Article

A Real-Time Safety Helmet Wearing Detection Approach Based on CSYOLOv3

Haikuan Wang ¹, Zhaoyan Hu ^{1,2,*}, Yuanjun Guo ^{2,*}, Zhile Yang ², Feixiang Zhou ³ and Peng Xu ⁴

¹ School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China; hkwang@shu.edu.cn (H.W.); zy.hu3@siat.ac.cn (Z.H.)

² Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; zl.yang@siat.ac.cn

³ School of Informatics, University of Leicester, Leicester LE1 7RH, UK; fz64@leicester.ac.uk

⁴ Haier Digital Technology (Shanghai) Co., Ltd., Shanghai 201600, China; xupeng.cosmo@haier.com

* Correspondence: yj.guo@siat.ac.cn

Received: 18 August 2020; Accepted: 18 September 2020; Published: 25 September 2020

Abstract: In the practical scenario of construction sites with extremely complicated working environment and numerous personnel, it is challenging to detect safety helmet wearing (SHW) in real time on the premise of ensuring high precision performance. In this paper, a novel SHW detection model on the basis of improved YOLOv3 (named CSYOLOv3) is presented to heighten the capability of target detection on the construction site. Firstly, the backbone network of darknet53 is improved by applying the cross stage partial network (CSPNet), which reduces the calculation cost and improves the training speed. Secondly, the spatial pyramid pooling (SPP) structure is employed in the YOLOv3 model, and the multi-scale prediction network is improved by combining the top-down and bottom-up feature fusion strategies to realize the feature enhancement. Finally, the safety helmet wearing detection dataset containing 10,000 images is established using the construction site cameras, and the manual annotation is required for the model training. Experimental data and contrastive curves demonstrate that, compared with YOLOv3, the novel method can largely ameliorate mAP by 28% and speed is improved by 6 fps.

Keywords: YOLOv3; cross stage partial network (CSPNet); safety helmet wearing (SHW) detection; spatial pyramid pooling (SPP); feature fusion

1. Introduction

With the expeditious evolvement of the construction industry in recent years, engineering construction projects can be seen everywhere across a city. The environment of a construction site is becoming more complicated and risky than ever before, and accidents happen frequently under some severe construction environments. Wearing a safety helmet when entering the construction site is a necessary protection measure for everyone, especially the workers who are more likely to be injured during the working process. However, with the increased difficulty of personnel management by humans, it is easy to cause safety accidents to the workers who are not wearing helmets or have non-standard operations.

With more and more attention attached to site safety, many research works have been done in the area of object detection using video surveillance with regard to large-scale construction places [1,2]. For example, Jie Li et al. [3] put forward an ingenious and pragmatic safety helmet wearing (SHW) detection modus on the basis of image preprocessing and machine learning. The accuracy of detection for the newly raised method is up to 80.7% and the frame rate is 7 fps. Dikshant Manocha et al. [4]

presented a helmet detection method for two-wheeler riders with the assistance of machine learning and a user interface was provided to pay challans. This technology firstly captures the real-time image of road traffic and detect the two wheelers among all vehicles on the road, secondly it processed to recognize whether riders are wearing helmet or not. Rattapoom Waranusast et al. [5] presented a system which is able to identify autobike riders and check if they were wearing safety helmets automatically at the same time. Furthermore, K-Nearest Neighbor(KNN) is adopted to make the classification between moving autocycles and other background objects on the road according to the features extracted from district attributes. The experimental data reveal that the average detection precision for various categories of lanes are all above 60% respectively.

While the above-mentioned methods meet the requirements in terms of accuracy, they are not suitable for the application of real-time detection. Besides, these methods used for SHW detection are all in the light of traditional machine learning methods. The features used for safety helmet detection are artificially selected and designed, and the obtained feature are not robust enough. Consequently, target detection on the basis of deep learning is developing rapidly and applied extensively in many fields. In 2014, the fast Region-CNN(RCNN) algorithm based on Visual Geometry Group Network16(VGG16) [6] is practically 9 times as fast as RCNN in respect of training speed, about 3 times as fast as that of SPP-net, while test speed is 213 times faster than RCNN, 10 times as fast as SPP-net. The mAP on Visual Object Classes 2012(VOC2012) is about 66%. Faster RCNN brought forward a fresh model architecture of RPN so that the full image convolutional features can be utilized together with the detection model, which has a frame rate of 5 fps on a GPU and achieves the detection accuracy of 73.25% mAP on PASCAL VOC2007.

