaluate the rationality and effectiveness of the proposed method. Table 2 shows the performance of the ablation study.

Table 2: Each module ablation contrast experiment

Model	DICA	Re-BiFPN	Precision (%)	Recall (%)	mAP (%)
YOLOv5s	×	×	84.2	87.4	92
YOLOv5s-D	√	×	90.9	89.5	94
YOLOv5s-R	×	√	93.3	88.7	93.1
YOLOv5s-DR	√	√	93.5	91.1	94.3

For the analysis of Table 2, adding the DICA module results in a 2% improvement in mAP compared to the original YOLOv5s. Using Re-BiFPN for feature fusion, mAP is improved by 1.1%. Combining the DICA and Re-BiFPN, the mAP is improved by 2.3%, and Recall is improved by 3.7%, indicating that the model's comprehensiveness and correctness are improved. The accuracy of different target scales has improved when comparing different classes of targets, as shown in Fig. 8. Adding the DICA module brings 2.2% and 1.6% improvement for small targets such as helmets and heads. The experimental results show that the DICA module preserves the detailed information while increasing the perceptual field, effectively improving the network's feature extraction capability.

Integrating the spatial information of the feature map with the channel information on a larger scale makes it easier to capture the feature information of small targets and effectively improves the network's attention to small targets. The Re-biFPN module brings 0.9% and 1.2% improvement to helmets and heads, respectively. The experimental results show that the improved Re-BiFPN obtains richer contextual information, enhances the feature representation of small targets, and effectively improves the detection accuracy of helmets and heads.

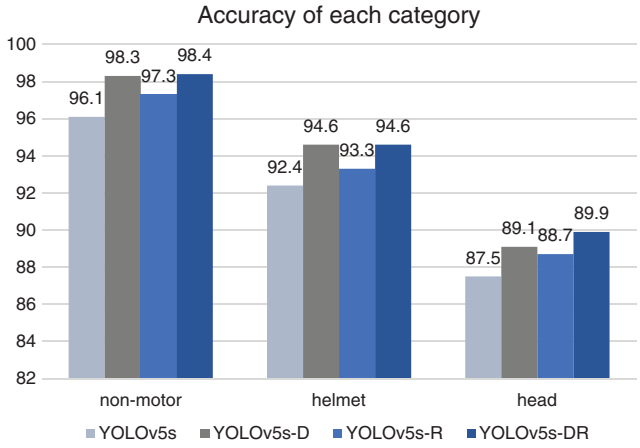
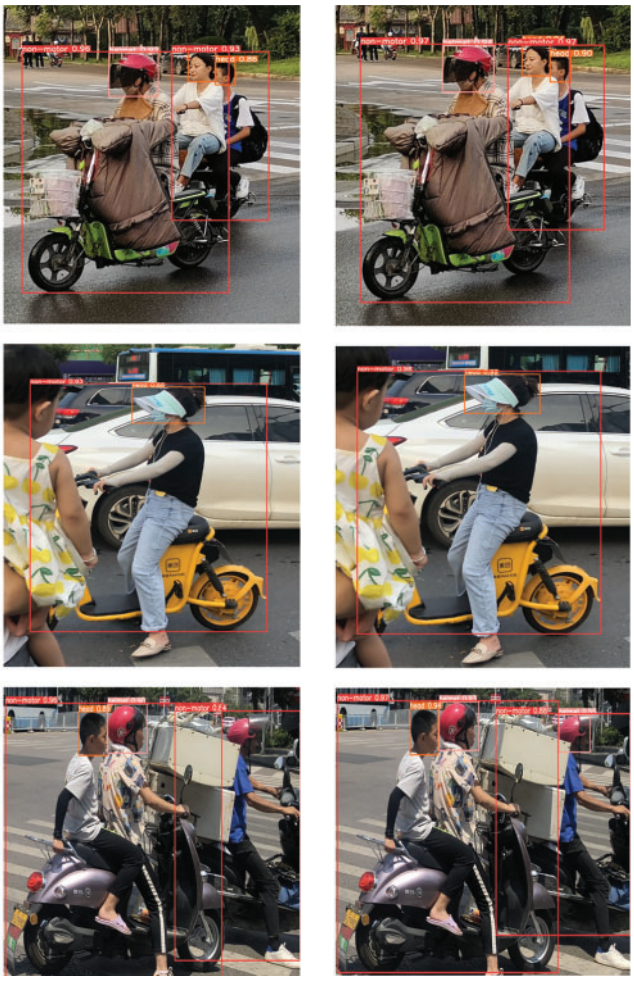


Figure 8: Comparison of the accuracy per category

The detection effects of YOLOv5s-DR and YOLOv5s models in the test set are compared, and the results are visualized. Fig. 9a is the detection effect of the YOLOv5s model, and Fig. 9b is the detection effect of the YOLOv5s-DR model. In Fig. 9a, the prediction probability of small targets such as helmets and heads reached 88% at the lowest level, and there is even a phenomenon that the detected object is lost. In Fig. 9b, the prediction probability of targets has been effectively improved, and small targets not detected by the YOLOv5s model are also correctly classified, especially in the case of helmet occlusion. In general, the YOLOv5s-DR model effectively improves the detection accuracy when targets are occluded, and the shapes of the helmet and the head are similar. The analysis shows that the addition of the two modules, DICA and Re-BiFPN, effectively reinforces the robustness of the model and can also achieve effective detection in complex situations.

