Computers, Materials & Continua
DOI: 10.32604/cmc.2022.026664
Article

Tech Science Press

Safety Helmet Wearing Detection in Aerial Images Using Improved YOLOv4

Wei Chen[1], Mi Liu[1,*], Xuhong Zhou[2], Jiandong Pan[3] and Haozhi Tan[4]

[1]College of Civil Engineering, Changsha University of Science and Technology, Changsha, 410114, China
[2]College of Civil Engineering, Chongqing University, Chongqing, 730000, China
[3]College of Mechanical and Electrical Engineering, Hunan Agricultural University, Changsha, 410128, China
[4]Department of Civil and Environmental Engineering, University of Auckland, Auckland, 1010, New Zealand
*Corresponding Author: Mi Liu. Email: liumicslgdx@163.com
Received: 31 December 2021; Accepted: 09 February 2022

Abstract: In construction, it is important to check whether workers wear safety helmets in real time. We proposed using an unmanned aerial vehicle (UAV) to monitor construction workers in real time. As the small target of aerial photography poses challenges to safety-helmet-wearing detection, we proposed an improved YOLOv4 model to detect the helmet-wearing condition in aerial photography: (1) By increasing the dimension of the effective feature layer of the backbone network, the model's receptive field is reduced, and the utilization rate of fine-grained features is improved. (2) By introducing the cross stage partial (CSP) structure into path aggregation network (PANet), the calculation amount of the model is reduced, and the aggregation efficiency of effective features at different scales is improved. (3) The complexity of the YOLOv4 model is reduced by introducing group convolution and the pruning PANet multi-scale detection mode for de-redundancy. Experimental results show that the improved YOLOv4 model achieved the highest performance in the UAV helmet detection task, that the mean average precision (mAP) increased from 83.67% of the original YOLOv4 model to 91.03%, and that the parameter amount of the model is reduced by 24.7%. The results prove that the improved YOLOv4 model can effectively respond to the requirements of real-time detection of helmet wearing by UAV aerial photography.

Keywords: Safety-helmet-wearing detection; unmanned aerial vehicle (UAV); YOLOv4

1 Introduction

Civil engineering is one of the oldest and most engaged industries, and its construction of engineering facilities reflects the development of the social economy, culture, science, and technology in

Each historical period. However, safety accidents often occur during the construction phase of civil engineering projects. According to statistics, one-fifth of the deaths of American workers in 2019 occurred in the construction industry, and 1,061 construction workers died. In 2018, 734 safety

accidents occurred in the construction and municipal sectors in China, resulting in 840 deaths, with casualties of construction workers having a major impact on the sustainable development of civil engineering.

The head is the most important organ of the human body, and as head armor, a safety helmet can effectively protect the head from severe impacts [1,2]. Although the safety protection provisions uphold wearing a safety helmet as a rule applicable to construction site staff, due to the insufficient safety awareness of some workers, there are still situations where safety helmets are not worn, resulting in many injuries and deaths in engineering accidents. Therefore, inspecting whether workers on the construction site wear safety helmets is of great significance to reducing casualties. At present, the detection method of helmet wearing is mainly manual inspection, and its degree of automation is low. Moreover, owing to the large area of the construction site, the complex environment, and the fatigue of inspectors, it is difficult to truly reflect the wearing of helmets of onsite workers. As a result, experts have conducted many studies on automated safety helmet detection methods.

In 2012, AlexNet [3] overcame the problem of traditional machine learning, which only contains shallow learning skills, and won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Afterward, deep learning became a research hotspot. Inspired by AlexNet, ZFNet [4], VGG [5], GoogleNet [6], ResNet [7], and other networks have been proposed to refresh the error rate on the ImageNet datasets to 3.6%, which is lower than the human error rate of 5.1% on the ImageNet datasets. Furthermore, the application of deep learning in various fields has also witnessed great development.

Many researchers have introduced the target detection algorithm in deep learning into safety-helmet detection. They have typically divided the workers on the construction site into two categories, those who wear helmets and those who do not wear helmets, so as to realize automatic detection of helmet wearing. Guo et al. [8] presented the Faster R-CNN algorithm based on a VGG feature extraction network to determine whether a worker is wearing a helmet. Chen et al. [9] adopted the k-means ++ algorithm to re-cluster small targets in safety helmets, and then used the improved Faster R-CNN algorithm to judge whether to wear a safety helmet. As a representative work of the two-stage target detection algorithm, the Faster R-CNN [10] algorithm needs to obtain an anchor box before classifying and locating the target. The detection accuracy is high, but the inference speed is slow, making it difficult to achieve real-time monitoring. Zhou et al. [11] presented the YOLOv3 algorithm, which integrates an attention mechanism and thereby solves the problem of small targets and easy occlusion of a safety helmet. Wang et al. [12] proposed an improved YOLOv3 algorithm to meet the requirements of real-time safety-helmet monitoring. Li et al. [13] proposed the SSD-Mobilenet algorithm to improve the accuracy of safety-helmet detection. The YOLOv3 [14] model and SSD [15] model used mentioned above are one-stage target detection algorithms, which can directly classify and locate targets at one time. Although the detection speed is fast and can meet the requirements of real-time monitoring, the error detection rate of a safety helmet facing a small target and a complex construction environment is higher. In addition, all these studies and other research on the safety-helmet detection algorithm of datasets obtained by the crawler and video site are applied to the video monitoring system. Compared with manual inspection, it can better save manpower, but the surveillance cameras are generally arranged in the entrances and exits and the material stacking site, so the real construction area makes it difficult to install video surveillance or carry out a full range of real-time monitoring of the workers under construction.