Currently, with the rapid development of artificial intelligence, target detection and recognition algorithm based on deep learning has attracted much attention for researchers both at home and abroad [7–14]. For instance, Wei Liu et al. [15] presented a fresh method named SSD for target detection in pictures with applying a single deep neural network of ECCV2016. The multistage feature maps are employed as the basis of classification and regression, achieving the multi-scale effect. Ross Girshick et al. [16] raised a novel detection network model called R-CNN which is simple and scalable, achieving a mAP of 53.3% in 2014. Neural network migration learning method is proposed in this paper when lacking of labeled data, adopting neural network trained in other large datasets, and then fine tune in small-scale specific datasets. In [17], the CNN network needs certain size image input, so that the certain length image representation was generated after that any size image passed through pooling layer in SPP-net, which improves the speed of R-CNN for detection above 24 times. In [18], a separate CNN model is utilized in You Only Look Once(YOLO) algorithm to obtain end-to-end target detection function. The initial image is firstly normalized to the size of 448×448 , and then it is transmitted to CNN network for feature extraction. In the end, the network prediction consequence are dealt with to detect object categories. According to the comparison with R-CNN model, YOLO is a centralized framework with faster image processing speed with the training procedure is the format of end-to-end meanwhile. Joseph Redmon et al. [19] proposed YOLOv3 algorithm, which has no pooling layer and full connection layer in the entire YOLOv3 model. In the procedure of forward propagation, the size alteration of tensor is implemented by transforming the stride of convolution kernel. It has an obvious architecture and excellent real-time property. Two methods are proposed in [20] to obtain better helmet detection performance, utilizing haar trait and circle hough transform for face detection separately. Madhuchhanda Dasgupta et al. [21] presented an architecture for helmet detection of Motorcyclists on the moving autocycles. In the presented method, YOLOv3 model was employed for the detection of motorcycle riders at first stage, while an algorithm structure on the basis of CNN was brought forward for SHW detection of Motorcyclists. Rohith et al. [22] intended to establish an intelligent system to determine whether a cyclist is wearing a helmet, which was applied as the basis of the law enforcement to impose a fine on the offender. Yang Bo et al. [23] adopted YOLOv3 model to fine-tune the datasets for electric power construction scene, with accuracy of 90 %. To summarize, the previous methods are effective for target detection with low speed requirement and with high

accuracy in an ideal environment. However, the accuracy and speed cannot meet the requirements in the actual engineering environment. In this paper, YOLOv3 is improved to handle various situations due to its strong robustness in terms of real-time detection performance.

In this paper, an innovative method based on CSYOLOv3 was put forward for SHW detection of video surveillance on construction site. With the purpose of evaluating the feasibility and stability of the newly raised method, various visual conditions of construction sites were considered for experiment. The overall structure of the paper is as follows. Section 2 presents the overall network structure of the original YOLOv3. In Section 3, the novel algorithm of CSYOLOv3 is illustrated in detail. Section 4 presents the experimental data and curves. Finally, the conclusion of the raised method will be expressed in Section 5.

2. Original YOLOv3 Model

2.1. The Structure of YOLOv3

YOLOv3 is a single-stage target detection model proposed by Redmon et al. [19] in 2018. The network structure of the algorithm is shown in Figure 1. It combines excellent methods such as residual network, feature pyramid and multi-feature fusion network, with high performance of detection speed and accuracy. The first highlight of YOLOv3 is that the novel darknet53 network is applied as the backbone feature extraction, which draws on the experience of residual network in RESNET [13] and has excellent effect. The residual structure of darknet53 includes four steps: First, a 3×3 convolution with stride 2 is recorded as feature layer X . Afterwards, one 1×1 convolution is executed to compress the number of channels to half of the original, after which a 3×3 convolution is performed to enhance feature extraction and expanded to the number of channels to obtain $F(x)$. Finally, X and $F(x)$ are stacked by residual structure. One of the biggest advantages is that the model is capable of enhancing the accuracy by reinforcing its network depth. Meanwhile, the inner residual block employs jump connection, which also alleviates the gradient disappearance puzzle raised by adding its network depth.

Figure 1. Original YOLOv3 network structure.