5.3 Performance Comparison of Different Models

To further analyze the detection performance of the YOLOv5s-DR model, our dataset is tested on other target detection models, such as YOLOv5s, Faster-RCNN, YOLOv4 and YOLOv7 [39]. Comparing the experimental results and analyzing them, our model is 2.3% higher than YOLOv5s in mAP, 33.2% higher than Faster-RCNN, and higher than YOLOv4 by 26.3%. Compared with YOLOv7, although the mAP is reduced, YOLOv5s-DR is slightly better in detection speed. Detailed metrics comparison is shown in Table 3. Experiments show that the YOLOv5s-DR model proposed in this paper effectually improves the detection accuracy by 94.3%. In terms of speed, it can recognize 81 images per second.



(a) YOLOv5s (b) YOLOv5s-DR

Figure 9: Detection results of YOLOv5s and YOLOv5s-DR

Table 3: Performances of all comparison models

Model	Param (M)	Speed (fps)	mAP (%)
YOLOv5s	7.1	90	92
Faster-RCNN	108	3	61.1
YOLOv4	244	11	68
YOLOv7	6.3	71	96.05
YOLOv5s-DR (ours)	9.7	81	94.3

6 Conclusions and Future Work

This paper proposes a helmet-wearing detection dataset for non-motor drivers called HWND. It contains 1332 helmet-wearing images taken by us and 764 images from a public dataset, and we annotate the dataset in detail. The article also proposes a non-motor drivers helmet wearing detection model YOLOv5s-DR. It adds a DICA mechanism combining dilated convolution and attention mechanisms to YOLOv5, which expands the perceptual field and facilitates the network to extract high-level semantic features. Besides, it enhances the network's ability to model the dependencies between channels, extracts more precise information, and effectively improves the feature representation capability of the model. Meanwhile, Re-BiFPN is used for feature extraction to enhance the feature fusion capability of the network, and the feature map acquires richer contextual information while reducing the computational effort. It is tested on the HWND dataset to verify the proposed method's performance. Extensive experiments demonstrate that the proposed method improves the accuracy of helmet detection, providing an effective solution for helmet-wearing detection of non-motor drivers in practical scenarios. We plan to produce a larger dataset and study a more lightweight helmet detection model for future work.

Acknowledgement: The authors would like to thank the support of Central South University of Forestry & Technology and the support of Natural Science Foundation of Hunan Province.

Funding Statement: This research was funded by Natural Science Foundation of Hunan Province under Grant NO: 2021JJ31142, author F. J, http://kjt.hunan.gov.cn/.

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

References

[1] L. Zhang, C. Li and H. Sun, "Object detection/tracking toward underwater photographs by remotely operated vehicles (ROVs)," *Future Generation Computer Systems-the International Journal of Escience*, vol. 126, no. 1, pp. 163–168, 2022.

[2] B. Wang, Y. Zhao and C. L. P. Chen, "Hybrid transfer learning and broad learning system for wearing mask detection in the COVID-19 era," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.

[3] K. Akyol and B. Şen, "Automatic detection of COVID-19 with bidirectional LSTM network using deep features extracted from chest X-ray images," *Interdisciplinary Sciences-Computational Life Sciences*, vol. 14, pp. 89–100, 2022. https://doi.org/10.1007/s12539-021-00463-2

[4] A. Ashraf, M. Imran, A. M. Qahtani, A. Alsufyani, O. Almutiry *et al.,* "Weapons detection for security and video surveillance using CNN and YOLO-V5s," *Computers, Materials & Continua*, vol. 70, no. 2, pp. 2761–2775, 2022.

[5] T. Zhou, B. Xiao, Z. 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.

[6] Z. Lu, S. Liang, Q. Yang and B. Du, "Evolving block-based convolutional neural network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–21, 2022.

[7] 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 (CVPR)*, Columbus, USA, pp. 580–587, 2014.

