Safety Helmet Detection Based on YOLOv5

Fangbo Zhou
School of Electrical and Electronic Engineering
Shanghai Institute of Technology
18855055936@163.com

Huailin Zhao
School of Electrical and Electronic Engineering
Shanghai Institute of Technology
zhao_huailin@yahoo.com

Zhen Nie
School of Electrical and Electronic Engineering
Shanghai Institute of Technology
1063967574@qq.com

2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA) | 978-1-7281-9004-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICPECA51329.2021.9362711

***Abstract*—As the most basic protection for workers, safety helmets have great significance to workers' lives. However, due to a lack of safety awareness, safety helmets are often not worn. With the continuous development of object detection technology, the YOLO series of algorithms with very high precision and speed has been used in various scene detection tasks. To establish a digital safety helmet monitoring system, we propose a safety helmet detection method based on YOLOv5 and annotate the 6045 collected data sets. Finally, we used the YOLOv5 model with different parameters for training and testing. The four models are compared and analyzed. Experimental results show that the average detection speed of YOLOv5s reaches 110 FPS. Fully meet the requirements of real-time detection. Using the trainable target detector's pre-training weight, the mAP of YOLOv5x reaches 94.7%, proving the effectiveness of helmet detection based YOLOv5.**

Keywords—Object detection, Safety helmet detection, YOLOv5, Real-time detection

I. Introduction

In recent years, the detection of scene objects in images [1-3] and videos [4] has become a research hot spot in computer vision. The object detection model can recognize multiple objects in an image and judge the object's category and location [5]. Applying object detection technology to product safety has always been an issue of concern to researchers [6-7]. As the most basic personal protective equipment for workers, safety helmets have significance in protecting workers' lives. However, some workers often do not wear safety helmets for work due to a lack of safety concepts. The detection of safety helmets has become an essential means of production safety monitoring, especially in construction sites, coal mines, and other workplaces with extensive applications.

Traditional object detection methods use a sliding window-based region selection strategy [8], which is not targeted and has high complexity. And the hand-designed feature extractor is not very robust to the diversity of targets. Thanks to the improvement of computers' performance, especially GPUs, the deep learning method [9] that required many calculations has once again attracted researchers' attention, especially the successful application of convolutional neural networks in the field of computer vision. Current object detection tasks mostly use methods based on convolutional neural networks, which have the advantage of not requiring manual feature design when extracting image features.

In recent years, many researchers have proposed a series of object detection algorithms based on deep learning. In 2015, Redmon et al. proposed the YOLO algorithm [10], which is much faster than other algorithms. Redmon et al. [11] proposed the YOLOv3 object detection algorithm in 2018, which further improved detection speed and accuracy. By 2020, the YOLO series of algorithms [12] have been developed to YOLOv5. This paper combines the requirements of the product safety field for helmet detection.

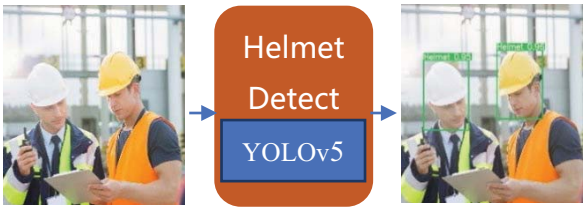

Fig. 1. The structure of the helmet detector. Input a picture, pass the helmet detector, output the detected helmet and its position information, and output its confidence.

As shown in Fig. 1, we use YOLOv5 as a detector to detect whether workers wear helmets. In summary, this paper has the following contributions:

1) We obtained 6045 pictures from the Internet and annotated them. The head without a helmet is annotated as "Alarm." And the head with a correct helmet is annotated as "Helmet."

2) We use the most advanced target detection algorithm to detect whether the worker is wearing a helmet correctly and achieve excellent results.

3) We trained and evaluated YOLOv5 (s, m, l, x) models with different depths and widths and compared them.

The rest of this paper is organized as follows. Section 2 introduces the work related to the object detection algorithm and helmet detection method. In section 3, we present the method of YOLOv5 in detail. Besides, we give the experimental details and analyze the experimental results in section 4. We conclude our work in section 5.

II. Related Work

In this section, we mainly review target detection algorithms and helmet detection methods.

A. Object detection algorithms

Object detection task is widely used in reality. Target detection is to find all objects of interest in the image. It includes two subtasks: determining the category and location of the object. We divide target detection algorithms into traditional methods [13-15] and methods based on convolutional neural networks.