Each convolution layer of DarkNet53 employs a unique structure (darknetconv2d). L2 regularization is used at each convolution operation, batch normalization and leaky ReLU activation function are performed after convolution being executed. Compared with the ordinary ReLU function that sets all negative values to zero, the Leaky ReLU activation function used in DarkNet53 allocates all negative values a non-zero slope, as shown in Equation (1).

$$f(x)_j = \begin{cases} x_j & \geq 0 \\ \frac{x_j}{a_j} & < 0 \end{cases} \quad (1)$$

The second highlight of YOLOv3 is the use of multi-scale features for prediction, extracting three different feature layers from DarkNet53 with shape size of (52, 52, 256), (26, 26, 512) and (13, 13, 1024). The three feature layers are convoluted for five times. One part of the processed results is utilized to output the prediction corresponding to the feature layer, and the other part is used to fuse with the previous feature layer after de-convolution operation. In addition, the way YOLOv3 predicts the coordinate position of the bounding box inherits the method in YOLOv2 [24], and *k*-means clustering is utilized to generate three categories of prior boxes with different sizes. Finally, each prediction border will generate four values, namely, the coordinate position of the upper left corner, along with the width and height of the border. Compared with other target detection models, YOLOv3 has the superiority of swift detection speed and high accuracy, but it still has some disadvantages when directly applied to the SHW detection task in complex scenes. Firstly, although the multi-scale prediction network is adopted in YOLOv3 to make full use of the receptive field, effectively alleviating the lack of scale invariance of convolution neural network, it also increases the amount of calculation, which leads to higher demands of hardware facilities; Secondly, although YOLOv3 improves the detection accuracy of small targets, it has the problem of insufficient shallow feature extraction. Thirdly, YOLOv3 still shows some detection deficiencies in the problems of existing occlusion, dense crowd and small scale targets in complex scenario. Therefore, the CSYOLOv3 model is put forward in this paper aiming to address the above problems.

2.2. Loss Function

The loss function of YOLOv3 is comprised of the prediction frame(*x,y*), the prediction frame size(*w,h*), the prediction class(*C*), and the prediction confidence(*confi*), as shown in Equation (2).

$$\text{Loss} = \frac{1}{n} \sum_{i=1}^n \text{loss}_{xy} + \frac{1}{n} \sum_{i=1}^n \text{loss}_{wh} + \frac{1}{n} \sum_{i=1}^n \text{loss}_C + \frac{1}{n} \sum_{i=1}^n \text{loss}_{confi} \quad (2)$$

where *n* is the total number of trained targets, and the components of loss function are described in detail as follows:

$$\text{loss}_{xy} = shw * (2 - w * h) * bst(\text{true}_{xy}, \text{pred}_{xy}) \quad (3)$$

$$\text{loss}_{wh} = shw * (2 - w * h) * 0.5 * S(\text{true}_{wh}, \text{pred}_{wh}) \quad (4)$$

$$\text{loss}_C = shw * bst(\text{true}_C, \text{pred}_C) \quad (5)$$

$$\text{loss}_{confi} = shw * bst(shw, \text{predshw}) + (1 - shw) * bst(shw, \text{predshw}) * \text{ignoreshw} \quad (6)$$

where *shw* indicates the point of the detected targets, and (*w,h*) represents the size of length and breadth about the forecast frame, *bst* denotes the binary cross entropy function, *S* means the function of variance. *true_{xy}* is the concrete target coordinate position on each image while *pred_{xy}* indicates the forecast position; *true_{wh}* is the actual groundtruth box size, *pred_{wh}* means the prediction box size; *true_C* refers to the realistic target class, *pred_C* represents the prediction class; *predshw* is the forecast object point, *ignoreshw* is associated with the value of IOU, if the value of IOU is less than the threshold, *ignoreshw* is 0.

3. Cross Stage YOLOv3 (CSYOLOv3) Model

Original YOLOv3 model has achieved good performance of target detection in common dataset such as COCO2007 and COCO2012. However, for a specific dataset of helmet, some improvements should be made to YOLOv3 to satisfy the requirements of realistic site scene. Improvements of original YOLOv3 include: The backbone network of DarkNet53 and the feature enhancement network.