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

[9] 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.
[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed *et al.,* "SSD: Single shot multibox detector," in *Proc. of the European Conf. on Computer Vision*, Cham, Switzerland, Springer, pp. 21–37, 2016.
[11] 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.
[12] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 6517–6525, 2017.
[13] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint, 2018. https://arxiv.org/abs/1804.02767
[14] A. Bochkovskiy, C. -Y. Wang and H. -Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint, 2020. https://arxiv.org/abs/2004.10934
[15] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li *et al.,* "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128837–128868, 2019.
[16] M. Tan, R. Pang and Q. V. Le, "EfficientDet: Scalable and efficient object detection," arXiv preprint, 2018. https://ui.adsabs.harvard.edu/abs/2019arXiv191109070T
[17] B. Yogameena, K. Menaka and S. S. Perumaal, "Deep learning-based helmet wear analysis of a motorcycle rider for intelligent surveillance system," *IET Intelligent Transport Systems*, vol. 13, no. 7, pp. 1190–1198, 2019.
[18] S. Chen, W. Tang, T. Ji, H. Zhu, Y. Ouyang *et al.,* "Detection of safety helmet wearing based on improved faster R-CNN," in *Int. Joint Conf. on Neural Networks (IJCNN)*, Glasgow, UK, pp. 1–7, 2020.
[19] Y. Gu, S. Xu, Y. Wang and L. Shi, "An advanced deep learning approach for safety helmet wearing detection," in *Int. Conf. on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, Atlanta, GA, USA, pp. 669–674, 2019.
[20] A. Shrivastava, A. Gupta and R. Girshick, "Training region-based object detectors with online hard example mining," arXiv preprint, 2016. https://ui.adsabs.harvard.edu/abs/2016arXiv160403540S
[21] P. Mohan, P. Narayan, L. Sharma and M. Anand, "Helmet detection using faster region-based convolutional neural networks and single-shot multibox detector," in *Int. Conf. on Smart Computing and Communications (ICSCC)*, Kochi, India, pp. 209–214, 2021.
[22] W. Chen, M. Liu, X. Zhou, J. Pan and H. Tan, "Safety helmet wearing detection in aerial images using improved yolov4," *Computers, Materials & Continua*, vol. 72, no. 2, pp. 3159–3174, 2022.
[23] W. Jia, S. Xu, Z. Liang, Y. Zhao, H. Min *et al.,* "Real-time automatic helmet detection of motorcyclists in urban traffic using improved YOLOv5 detector," *IET Image Processing*, vol. 15, no. 14, pp. 3623–3637, 2021.
[24] G. Han, M. Zhu, X. Zhao and H. Gao, "Method based on the cross-layer attention mechanism and multiscale perception for safety helmet-wearing detection," *Computers and Electrical Engineering*, vol. 95, no. C, pp. 13, 2021.
[25] R. Cheng, X. He, Z. Zheng and Z. Wang, "Multi-scale safety helmet detection based on SAS-YOLOv3-Tiny," *Applied Sciences*, vol. 11, no. 8, pp. 3652, 2021.
[26] Z. Daquan, Q. Hou, Y. Chen, J. Feng and S. Yan, "Rethinking bottleneck structure for efficient mobile network design," arXiv preprint, 2020. https://ui.adsabs.harvard.edu/abs/2020arXiv200702269D
[27] L. Zeng, X. Duan, Y. Pan and M. Deng, "Research on the algorithm of helmet-wearing detection based on the optimized yolov4," *The Visual Computer*, pp. 1–11, 2022. https://doi.org/10.1007/s00371-022-02471-9
[28] C. -Y. Wang, H. -Y. M. Liao, I. H. Yeh, Y. -H. Wu, P. -Y. Chen *et al.,* "CSPNet: A new backbone that can enhance learning capability of CNN," arXiv preprint, 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv191111929W

[29] M. Sadiq, S. Masood and O. Pal, "FD-YOLOv5: A fuzzy image enhancement based robust object detection model for safety helmet detection," *International Journal of Fuzzy Systems*, vol. 24, no. 5, pp. 2600–2616, 2022.
[30] L. Wang, Y. Cao, S. Wang, X. Song, S. Zhang *et al.,* "Investigation into recognition algorithm of helmet violation based on YOLOv5-CBAM-DCN," *IEEE Access*, vol. 10, pp. 60622–60632, 2022.
[31] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang *et al.,* "Deformable convolutional networks," arXiv preprint, 2017. https://ui.adsabs.harvard.edu/abs/2017arXiv170306211D
[32] S. Woo, J. Park, J. -Y. Lee and I. S. Kweon, "CBAM: Convolutional block attention module," arXiv preprint, 2018. https://ui.adsabs.harvard.edu/abs/2018arXiv180706521W
[33] Z. Li, W. Xie, L. Zhang, S. Lu, L. Xie *et al.,* "Toward efficient safety helmet detection based on YoloV5 with hierarchical positive sample selection and box density filtering," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022.
[34] "Bikes helmets dataset," *Make ML*. https://makeml.app/datasets/helmets
[35] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, "Squeeze-and-excitation networks," arXiv preprint, 2017. https://ui.adsabs.harvard.edu/abs/2017arXiv170901507H
[36] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao *et al.,* "Dual attention network for scene segmentation," arXiv preprint, 2018. https://ui.adsabs.harvard.edu/abs/2018arXiv180902983F
[37] C. Ding, Y. Chen, R. Li, D. Wen, X. Xie *et al.,* "Integrating hybrid pyramid feature fusion and coordinate attention for effective small sample hyperspectral image classification," *Remote Sensing*, vol. 14, no. 10, pp. 2355, 2022.
[38] Q. Hou, D. Zhou and J. Feng, "Coordinate attention for efficient mobile network design," arXiv preprint, 2021. https://ui.adsabs.harvard.edu/abs/2021arXiv210302907H
[39] C. -Y. Wang, A. Bochkovskiy and H. -Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," arXiv preprint, 2022. https://ui.adsabs.harvard.edu/abs/2022arXiv220702696W