As a highly intelligent equipment in the new era, Unmanned aerial vehicle (UAV) has the characteristics of strong mobility, a low cost, and high data accuracy. It has been widely used in agriculture [16], medical treatment [17], communications [18], transportation [19], geographic mapping

[20], and emergency rescue [21]. We proposed using UAV aerial photography and the YOLOv4 [22] algorithm to monitor the conditions of workers wearing safety helmets on construction sites in real time. Aerial photography by UAV can cover the whole construction site and achieve a full range of real-time monitoring of staff wearing safety hats. In addition, it can track workers in real time and check whether workers enter dangerous sites or have non-standard dangerous construction behaviors, which is conducive to the automatic management of personnel. The YOLOv4 algorithm has obvious advantages in aerial shooting of small targets and real-time monitoring, mainly owing to the CSPDarknet53 backbone and spatial feature pyramid in the algorithm. However, in practical application, the YOLOv4 model has the following problems:

(1) Compared with other safety helmet datasets, the aerial helmet image is small. Owing to the small number of pixels, it is easy to cause the loss of helmet characteristic information in the process of downsampling.
(2) The construction site is complex, and there are many pieces of equipment, so the safety helmet is easily blocked by objects.
(3) Aerial images are greatly affected by angle and illumination.

If the YOLOv4 algorithm is used directly, the above problems easily occur. Therefore, according to the characteristics of the UAV aerial safety helmet dataset, we improved the backbone network and neck network of YOLOv4, and the improved model adopts group convolution to reduce model parameters. An enhanced YOLOv4 model is proposed to improve the detection accuracy of safety-helmet wearing under the background of a construction site without affecting the detection speed.

The rest of this paper is organized as follows. Section 2 introduces the related work, including the UAV aerial safety helmet dataset, YOLOv4, group convolution, and path aggregation network (PANet). Section 3 introduces the method proposed in this paper. Section 4 introduces the details of the experiment, including the experimental dataset, experimental environment, and evaluation indicators. Section 5 introduces the relevant experiments and a discussion of the experimental results. Section 6 summarizes the work of this paper.

2 Related Work

2.1 Scale Distribution of UAV Aerial Safety Helmet Dataset

The problem of safety-helmet-wearing detection has existed for many years. Different emerging technologies have had a significant impact on the public helmet datasets and have obtained good detection results. Han et al. [23] used the cross-layer attention mechanism to further refine the feature information of the target region to improve the SSD detection accuracy. Wu et al. [24] introduced the dense connection idea of dense convolutional network (DenseNet) into the YOLOv3 backbone network to improve the model detection accuracy. However, there is a great difference between the target scale distribution in the traditional public datasets and the UAV aerial helmet dataset used in the work of this paper. Therefore, we obtained statistics on the ratio of the average area of the detected object to the average area of the whole image in this work's UAV aerial safety helmet dataset and several groups of common helmet datasets.

As shown in Fig. 1, the proportion of the average target area and the average area of the whole image is only 0.0004 in the UAV aerial safety helmet dataset. Unlike in previous work, the target scale of most of our dataset is a small scale distribution. Therefore, based on YOLOv4, we adjusted the sensitivity of the model to targets of different scales to make the model more focused on small

target detection and thereby improve the detection accuracy of the YOLOv4 model in the UAV helmet detection task.

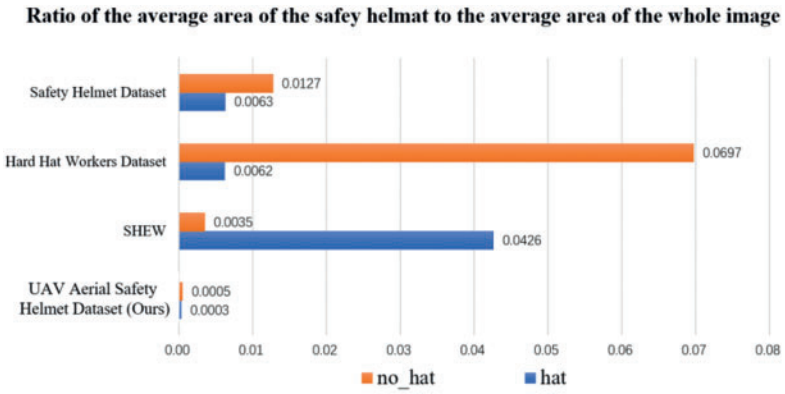


Figure 1: Ratio of the average area of the safety helmet to the average area of the whole image

2.2 YOLOv4

According to the requirement of real-time detection of inference speed, the one-stage detection algorithm was selected as the basic algorithm in this work. SDD [16] and You only look once (YOLO)-series [14,22,25,26] algorithms are representative works of one-stage detection algorithms. The SDD algorithm does not reuse high-resolution low-level information, resulting in insufficient low-level feature information and poor detection effect on small targets. The YOLOv4 [22] algorithm integrates the main characteristics of YOLOv1 [25], YOLOv2 [26], YOLOv3 [14], and other YOLO models, and has good performance in inference speed and detection accuracy. Compared with the predecessor YOLOv3, the average precision (AP) and frame per second (FPS) tested on the Microsoft COCO (MS COCO) dataset were improved by 10% and 12% respectively. Therefore, we chose YOLOv4 as the basic algorithm, whose network structure is mainly composed of three parts: Backbone, Neck and Head, as shown in Fig. 2.