Traditional object detection algorithm: The traditional object detection algorithm mainly consists of three steps: (1) Using sliding window frames of different sizes and proportions to slide on the input picture with a certain step length and as a candidate area; (2) Perform feature extraction on the local information of each candidate region; traditional methods include color-based, texture-based, shape-based methods, and some middle-level or high-level semantic features. The final target that needs to be detected is the final output result of the algorithm. (3) Use classifiers for recognition, such as SVM [16] models. After the check box is judged, a series of candidate boxes that may be the detection target will be obtained, and these candidate boxes may have some overlapping conditions. At this time, NMS [17] is needed to merge the candidate frames.

Although traditional detection algorithms can achieve good results in specific scenarios, in complex environments, such as weather changes, uneven distribution of workers, and different helmets, its accuracy is difficult to guarantee. Its generalization ability is poor. Besides, traditional manual design features require a lot of prior knowledge. The three-part detection process is cumbersome and computationally expensive and cannot meet the real-time monitoring.

Object detection methods based on convolution neural networks are mainly divided into two categories: one is a two-stage approach represented by RCNN [1]. The other is a regression-based target detection algorithm represented by YOLO [10]. In theory, the two-stage algorithm is more accurate than the one-stage algorithm. Represented by Faster-RCNN [13], the two-step target detection algorithm consists of a convolutional layer, a region candidate network, a region of interest pooling layer, and a classification layer. The convolutional layer comprises a set of basic convolutional layers, activation layers, and pooling layers, which are used to extract features and generate feature maps. The region candidate box network is used primarily to create region candidate boxes. The region of interest pooling layer is responsible for Collect feature maps and regional candidate frames; the classification layer determines the target category and corrects the position of the candidate frames according to the regional candidate frames.

The one-stage target detection algorithm truly achieves end-to-end training. Represented by YOLO, the regression-based target detection algorithm completes the target category's determination and the positioning of the target at one time. The entire network structure is composed of only convolutional layers and the input image. After the convolution operation, the target category and position are directly returned. Therefore, the one-stage target detection algorithm is faster than the two-stage target detection algorithm, especially YOLOv5, which has reached an advanced speed and accuracy level.

B. Safety Helmet Detection

Research on the detection of safety helmets has attracted many researchers. Wen et al. [18] proposed a circle/arc detection method based on improved Hough transform and applied it to detecting safety helmets in ATM monitoring systems. Rubaiya et al. [19] combined the image's frequency domain information with the Histogram of Oriented Gradient [20], a popular human detection algorithm, for construction worker detection. And combined color-based feature extraction technology and cyclic Hough transform feature extraction technology to detect helmets' use by construction workers. Li et al. [21] adopted the ViBe background modeling algorithm. According to the segmentation result of the moving target, the real-time human body classification framework is used to accurately and quickly locate pedestrians in the substation. Finally, according to pedestrian detection results, head positioning, color space transformation, and color feature recognition are used to detect wearing helmets.

However, the above methods based on traditional target detection can only be applied to specific scenarios, and the accuracy is not high. With the development of target detection technology, Wu et al. [22] improved the YOLOv3 algorithm to effectively deal with helmet coloring, partial occlusion, many targets, and low image resolution.

III. METHOD

In this section, we first introduce the overall structure of the network. Then we introduce in detail the details of our modified classifier, and evaluation metrics of the dataset.

A. YOLOv5 network

As shown in Fig. 2, the network structure of yoloV5 is divided into three parts, backbone, neck, and output. In the backbone, the input image with 640×640×3 resolution goes through the Focus structure. Using the slicing operation, it first becomes a 320×320×12 feature map, and then after a convolution operation of 32 convolution kernels, it finally becomes a 320×320×32 feature map. The CBL module is a basic convolution module. A CBL module represents Conv2D + BatchNormal + LeakyRELU.

The BottleneckCSP module mainly performs feature extraction on the feature map, extracting rich information from the image. Compared with other large-scale convolutional neural networks, the BottleneckCSP structure can reduce gradient information duplication in convolutional neural networks' optimization process. Its parameter quantity occupies most of the parameter quantity of the entire network. By adjusting the width and depth of the BottleneckCSP module, four models with different parameters can be obtained, namely YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The SPP module mainly increases the receptive field of the network and acquires features of different scales.