3.1. Improved Backbone

So as to further reinforce the feature extraction network of YOLOv3, the cross stage partial network (CSPNet) is introduced in this paper. CSPNet is a fresh backbone network proposed by Wang et al. [25] that can be used to heighten the learning capability of CNN. Its main advantages are: Firstly, it can enhance the learning ability of CNN. Secondly, it can eliminate the algorithm structure with high computational power consumption; thirdly, it can reduce the memory cost. This paper applies the CSPNet structure to the DarkNet53 network, and then the CSPDarknet53 network is constructed. The network structures of DarkNet53 and CSPDarknet53 are shown in Figure 2.

Figure 2. The network structure of DarkNet53 and CSPDarkNet53.

Through the comparison with the original DarkNet53 network in Figure 2a, the network structure of CSPDarknet53 [26] in Figure 2b is not complicated. The main difference is that the original residual block is divided into Shortconv part and Mainconv part. In fact, the Shortconv part is to generate a large residual edge, which is directly connected to the last part of the structure after convolution processing. As the main part of the novel backbone, Mainconv continues to stack n times of residual blocks, specifically, the number of channels in feature maps is adjusted by a 1×1 convolution, and then the feature extraction is enhanced by a 3×3 convolution, after which the above output and the small residual edges are stacked together. Afterwards, the channel number is adjusted to the same as the Shortconv part through a 1×1 convolution. Finally, the Shortconv and the Mainconv part are stacked in the CSPDarkNet53, where the value of n is 1, 2, 8, 8, 4. Meanwhile, the activation function employed in convolution block is optimized in this paper, in which the activation function is changed from

LeakyReLU to Mish. Concretely, the convolution block is changed from DarknetConv2D_BN_Leaky to DarknetConv2d_BN_Mish.

Mish is a novel self-regularized nonmonotone activation function of neural network proposed by diganta Misra [27], with several characteristics such as no upper bound, with lower bound, smooth and nonmonotonic. Among them, “no upper bound” effectively avoids the problem of gradient vanishing, and “with lower bound” makes the network regularization strengthened. The results show that “smooth” can extract more advanced potential features to obtain better generalization ability, while “nonmonotone” can retain smaller negative input, which improves the interpretability and gradient flow of the neural network. The Mish activation function is shown in Equation (7).

$$Mish = x \times \tanh(\ln(1 + e^x)) \quad (7)$$

After the improvement of CSPDarknet53 network, the feature enhancement network module is introduced in this paper to further strengthen the network feature representation.

3.2. Feature Enhancement Network

Spatial pyramid pooling (SPP) is a model raised by He Kaiming et al. [25] to address the problem with regard to the input images with different sizes of the neural networks. Its main idea is to splice feature maps of arbitrary size into a fixed length feature vector through multi-scale pooling operation. Different from the purpose of he Kaiming et al., the SPP structure is introduced in this paper in order to further obtain multi-scale local feature information. Thus it is fused with the global feature information to obtain more abundant feature representation, thereby improving the prediction accuracy. Due to a series operations of convolution and down-sampling, the global semantic information of CSPDarknet53 is very rich. Therefore, in order to obtain more local features, the SPP structure is added into the convolution of the last feature layer in CSPDarknet53, which is shown in Figure 3.

It can be seen from Figure 3 that the specific steps of the improved SPP network structure are as follows: Firstly, the network structure of 13×13 feature layer is convoluted for three times, and then three different scale pooling layers are used for maximum pooling, of which the pooling kernel size is 13×13 , 9×9 and 5×5 , while the stride is 1. Finally, the input global feature maps and the three local feature maps after pooling are stacked, then three times convolution followed. The use of SPP structure can greatly strengthen the receptive field of the last feature layer, separate the most significant context features to obtain more abundant local feature information.

Considering that the details and location information of low-level feature layer are generally rich, however, with the deepening of feature layers, the detail information is decreasing, while the semantic information is increasing. Thus, the higher the feature layer is, the richer the contained semantic information will be. After the SPP structure is added into the CSPDarknet53, the multi-scale prediction network is improved by combining the feature fusion strategy, from which feature representation is enhanced through top-down and bottom-up fusion strategies [28] to further realize feature reuse. The structure of the improved multi-scale prediction network is expressed in Figure 4.