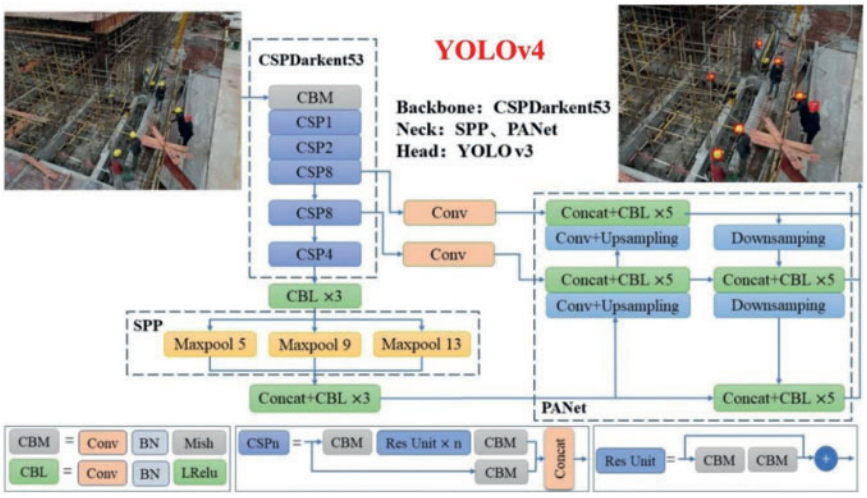


Figure 2: Safety-helmet detection based on YOLO v4

Backbone, compared with the YOLOv3 algorithm, utilizes CSPDarknet53 as the backbone network of feature extraction. Moreover, YOLOv4 combinate the cross stage partial (CSP) [27] structure and residual [28] network, the problem of gradient information repetition in the Darkent53 network is eliminated, and more advanced feature information is output. The feature extraction capability of the backbone network is strengthened while reducing model parameters. Neck is mainly composed of spatial pyramid pooling (SPP) [29] and path aggregation network (PANet) [30]. SPP adopts three pooling layers of different scales: 5×5, 9×9, and 13×13. After the max pooling of input features, it can greatly increase the receptive field, eliminate the influence of inconsistencies in input scales, and produce output of a fixed length. PANet was adopted as the feature fusion structure of the model, and the bottom-up path was added on the basis of the feature pyramid networks (FPN) [31] of YOLOv3 to improve the feature reuse capability of the model. Finally, the target was detected by the detection head of YOLOv3, and the feature map of three scales was output: 13×13, 26×26, and 52×52.

The main modules are as follows:

(1) CBM, which includes the convolution layer, batch normalization layer, and Mish activation function.
(2) CBL, which includes the convolution layer, batch normalization layer, and LeakyRelu activation function. Unlike the CBM module, the LeakyRelu activation function has less computation than the Mish function, but it is not as nonlinear as Mish functions.
(3) CPS is the main feature extraction structure of YOLOv4, which divides the input features into two parts. One feature is extracted by stacking multiple residual structures, while the other input feature, as the large residual edge, is stacked with the output feature of the residual block. This way, the model can simplify the occupation of video memory and improve the feature extraction ability of the model.

The original YOLOv4 backbone network has strong feature extraction ability, and the multi-scale target detection ability was improved through the idea of multi-scale detection. However, in the application of UAV to safety-helmet detection, helmet target features are small and dense, and the detection accuracy of the original YOLOv4 model is poor. Therefore, we need a model that pays more attention to small target detection to enhance the ability of UAV to detect helmets.

2.3 Group Convolution

Group convolution first appeared in AlexNet [3] and was carried forward in ResNeXt [32]. It groups the input channel and the convolution kernel channel, and uses local correlation of channels to reduce model complexity.

As shown in Fig. 3, the size of the input feature map is $H_{in} \times W_{in} \times D_{in}$, the size of the convolution kernel is $h_1 \times w_1 \times D_{in}$, the quantity of the convolution kernel is D_{out}, and the size of the output feature map is $H_{out} \times W_{out} \times D_{out}$. If the general convolution operation is adopted, the number of parameters is:$h_1 \times w_1 \times D_{in} \times D_{out}$.

If group convolution operation is adopted, the input feature map is divided into "g" parts according to the channels. The size of each group of input feature map is $H_{in} \times W_{in} \times D_{in}/g$, the convolution kernel is $h_1 \times w_1 \times D_{in}/g$, and the size of each output feature map is $H_{out} \times W_{out} \times D_{out}/g$. Finally, each group of output feature graphs is spliced according to the corresponding channel position, and the final output feature graph is $H_{out} \times W_{out} \times D_{out}$. The number of parameters is:$h_1 \times w_1 \times D_{in} \times D_{out} \times \frac{1}{g}$.

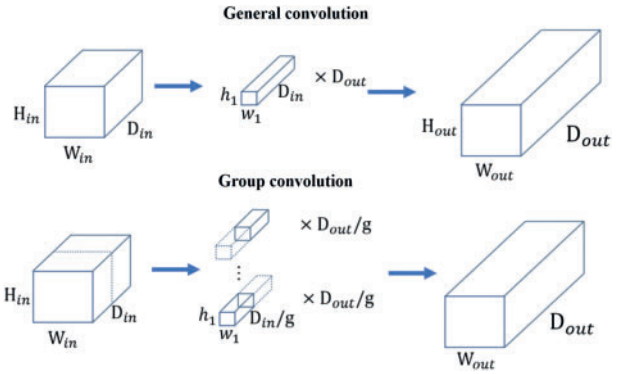


Figure 3: The comparison between general convolution and group convolution

According to the calculation, the parameter number of group convolution is 1/g of general convolution, which effectively reduces the model complexity. Compared with depthwise separable convolution, group convolution is concise, and modularization is suitable for de-redundant operation of obese models. Considering that different effective feature layers in the YOLOv4 backbone network have different feature utilization rates for small targets, we introduced group convolution into the model backbone network to ensure the accuracy of the model and reduce the complexity of the backbone network.

2.4 PANet

Among current target detection algorithms, the FPN [31] is a common module in the neck network and aggregates feature layers of different scales through upsampling to generate feature maps with stronger expression ability. However, PANet [30] in the YOLOv4 model adopts multi-path aggregation, which not only performs top-down feature aggregation but also adds bottom-up secondary fusion through downsampling, thus enriching the semantics of the effective feature layer. However, in UAV detection tasks, the effective weights of the deep effective feature layer and shallow effective feature layer are not the same, so we proposed a more concise PANet, which preserves the deep semantic information while enhancing the semantic information of shallow features.