YOLOv5 also adds a bottom-up feature pyramid structure based on the FPN structure. With this combination operation, the FPN layer conveys strong semantic features from top to bottom, and the feature pyramid conveys robust positioning features from the bottom up. Combine feature aggregation from different feature layers to improve the network's ability to detect different scales' targets. At the end of the figure, output the classification results and object coordinates.

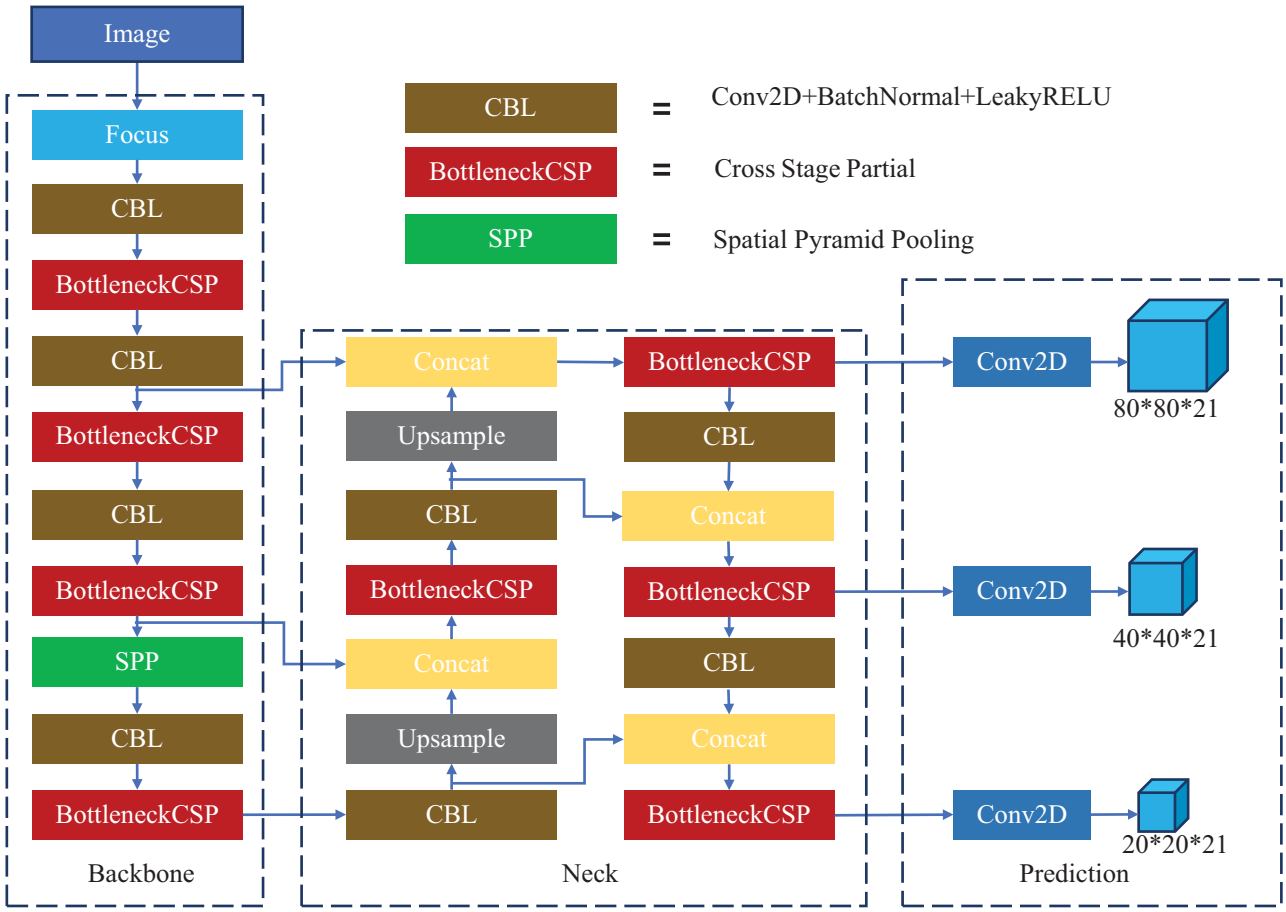

Fig. 2. YOLOv5 network structure

B. Classifier modification

For the COCO dataset [23], there are 80 object categories, and the dimension of the output tensor is $3 \times (5 + 80) = 255$, where 3 represents the three template boxes for each grid prediction. And 5 represents each prediction box's coordinates (x, y, w, h) and confidence (confidence, c).