As can be seen from Figure 4, the specific improvements of the multi-scale prediction structure in this paper are: Firstly, three effective feature layers are extracted from the CSPDarknet53 backbone, which are respectively recorded as Large Feature Layer (LFL), Medium Feature Layer (MFL) and Small Feature Layer (SFL); Secondly, SFL0 is then convoluted three times, and the spatial pyramid pooling is performed so as to obtain SFL1, MFL1 is obtained by fusing the results of SFL1 that performing once convolution and upsampling with that of MFL0, and then LFL1 is obtained by fusing the results of MFL1 that executing once convolution and up-sampling with the result of LFL0 that only executing once convolution, so as to complete the feature fusion from bottom to top. Afterwards, FLF2 is achieved by five times convolution for LFL1, in order to get MFL2 by fusing the result of LFL2 that performing once downsampling with the result of MFL1 that executing five convolutions. Then, MFL2 is further downsampled, of which the result is fused with SFL1 to get SFL2, thus completing the feature fusion

from top to bottom. Finally, LFL2, MFL2 and SFL2 are obtained by fusing the three initial effective feature layers LFL0, MFL0 and SFL0 by bottom-up and top-down fusion, which are input into YOLO Head for prediction after five convolutions. Among them, conv_1 represents a convolution with the size of 1×1 .

Figure 3. Improved spatial pyramid pooling (SPP) network structure.

Figure 4. Improved multi-scale prediction network.

It can be seen from Figure 4 that the multi-scale prediction network is improved by using the top-down and bottom-up feature fusion strategies. While the computational complexity is added to a certain extent, the prediction accuracy has been significantly improved. Therefore, in general, this paper is meaningful for the improvement of multi-scale prediction network. After improving

the backbone network, feature enhancement network and multi-scale prediction network, the overall structure of YOLOv3 is displayed in Figure 5.

Figure 5. The CSYOLOv3 model structure.

4. Experiment

4.1. Dataset, Training Method and Environment

The dataset adopted in the part of experiments is self-built based on the practical construction project, and the format of the dataset is COCO. Videos were taken by the surveillance cameras from a bird eye's view, which may include massive redundant information such as lots of frames without people or with similar content. Accordingly, the videos were collected from ten cameras located in ten different positions and one picture was taken from every ten frames in the videos at 60 frame rate, so as to effectively avoid the problem of data redundancy. Meanwhile, it is essential to make data enrichment to expand the number of images including people wearing safety helmet. The hardware employed in this section is a PC workstation with Intel Core i7-6050X CPU @ 3.00GHz, NVIDIA GTX 1080 Ti graphics card, 10 GB(11178 MiB), Ubuntu 16.04 system. Python is chosen as the experimental programming language due to its simplicity and powerful functions, and GPU accelerated library is CUDA10.2 and CUDNN7.6.5. In the experiment, image dataset was divided into training set and test set according to the ratio of nine to one in a stochastic way. First of all, the YOLOv3 pre-trained model [18] should be loaded for training with two steps: The first step trains the training set for 50 epochs, with the batchsize set as 16, ADAM optimization algorithm as the optimizer and the learning rate of 0.001; Secondly, the training set is trained for 100 epochs, with the value of batchsize set as 8 and the learning rate set to 0.0001. After each training period, AP, mAP, and loss of the test set are computed, the weights files are preserved every ten epochs. At the same time, the model with the highest value of mAP and the lowest value of loss are preserved.

4.2. Evaluation Criteria

4.2.1. Mean Average Precision

The Mean Average Precision (mAP) is a critical evaluation index owing to its capability to evaluate the quality of detection algorithms. The Mean Average Precision (mAP) is a critical evaluation index owing to its capability to evaluate the quality of detection algorithms, of which the value is the acreage surrounded by the Precision-Recall (P-R) curve and the coordinate axis. The horizontal and vertical of the coordinate axis are $Recall_{SHW}$ and $Precision_{SHW}$ separately. The P-R curve is illustrated by choosing the corresponding precision and recall proportion for distinct thresholds. The formula of $Precision_{SHW}$ is illustrated in Equation (8):

$$Precision_{SHW} = \frac{TP_{SHW}}{TP_{SHW} + FP_{SHW}} \quad (8)$$

where TP_{SHW} means the number that is classified to positive samples accurately, and FP_{SHW} is the number that is classified to positive samples by mistake. The $recall_{SHW}$ is computed according to Equation (9):

$$Recall_{SHW} = \frac{TP_{SHW}}{TP_{SHW} + FN_{SHW}} = \frac{TP_{SHW}}{Precision_{SHW}} \quad (9)$$

where FN_{SHW} is the positive samples, but it is classified into negative samples erroneously. As a result of the beyond formula, the P-R curve can be drafted to compute the AP values of every target categories, and the mAP value of the entire model can be obtained by calculating the average value of the AP values of every target class.