3 Methods

In the work of this paper, YOLOv4 was improved based on the target scale distribution characteristics of the UVA aerial safety helmet dataset.

(1) For the UAV aerial safety helmet dataset, small and medium-sized targets occupy the main distribution. By adjusting the effective feature receptive field of the backbone network, more fine-grained effective feature maps are retained. The main modification is to increase the number of shallow feature layers; remove the CSP4 structure of the last layer in backbone; set the dimensions of effective feature layers as 26×26, 52×52, and 104×104; reduce the receptive field of effective feature layers; and retain more small target feature information.

(2) In order to reduce the model complexity, we replaced the general convolution in the residual structure in the effective feature layer with a large receptive field with group convolution. The CSP8 of the last two layers in backbone was replaced with G-CSP8 to reduce the model complexity without affecting the model progress.

(3) In order to improve the model reasoning speed, we adjusted the number of detection heads in PANet, retained the deep feature information fusion, and eliminated the influence of the deep detection heads on the model reasoning speed.

3.1 Improvement of Backbone

In this study, owing to the dominant role of small targets in the UAV aerial safety helmet dataset, the original YOLOv4 model performs multi-scale detection of targets, and its detection scale cannot fully adapt to the UAV dataset. In order to solve the problem of matching detection target and detection head scale, we changed the number of backbone network layers of the original model; adjusted the structure of CSP1, CSP2, CSP8, CSP8, and CSP4 in the original CSPDarkne53 to CSP2, CSP4, CSP8, and CSP8; and trimmed the effective feature layer of 13x13. The deep effective feature layer with too large a receptive field and a lack of fine-grained semantic information was abandoned. The C3 layer (104×104) with a smaller receptive field and more fine-grained feature information was added as the effective feature layer, and the effective feature layer was adjusted to 26×26, 52×52, and 104×104. In this way, the scale matching between the detection head and the detection model can be guaranteed, the utilization rate of effective features in the backbone network can be improved, and the detection effect of the model can be enhanced.

3.2 Improvement of the De-Redundancy

Although adding a shallow effective feature layer can improve the utilization of effective features in the backbone network, the finer-grained detection layer also has larger model parameters, resulting in an increase in the number of model parameters. For UAV detection tasks, we only need to enhance the feature utilization of small target, and the huge YOLOv4 model is too redundant for UAV detection tasks. In order to ensure the detection accuracy of the model, the improved model retains the shallow effective feature layer and only carries out de-redundancy for the deep effective feature layer. We reduced the number of parameters of deep effective features by group convolution, and replaced the residual block in the CSP8 module of the original backbone network with group convolution to obtain G-CSP8, whose grouping number g is 32. These changes can effectively reduce the redundancy of the model and improve the calculation speed of the model. The improved backbone structure is shown in Fig. 4.

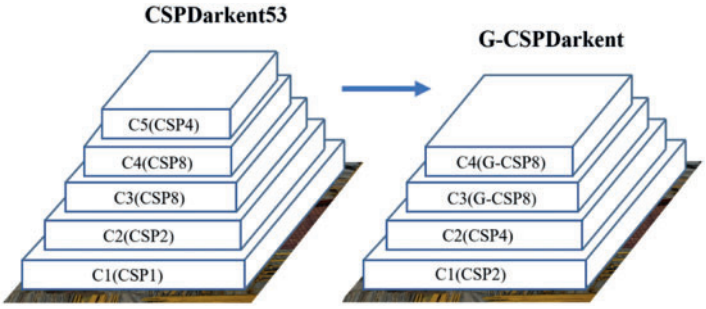


Figure 4: The improved backbone structure

3.3 Improvement of PANet

Adding group convolution to the backbone network can effectively adjust the parameters of the backbone network, but the phenomenon of increasing the computation amount of the neck network

caused by adding a large dimensional shallow effective feature layer is not alleviated. To solve this problem, we made the improvements shown in Fig. 5. Specifically, the CSP structure was added into PANet to alleviate the problem of repeated gradient information in the neck network and reduce the calculation amount of model. We also fully considered the matching problem of the target scale and model detection scale in the UAV aerial safety helmet dataset. Our approach was to retain the characteristics of the deep layer effective features information (C4) and shallow layer feature for feature fusion effectively. Moreover, trimming the detection head of the deep effective feature layer (C4) and only the effective feature layer of C3 and C2 layer is used for detection, which further improves the model reasoning speed and alleviates the redundancy problem caused by adjusting the backbone network to the model.