We have two types of objects in the helmet detection scene, respectively, with helmet and without the safety helmet, so we need to modify the YOLOv5 classifier. The output dimension becomes $3 \times (5 + 2) = 21$. Adapt to the helmet detection scene. By modifying, we can reduce the number of network parameters, reduce computational overhead, and improve detection accuracy and speed.

C. Dataset introduction

Our data set is 6045 pictures collected by crawlers. As shown in Fig 3, the images we selected contain various scenes and helmets of different scales. Besides, we also added some classroom scenes to increase the number of negative samples. And use the labeling software to annotate the object and category, annotate “Helmet” (positive sample) for people wearing helmets, and “Alarm” (negative example) for people wearing helmets. Finally, the labeled pictures are divided into a training set and test set according to the ratio of 2:1.

D. Metrics

The experimental results adopt the average precision mean mAP and the number of frames per second (FPS). For each category of object, first, calculate the Precision rate and Recall rate.

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

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

Fig. 3. Some samples in the dataset

$$AP = \frac{1}{11} \sum_{recall \in [0, 0.1, \cdots, 1]} Precision\ (recall) \qquad (3)$$

TP is a true positive example; FP is a false positive example; FN is a false negative example; AP is the average accuracy of a certain category, and mAP is the average of APs in all categories.

IV. Experiment

In this section, we introduce the experiment's details, and then we show the training results without using pre-training weights and compare the four models of YOLOv5. Finally, we show the experimental result using transfer learning.

A. Experimental details

In our experiment, all the training set pictures are randomly cropped into 640*640 size. And all perform data enhancement methods such as random flip, geometric distortion, illumination distortion, image occlusion, random erase, cutout, mixup, etc. During the testing phase, our pictures were scaled to 640*640, which is convenient for us to evaluate the network. Our experiment's operating system is ubuntu18.04, using the Pytorch framework, and training and testing on a single NVIDIA GTX1660 GPU. We use SGD optimizer to optimize our network. Some experimental hyperparameters are shown in Table I.

TABLE I SOME EXPERIMENTAL HYPERPARAMETERS

Parameter	value
Learning rate	0.01
Learning rate decay	0.999
Learning rate decay step	1
Weight rate decay	5e-4
Momentum	0.937
Batch size	16
Number of iterations	100

B. Experimental results

We use YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x models to train the helmet training data set. And after each round of training is evaluated, the evaluation results are shown in Fig. 4.

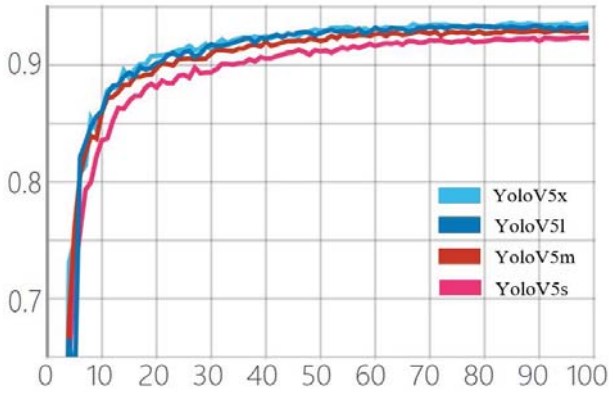

Fig. 4. Without pre-training weights, test accuracy for YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x.

It can be found that in the initial training phase, as the training time increases, the four models all increase rapidly, and the mAP of YOLOv5x with the enormous largest of parameters increases the fastest. As the training time increases, the mAP values of the four models gradually converge. To be able to compare the shortcomings of the four models more quantitatively. We take the best weights of the four models and evaluate them. The evaluation results are shown in Table II. It can be seen that the performance of YOLOv5m is improved by 0.8% compared to YOLOv5s. As the parameter quantity increases, the advantages gradually decrease, and YOLOv5x is only 0.1% higher than YOLOv5l, and the inference speed is twice as slow.

TABLE II. THE PERFORMANCE OF DIFFERENT MODELS WITHOUT PRE-TRAINING WEIGHTS.

Model	mAP	FPS
YOLOv5s	92.3	110
YOLOv5m	93.1	64
YOLOv5l	93.5	37
YOLOv5x	93.6	21

To better display our experiments' results, we found some image in the test set to visualize the results. As shown in Fig. 5, since the convolutional neural network algorithm does not need to extract features manually and enhance the generalization ability, YOLOv5 has a good detection effect on the different colored helmets in (a) and the gray-scale picture in (b). For (c) and (d) that contain many targets, YOLOv5 can also detect them. Especially in the case of mutual occlusion in (d), the model can also determine whether to wear a helmet. Besides, YOLOv5 also detects people who are not wearing helmets (e) and (f). Although the person in (f) was wearing a hat, it was also detected that he did not wear a helmet.