4.2.2. FPS

In addition to the evaluation index of detection precision, one critical metrics of speed is required in detection algorithm on practical construction plant. Fast speed is extremely more significant than the accuracy for some application scenarios with highly real-time requirements. In this paper, the real-time performance of SHW detection is critical for site manager to monitor the violation of the workers anytime and anywhere from surveillance cameras on the sites. In addition, the real-time detection video will be used as the basis for warning, so as to avoid the recurrence of similar situation. The common evaluation indicator of speed is Frame Per Second (FPS), which means the number of images that can be dealt with per second.

4.3. Experimental Results and Analysis

4.3.1. Comparison of Results

In this paper, the novel CSYOLOv3 algorithm achieved high performance in SHW detection of actual construction site compared with the original YOLOv3 model, which can be seen from Table 1. Moreover, the comparison with tiny-YOLOv3 was made in order to ensure the reliability of the proposed algorithm. For the SHW detection, YOLOv3 only obtains the mAP value of 42.5% while the CSYOLOv3 in this paper achieves the mAP value of 67.05%, and tiny-YOLOv3 reaches the mAP value of 38%. In terms of detection speed, the frame rate of the CSYOLOv3 has reached 25, while the original YOLOv3 has reached 19. Tiny-YOLOv3 is faster than the other two models with fps 27. Therefore, in a comprehensive view, the performance of CSYOLOv3 is most suitable for practical engineering application in aspects of accuracy and speed.

Table 1. Statistics of test results for different postures.

Network Structure	Frames Per Second(FPS)	mAP(%)
YOLOv3	19	42.5
CSYOLOv3	25	67.05
tiny-YOLOv3	27	38

The training loss curves of the two algorithms are demonstrated in Figure 6a, of which the horizontal axis and vertical axis represent training time and the loss value of the final training respectively. Figure 6a expresses that CSYOLOv3 model's loss value is smaller than that of the original YOLOv3 eventually, among which the loss value of CSYOLOv3 model is 0.713 while that of original YOLOv3 model is above 2.369. Two critical performance indicators are precision and recall respectively, with the curves illustrated in Figure 6b,c. The horizontal axis in Figure 6b,c expresses the training time, while the vertical axis represents the value of precision and recall respectively. Similarly, it can be seen that the two indicators of CSYOLOv3 outperform that of YOLOv3, with precision 85.41% for CSYOLOv3 and 76.78% for YOLOv3, and recall 67.3% for CSYOLOv3 and 45.58% for YOLOv3. The mAP curves of the two model is displayed in Figure 6d. Figure 6d displays that the highest mAP value of the CSYOLOv3 model is 67.05%, while the highest mAP value of original YOLOv3 model is 42.5%, demonstrating that the improved algorithm has better detection performance than YOLOv3.

(a) The training loss curves. **(b)** The precision curves. **(c)** The recall curves. **(d)** The mAP curves.**Figure 6.** The experimental data curves of YOLOv3 and CSYOLOv3.

4.3.2. Real Construction Site Application

In order to illustrate the practicability, the newly raised detection model is applied to four real-system cases in an actual complex construction site scene, including normal detection, occlusion detection, dense crowd detection and small-scale face detection. The specific detection example effect is shown in Figures 7–10. As can be seen from Figure 7, for normal safety helmet detection, the proposed CSYOLOv3 and YOLOv3 algorithm have all achieved good SHW detection results, safety helmet

wearing can be identified correctly. Moreover, this proposed algorithm has a significant improvement compared with YOLOv3 in terms of detection accuracy. Some complex scenes are considered as follows: Occlusion, dense crowd and small-scale SHW detection.

Figure 7. Results of normal safety helmet wearing (SHW) detection.

Figure 8. Results of occlusion SHW detection.

Figure 9. Results of dense SHW detection.

Figure 10. Results of small size SHW detection.