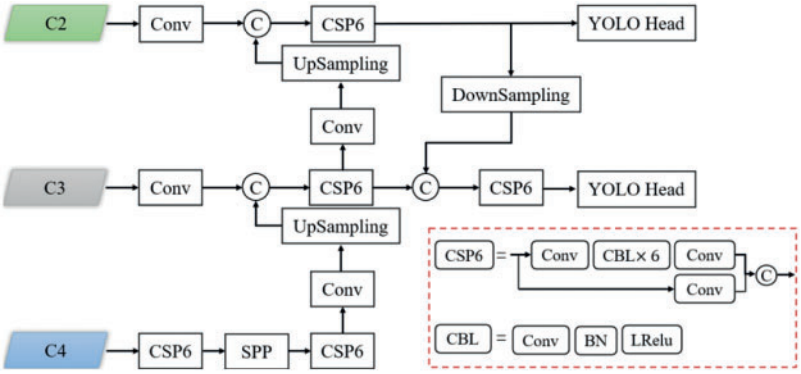


Figure 5: Improved PANet structure

3.4 The Overall Structure Design of Improved YOLOv4 Algorithm

For the UAV aerial safety helmet dataset, we made the above improvements to the YOLOv4 model. The overall structure of the algorithm is shown in Fig. 6. Compared with the original model, the main improvements are as follows: We adjusted the dimension and number of effective feature layers in the backbone network, reduced the model receptive field, and improved the utilization rate of small target features in the backbone network. The specific modification was to cut the 13x13 effective feature layer and add a 104x104 effective feature layer. The structure of backbone network CSP1, CSP2, CSP8, CSP8, and CSP4 was adjusted to CSP2, CSP4, CSP8, and CSP8. At the same time, group convolution was added to alleviate the redundant phenomenon of the backbone network caused by increasing the dimension of effective feature layer. The local correlation adjustment model of the channel was used to adjust the calculation parameters. In the backbone network, group convolution was mainly added in the convolution layer of the large receptive field; that is, the CPS8 structure of C3 and C4 layers was adjusted to G-CSP1_8. In the neck network, the general convolution in the original model was added to the CSP structure to improve the performance of feature aggregation and reduce model parameters. At the same time, only the feature fusion part of the C4 effective feature layer was retained, and the detection head part of the C4 layer was trimmed to improve the overall reasoning speed of the model.

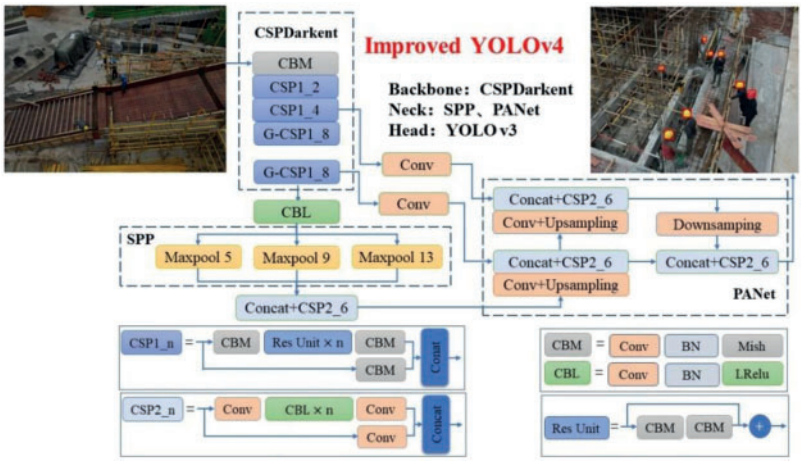


Figure 6: Safety-helmet detection based on the improved YOLOv4

4 Experimental Design

4.1 Datasets

4.1.1 Dataset Acquisition

The experimental research area where the work of this paper was conducted is located in the engineering project department of the student engineering training center of the Changsha University of Science and Technology. In April 2021, the safety-helmet image was obtained by DJI Mavic 2 UAV. The effective pixel of UAV was 20 million, the equivalent focal length of the camera lens was 28 mm, the maximum image size was 5472 × 3648, and the flight speed was 2 m/s. There were three batteries, each lasting about 30 min, and the imaging angle was between −45° and −90°. During the flight, the vertical distance between the UAV and workers was 8–15 m, and 3,399 images of construction site personnel were collected.

We obtained the helmet dataset according to the VOC2007 standard, the main work was as follows:

(1) The objects were divided into two categories: those who wore a helmet and those who did not wear a helmet, and the labels were defined as hat and no_hat, respectively. Labeling was used to mark, and the corresponding labeling file was obtained, which contained the name, size, targets, and basic information of each target of the corresponding picture.
(2) We split the training dataset and test dataset in an 8:2 ratio. A total of 3,399 target images were collected, including 2,719 training datasets and 680 test datasets.
(3) Generate the training dataset and test dataset corresponding to the index file in the .txt format.

4.1.2 Data Augmentation

Data augmentation is a common technique enhancement method in the absence of data. In the work of this paper, random rotation (the random factor was 0.5), random scaling (the resizing range was 0.25–2), color gamut adjustment (the hue range was −0.1–0.1, the saturation range was 1–1.5, and the lightness range was 1–1.5, the color gamut adjustment was controlled by the three parameters mentioned above) were used, and prior knowledge was introduced to prevent overfitting of the model.

4.2 Experiment Environment

The proposed model is an improvement based on YOLOv4, and its experimental environment is shown in Tab. 1.

Table 1: Experimental environment

Configuration	Parameter
CPU	AMD 5600X
GPU	Nvidia GeForce RTX 2080TI
Accelerated environment	CUDA10.0 CUDNN7.5.0
Development environment	Pycharm2020.1.3 Pytorch1.2.0
Operating system	Windows 10

The improved YOLOv4 model adopts SGD training. For the training, the input image size was 416×416, the batch size was 8, the momentum was 0.9, and the weight decay was 0.001. the initial learning rate was 0.0001, which dynamically dropped to 0.00001 during training. After 400 epochs, the model reached the fitting state.

4.3 Evaluation Index

We evaluated all models based on the following five indicators: (1) precision, (2) recall, (3) mean average precision (mAP), (4) parameter number, and (5) FPS. The specific formulas for the first two are as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

where TP represents the number of positive cases correctly divided, FP represents the number of positive cases incorrectly classified, and FN represents the number of negative cases incorrectly classified.

5 Results and Analysis

To obtain the results and discussion, we carried out an independent comparative test analysis from three aspects, the backbone network, de-redundant operation, and feature fusion improvement, to reflect the effectiveness of the improved module.