TABLE III. THE PERFORMANCE OF DIFFERENT MODELS WITH PRE-TRAINING WEIGHTS.

Model	mAP	Improved
YOLOv5s	93.6	1.3
YOLOv5m	94.3	0.8
YOLOv5l	94.4	0.9
YOLOv5x	94.7	0.9

Deep Learning Often Requires A Lot Of Data To Enhance The Network's Generalization Ability, But Our Data Set Is Limited. Therefore, To Enhance The Model's Generalization Ability, We Adopt The Transfer Learning [24] Method And Use Pre-Training Weights Provided By Yolov5. We Take The Best Weights Of The Four Models And Evaluate Them. The Evaluation Results Are Shown In Table III. It Can Be Seen That After Using And Training The Weights, Map Increases With The Increase Of The Number Of Parameters. Improved Is The Improvement Result Of Using Pre-Training Weights Compared To Not Using Pre-Training Weights.

Fig. 5. Visualization results form test dataset

V. Conclusion

This work introduces the YOLOv5-based safety helmet detection method in detail, including the YOLOv5 network structure, classifier settings, and data set processing. Besides, experiments were carried out using YOLOv5 models (s, m, l, x) with different parameters. As a result, a YOLOv5-based helmet detection method has achieved excellent results in test accuracy and speed. After using the pre-training weights, the map values of various models increased by about 1%. In the next work, we will collect a large number of positive and negative samples of real scenes to improve the network's generalization ability.

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region based convolutional networks for accurate object detection and segmentation," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 1, pp. 142–158, 2016.

[2] R. Girshick, "Fast r-cnn," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, 2015, pp. 91–99.

[4] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang et al., "T-cnn: Tubelets with convolutional neural networks for object detection from videos," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 10, pp. 2896–2907, 2018.

[5] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," International journal of computer vision, vol. 88, no. 2, pp. 303–338, 2010.

[6] R. Waranusast, N. Bundon, V. Timtong, and et al, Machine vision techniques for motorcycle safety helmet detection, 2013 28th International Conference on Image and Vision Computing New Zealand (IVCNZ 2013), 35-40, 2013.

[7] R. R. V. e Silva, K. R. T. Aires, R. M. S. Veras, Helmet detection on motorcyclists using image descriptors and classifiers, 2014 27th SIBGRAPI Conference on Graphics, Patterns and Images, 141-148, 2014.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Computer Vision and Pattern Recognition, CVPR 2005. IEEE Computer Society Conference on, vol. 1. IEEE, 2005, pp. 886–893.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks,"in Advances in neural information processing systems, 2012, pp. 1097–1105.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[11] "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[12] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," arXiv preprint, 2017.

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 1. IEEE, 2001.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1. IEEE, 2005, pp. 886–893.

[15] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008, pp. 1–8.

[16] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," in Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011, pp. 89–96.

[17] C. Papageorgiou and T. Poggio, "A trainable system for object detection," International journal of computer vision,vol. 38, no. 1, pp. 15–33, 2000.

[18] Wen, Che-Yen, Shih-Hsuan Chiu, Jiun-Jian Liaw, and Chuan-Pin Lu. "The safety helmet detection for ATM's surveillance system via the modified Hough transform." In IEEE 37th Annual 2003 International Carnahan Conference on Security Technology, 2003. Proceedings., pp. 364-369. IEEE, 2003.

[19] Rubaiyat, Abu HM, Tanjin T. Toma, Masoumeh Kalantari-Khandani, Syed A. Rahman, Lingwei Chen, Yanfang Ye, and Christopher S. Pan. "Automatic detection of helmet uses for construction safety." In 2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW), pp. 135-142. IEEE, 2016.

[20] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009, pp. 32–39.

[21] Li, K., Zhao, X., Bian, J. and Tan, M., 2018. Automatic Safety Helmet Wearing Detection. arXiv preprint arXiv:1802.00264.

[22] Wu, Fan, Guoqing Jin, Mingyu Gao, H. E. Zhiwei, and Yuxiang Yang. "Helmet detection based on improved yolo v3 deep model." In 2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC), pp. 363-368. IEEE, 2019.

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in European conference on computer vision. Springer, 2014, pp. 740–755.

[24] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8697–8710.