Result of detection under occlusion is given in Figure 8, it can be seen from Figure 8a that five people were detected by YOLOv3 in total, and the highest prediction accuracy rate is 94%. Moreover, another person wearing a safety helmet in the picture is not detected. As can be seen from Figure 8b, CSYOLOv3 has detected all six SHW targets in occlusion case. In particular, the prediction accuracy of six person detection result is more than 93%, even close to 100%. In a scenario of crowded staff

as shown in Figure 9, seven SHW targets have been identified through YOLOv3 in Figure 9a, while eight people with safety helmet are detected by CSYOLOv3 in Figure 9b. In addition, the prediction accuracy of more than half of the detection boxes is more than 95%. As Figure 10 shows, CSYOLOv3 algorithm outperforms original YOLOv3 in the construction site scene of small target detection.

It can be seen from Figure 10a that three SHW targets have been detected based on Original YOLOv3, while a total number of four SHW targets have been detected correctly based on the CSYOLOv3 in Figure 10b. Furthermore, the prediction accuracy of the proposed algorithm has also been improved to a certain extent. In the aspect of real-time performance, the speed of per image processing was calculated based on the above three algorithms for the four SHW situation mentioned above. Experimental data indicates that the speed of CSYOLOv3 is generally faster than that of YOLOv3, and slightly lower than that of tiny-YOLOv3. Specifically, the average processing time of per image based on CSYOLOv3 is 0.04 s while that of YOLOv3 is 0.053 s, which can be seen from Table 2. Among the three methods compared, the real-time performance of tiny-YOLOv3 is best with the processing time of each image only 0.037 s. In conclusion, compared with the original YOLOv3 algorithm, the SHW detection performance of the CSYOLOv3 is significantly improved.

Table 2. Time of per image processing for four SHW situation.

Method \ Situation	Normal SHW	Occlusion SHW	Dense SHW	Small Size SHW
YOLOv3	0.0526 s	0.0526 s	0.0528 s	0.0523 s
CSYOLOv3	0.0400 s	0.0397 s	0.0405 s	0.0396 s
tiny-YOLOv3	0.0371 s	0.0371 s	0.0374 s	0.0369 s

5. Conclusions

In this paper, on the basis of the original YOLOv3 model, the CSYOLOv3 algorithm is proposed to address the problems of SHW detection under different scenarios of occlusion, dense crowd, small-scale target and so on. Firstly, so as to reduce the computing consumption of the detection model and ameliorate the training speed, CSPNetwork (Cross Stage Partial Network) is introduced to construct CSPDarknet53, which is the improvement to Darknet53. Then, the improved spatial pyramid pooling (SPP) structure is introduced, and the multi-scale prediction network is strengthened through the top-down and bottom-up feature fusion strategies, in order to realize the feature enhancement. Finally, training and testing are carried out on the newly labeled dataset. The experimental data and contrast curves reveal that the presented method can productively strengthen the accuracy and the speed of SHW detection in complex construction site scenes, and the average accuracy value reaches above 90%, and FPS reaches above 20. Since the dataset in this paper are collected under ideal illumination conditions, the primary work in the future is to further expand the dataset on this basis. Considering the illumination factor, some optimization and improvement of the network structure can be done to construct a better lightweight model for real-time target detection in a construction site; thus, the accuracy and real-time performance of the model can be improved to a greater extent.

Author Contributions: Conceptualization, H.W.; Data curation, Z.H.; Formal analysis, F.Z.; Investigation, Y.G. and Z.Y.; Methodology, P.X., H.W., Z.H., Z.Y. and F.Z.; Resources, H.W., P.X. and Z.Y.; Software, H.W. and F.Z.; Validation, Y.G. and F.Z.; Visualization, Z.H.; Writing—original draft, H.W., Z.H. and Y.G.; Writing—review & editing, H.W., Z.H., P.X., Y.G. and Z.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Science Foundation of China (52077213, 62003332, 61473182, 51607177, 61433012, U1435215), Key Project of Science and Technology Commission of Shanghai Municipality under Grant No. 16010500300, Shenzhen Science and Technology Innovation Commission application demonstration project (No. KJYY20160608154421217), Natural Science Foundation of Guangdong Province (2018A030310671); Outstanding Young Researcher Innovation Fund of Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences (201822).

Conflicts of Interest: The authors declare that there is no conflict of interest regarding the publication of this paper.