5.1 Performance Comparison of the Improved Backbone

In the original YOLOv4 model, the detection of the UAV aerial safety helmet dataset often suffers from omission and misdetection. The main reason is that the backbone network's insufficient feature utilization of small targets and the detection scale cannot match well with the scale distribution of the small target detection task. Considering the scale distribution characteristics of the UAV aerial

safety helmet dataset, we gave up the coarse-grained deep effective feature layer and added the fine-grained shallow effective feature layer. In addition to reducing the computational load of the model, the utilization rate of the backbone network for small target features is improved this way. In order to verify the effectiveness of the improved backbone network, we conducted ablation experiments on the improved backbone network with CSPDarknet53, ResNeXt50, DenseNet, and ResNeSt50. The experimental results are shown in Tab. 2.

Table 2: Comparison experiment of the backbone network

Method	Backbone	Input size	FPS	mAP (%)
YOLOv4	CSPDarknet53	416 × 416	47.1	83.67
YOLOv4	CSPDarknet53	608 × 608	40.5	87.58
YOLOv4	ResNeXt50	416 × 416	52.6	81.34
YOLOv4	DenseNet	416 × 416	71.7	77.51
YOLOv4	ResNeSt50	416 × 416	41.3	87.61
YOLOv4	Ours	416 × 416	39.2	89.30

In the same input dimension, the improved backbone network improved the mAP in the test dataset by 5.63% over CSPDarknet53, 7.96% over ResNeXt50, 11.79% over DenseNet, and 1.69% over ResNeSt50. For the larger input scale (608 × 608) of CSPDarknet53, the improved backbone network detection performance also improved by 1.72%. The experimental results indicate that the model detection performance of the improved backbone network is better than that of the classical backbone network. Thus, adding a more fine-grained effective feature layer can effectively enhance the utilization rate of small target features, improve the detection performance of the model for small target data tasks, and reduce the error detection of small targets and dense targets in UAV detection.

5.2 Ablation Experiment of Improved Module

As the scale distribution of the UAV dataset is relatively simple, the addition of a new shallow effective feature layer will inevitably lead to network enlargement and increase calculation too much. In order to balance the influence of backbone network improvement on the number of model parameters, considering the different proportions of the fine-grained semantic information in effective feature layers at different scales, we introduced group convolution into the last two layers of the backbone network to reduce the complexity of the backbone network. At the same time, we trimmed the PANet part by considering the distribution of the UAV aerial safety helmet dataset. we explored the impact of backbone network improvement, group convolution, and PANet improvement on network detection performance, The experimental results are shown in Tab. 3 and Fig. 7, where BNI represents backbone network improvement, GC represents group convolution, and PI represents PANet improvement.

After the addition of the improved backbone network, the mAP value of the model increased from 83.67% to 89.30%. Experimental results show that the improved backbone network could effectively strengthen the model to propose fine-grained features. The main reason is to add a large dimension of the effective feature layer and improve the utilization rate of fine-grained features of the model. However, with the increase of effective feature layer, the number of model parameters increased by about 29%, and the inference speed increased from 21.2 to 26.8 ms. Therefore, de-redundant modules were introduced in the next experiment to reduce the number of model parameters. According to the experiment, the number of model parameters added with group convolution decreased from 82.5

to 59.7 M, and the inference speed decreased from 26.8 to 25.2 ms. Moreover, we introduced the improved PANet module, and the model mAP increased from 89.67% to 91.03%, the number of model parameters decreased from 59.7 to 48.1 M, and the inference speed decreased from 25.2 to 23.2 ms. The experimental results indicate that the improved YOLOv4 model can effectively improve the model's detection efficiency of small targets in the UAV aerial safety helmet dataset.

Table 3: Ablation experiment of the improved module

Algorithms	BNI	GC	PI	Input size	FPS	mAP (%)	Param (M)
YOLOv4				416 × 416	47.1	83.67	63.9
YOLOv4	✓			416 × 416	39.2	89.30	82.5
YOLOv4	✓	✓		416 × 416	39.5	89.67	59.7
YOLOv4	✓	✓	✓	416 × 416	42.6	91.03	48.1

Notes: BNI: Backbone network improvements, GC: Group convolution, PI: PANet improvements.

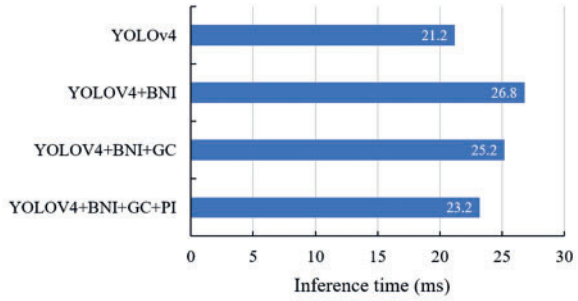


Figure 7: Contrast graph of inference time

5.3 Comparative Experiment of Different Models

Under the premise of better detection accuracy, the improved YOLOv4 model alleviated the influence of the improved backbone network on the model complexity and effectively improved the inference speed. Figs. 8a and 8b show the PR curves of hat and no_hat of improved YOLOv4 and YOLOv4, YOLOv3, respectively. In the UAV helmet-detection task, the PR curves of target hat and no_hat detected by the improved YOLOv4 model were higher than those of other models, which indicates that the performance of the improved YOLOv4 model is higher than that of other models.

Fig. 9 shows visualization test results before and after improvement, where (a), (c), (e), (g) represent the detection results of YOLOv4, (b), (d), (f), and (h) represent the detection results of the improved YOLOv4, the red box represents those who wore a safety helmet, and the blue box represents those who did not wear safety helmet. In addition, (a) and (b) represent detection under normal conditions, (c) and (d) represent detection under shielding conditions, (e) and (f) represent remote detection, and (g) and (h) represent detection under intense illumination. As can be seen from the figure: for (a) and (b), the YOLOv4 model not only missed detection but also mistakenly detected idle safety helmets. However, the improved YOLOv4 model could effectively reduce the missed detection and accurately distinguish idle safety helmets. For (c), (d), (e) and (f), the YOLOv4

model had insufficient detection ability for the obscured target, while the improved YOLOv4 model could accurately identify the obscured helmet; and in (e) and (f), the improved YOLOv4 model could identify the helmet with few pixels at a distance. For (g) and (h), the YOLOv4 model was affected by light, resulting in a large number of missed and false detections, while the improved YOLOv4 model still maintained a high detection accuracy under strong light. The results indicate that the improved YOLOv4 can effectively deal with the task of helmet detection by UAV aerial photography.

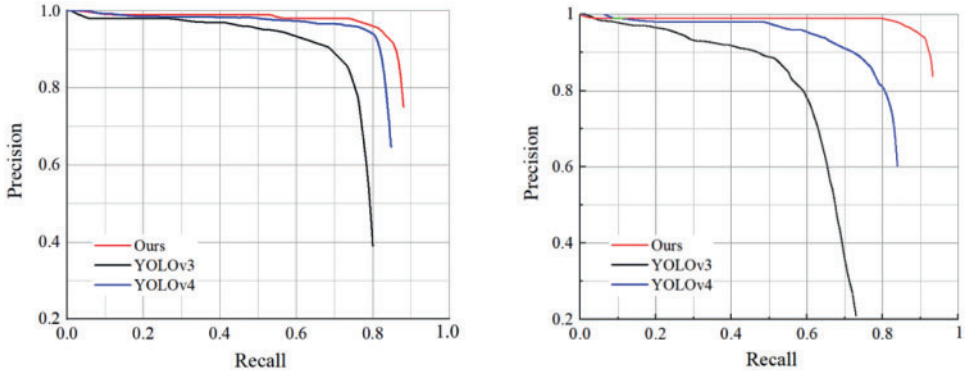


Figure 8: Contrast graph of PR curves. (a) PR curves of hat of different models (b) PR curves of no_hat of different models

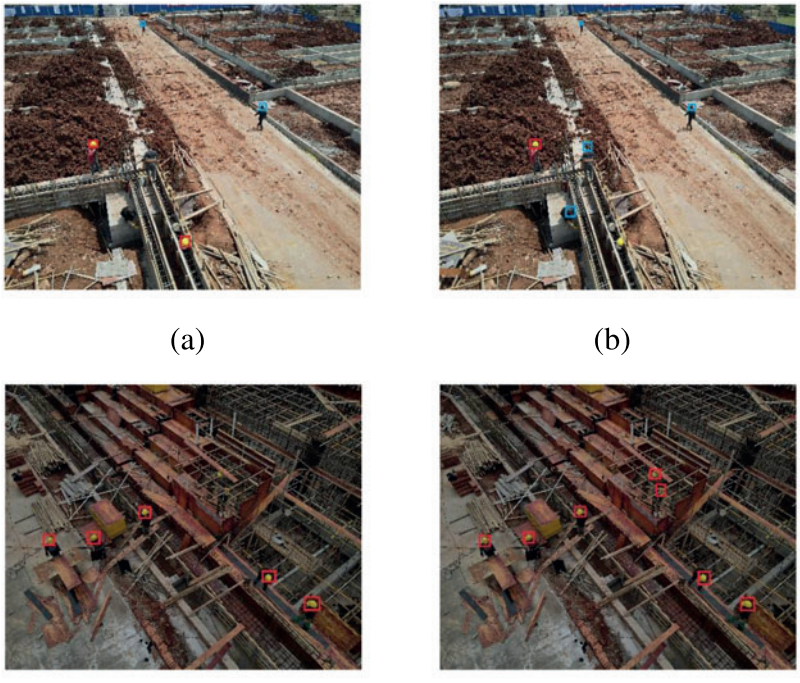
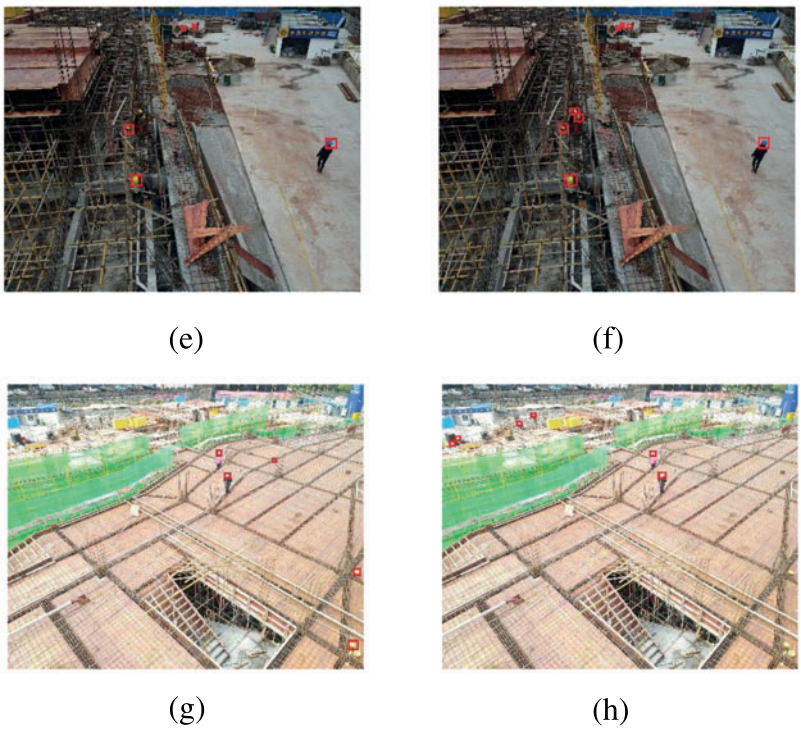


Figure 9: (Continued)