References

- Braun, A.; Borrmann, A. Combining inverse photogrammetry and BIM for automated labeling of construction site images for machine learning. *Autom. Constr.* **2019**, *106*, 102879. [\[CrossRef\]](#)
- Wang, C.; Li, J.; Wang, T.; Shi, P.; Snoussi, H.; Su, X. Abnormal object detection and recognition in the complex construction site via cloud computing. In Proceedings of the Conference on Research in Adaptive and Convergent Systems, Chongqing, China, 24–27 September 2019; pp. 71–75.
- Li, J.; Liu, H.; Wang, T.; Jiang, M.; Wang, S.; Li, K.; Zhao, X. Safety helmet wearing detection based on image processing and machine learning. In Proceedings of the 2017 Ninth International Conference on Advanced Computational Intelligence (ICACI), Doha, Qatar, 4–6 February 2017; pp. 201–205.
- Manocha, D.; Purkayastha, A.; Chachra, Y.; Rastogi, N.; Goel, V. Helmet Detection Using ML & IoT. In Proceedings of the 2019 International Conference on Signal Processing and Communication (ICSC), Noida, India, 7–9 March 2019; pp. 355–358.
- Waranusast, R.; Bundon, N.; Timtong, V.; Tangnoi, C.; Pattanathaburt, P. Machine vision techniques for motorcycle safety helmet detection. In Proceedings of the 2013 28th International Conference on Image and Vision Computing New Zealand (IVCNZ 2013), Wellington, New Zealand, 27–29 November 2013; pp. 35–40.
- Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
- Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; Jiang, Y. Acquisition of localization confidence for accurate object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 784–799.
- Sindagi, V.A.; Patel, V.M. Generating high-quality crowd density maps using contextual pyramid cnns. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1861–1870.
- Liu, T.; Luo, W.; Ma, L.; Huang, J.; Stathaki, T.; Dai, T. Coupled Network for Robust Pedestrian Detection with Gated Multi-Layer Feature Extraction and Deformable Occlusion Handling. *arXiv* **2019**, arXiv:1912.08661.
- Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9308–9316.
- Qin, Z.; Li, Z.; Zhang, Z.; Bao, Y.; Yu, G.; Peng, Y.; Sun, J. ThunderNet: Towards Real-time Generic Object Detection. *arXiv* **2019**, arXiv:1903.11752.
- Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 142–158. [\[CrossRef\]](#) [\[PubMed\]](#)
- He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
- Luo, H.W.; Zhang, C.S.; Pan, F.C.; Ju, X.M. Contextual-YOLOV3: Implement better small object detection based deep learning. In Proceedings of the 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDI), Taiyuan, China, 8–10 November 2019; pp. 134–141.
- Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2016; pp. 21–37.
- Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
- He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [\[CrossRef\]](#) [\[PubMed\]](#)
- Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
- Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
- Doungmala, P.; Klubsuwan, K. Helmet Wearing Detection in Thailand Using Haar Like Feature and Circle Hough Transform on Image Processing. In Proceedings of the 2016 IEEE International Conference on Computer and Information Technology (CIT), Nadi, Fiji, 8–10 December 2016; pp. 611–614.

21. Dasgupta, M.; Bandyopadhyay, O.; Chatterji, S. Automated Helmet Detection for Multiple Motorcycle Riders using CNN. In Proceedings of the 2019 IEEE Conference on Information and Communication Technology, Durban, South Africa, 6–8 March 2019; pp. 1–4.
22. Rohith, C.A.; Nair, S.A.; Nair, P.S.; Alphonsa, S.; John, N.P. An Efficient Helmet Detection for MVD using Deep learning. In Proceedings of the 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 23–25 April 2019; pp. 282–286.
23. Bo, Y.; Huan, Q.; Huan, X.; Rong, Z.; Hongbin, L.; Kebin, M.; Weizhong, Z.; Lei, Z. Helmet Detection under the Power Construction Scene Based on Image Analysis. In Proceedings of the 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), Dalian, China, 19–20 October 2019; pp. 67–71.
24. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
25. Wang, C.Y.; Mark Liao, H.Y.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of cnn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Glasgow, UK, 23–28 August 2020; pp. 390–391.
26. Xiao, D.; Shan, F.; Li, Z.; Le, B.T.; Liu, X.; Li, X. A target detection model based on improved tiny-yolov3 under the environment of mining truck. *IEEE Access* **2019**, *7*, 123757–123764. [[CrossRef](#)]
27. Misra, D. Mish: A self regularized non-monotonic neural activation function. *arXiv* **2019**, arXiv:1908.08681.
28. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).