(e) (f)

(g) (h)

Figure 9: Visualization test results before and after improvement (a), (c), (e), (g) represent the detection results of YOLOv4, (b), (d), (f), (h) represent the detection results of the improved YOLOv4

6 Conclusion

In order to overcome the challenge of small target detection in monitoring helmet wearing by UAV, an improved YOLOv4 model was proposed to check helmet wearing under the condition of UAV aerial photography. We made the model more sensitive to small targets by improving the backbone network. To ensure the simplicity of the model, the neck network was pruned, and the de-redundant operations such as group convolution were introduced. Experimental results show that the detection accuracy of the improved YOLOv4 network was improved, and the mAP increased from 83.67% to 91.03%. The results prove that our improved strategy can effectively retain small target feature information and improve the utilization rate of small target effective feature. The number of parameters of the improved network model was less than 24.7% of the original model, which ensures the inference speed of the model. Therefore, the improved YOLOv4 model can effectively cope with the problem of the small size of the target detected by UAV aerial photography. It can also complete real-time monitoring of the wearing of safety helmets at construction sites based on a UAV to ensure the effective implementation of safety specifications in the construction process and reduce the casualty rate.

Acknowledgement: The author would like to thank the support of Changsha University of Science and 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 51408063, author W. C, http://www.nsfc.gov.cn/; in part by the Outstanding Youth

Scholars of the Department of Hunan Provincial under Grant 20B031, author W. C, http://kxjsc.gov.hnedu.cn/.

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

References

[1] A. H. M. Rubaiyat, T. T. Toma, M. Kalantari-Khandani, S. A. Rahman, L. Chen *et al.,* "Automatic detection of helmet uses for construction safety," in *2016 IEEE/WIC/ACM Int. Conf. on Web Intelligence Workshops (WIW)*, Omaha, NE, USA, pp. 135–142, 2016.

[2] B. L. Suderman, R. W. Hoover, R. P. Ching and I. S. Scher, "The effect of hardhats on head and neck response to vertical impacts from large construction objects," *Accident Analysis & Prevention*, vol. 73, pp. 116–124, 2014.

[3] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[4] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conf. on Computer Vision*, Zurich, Switzerland, pp. 818–833, 2014.

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

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed *et al.,* "Going deeper with convolutions," in *2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 1–9, 2015.

[7] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778, 2016.

[8] S. Guo, D. Li, Z. Wang and X. Zhou, "Safety helmet detection method based on faster R-CNN," in *Int. Conf. on Artificial Intelligence and Security*, Singapore, Springer, pp. 423–434, 2020.

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

[10] 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.

[11] Q. Zhou, J. Qin, X. Xiang, Y. Tan and N. N. Xiong, "Algorithm of helmet wearing detection based on AT-YOLO deep mode," *Computers, Materials & Continua*, vol. 69, no. 1, pp. 159–174, 2021.

[12] H. Wang, Z. Hu, Y. Guo, Z. Yang, F. Zhou *et al.,* "A Real-time safety helmet wearing detection approach based on CSYOLOv3," *Applied Sciences*, vol. 10, no. 19, pp. 6732, 2020.

[13] Y. Li, H. Wei, Z. Han, J. Huang and W. Wang, "Deep learning-based safety helmet detection in engineering management based on convolutional neural networks," *Advances in Civil Engineering*, vol. 2020, pp. 1–10, 2020.

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

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

[16] G. Wang, Y. Han, X. Li, J. Andaloro, P. Chen *et al.,* "Field evaluation of spray drift and environmental impact using an agricultural unmanned aerial vehicle (UAV) sprayer," *Science of the Total Environment*, vol. 737, pp. 139793, 2020.

[17] Y. Xue, B. Onzo, R. F. Mansour and S. Su, "Deep convolutional neural network approach for covid-19 detection," *Computer Systems Science and Engineering*, vol. 42, no. 1, pp. 201–211, 2022.

[18] A. Noorwali, M. A. Javed and M. Z. Khan, "Efficient uav communications: Recent trends and challenges," *Computers, Materials & Continua*, vol. 67, no. 1, pp. 463–476, 2021.

[19] W. Sun, L. Dai, X. R. Zhang, P. S. Chang and X. Z. He. "RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring," *Applied Intelligence*, vol. 51, pp. 1–6, 2021. [Online]. Available: https://dx.doi.org/10.1007/s10489-021-02893-3.

[20] F. Nex and F. Remondino, "UAV for 3D mapping applications: A review," *Applied Geomatics*, vol. 6, no. 1, pp. 1–15, 2013.

[21] S. Goudarzi, S. A. Soleymani, M. H. Anisi, D. Ciuonzo, N. Kama *et al.,* "Real-time and intelligent flood forecasting using uav-assisted wireless sensor network," *Computers, Materials & Continua*, vol. 70, no. 1, pp. 715–738, 2022.

[22] 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.

[23] 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 & Electrical Engineering*, vol. 95, pp. 107458, 2021.

[24] 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 (ICNSC)*, Canada, pp. 363–368, 2019.

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

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

[27] C. Wang, H. Liao, Y. Wu, P. Chen, J. Hsieh *et al.,* "CSPNet: A new backbone that can enhance learning capability of CNN," in *2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, pp. 390–391, 2020.

[28] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, USA, pp. 770–778, 2016.

[29] K. He, X. Zhang, S. Ren and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[30] K. Wang, J. H. Liew, Y. Zou, D. Zhou and J. Feng, "PANet: Few-shot image semantic segmentation with prototype alignment," in *2019 IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Seoul, Korea, pp. 9197–9206, 2019.

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

[32] S. Xie, R. Girshick, P. Dollar, Z. Tu and K. He, "Aggregated residual transformations for deep neural networks," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, USA, pp. 1492–1500, 2017.