Hindawi
Complexity
Volume 2020, Article ID 6662603, 8 pages
https://doi.org/10.1155/2020/6662603

Research Article

Towards Pedestrian Target Detection with Optimized Mask R-CNN

Dong-Hao Chen, Yu-Dong Cao, and Jia Yan

School of Electronics and Information Engineering, Liaoning University of Technology, Jinzhou, Liaoning 121001, China

Correspondence should be addressed to Yu-Dong Cao; caoyd@lnut.edu.cn

Received 4 November 2020; Revised 4 December 2020; Accepted 10 December 2020; Published 22 December 2020

Academic Editor: M. Irfan Uddin

Aiming at the problem of low pedestrian target detection accuracy, we propose a detection algorithm based on optimized Mask R-CNN which uses the latest research results of deep learning to improve the accuracy and speed of detection results. Due to the influence of illumination, posture, background, and other factors on the human target in the natural scene image, the complexity of target information is high. SKNet is used to replace the part of the convolution module in the depth residual network model in order to extract features better so that the model can adaptively select the best convolution kernel during training. In addition, according to the statistical law, the length-width ratio of the anchor box is modified to make it more accord with the natural characteristics of the pedestrian target. Finally, a pedestrian target dataset is established by selecting suitable pedestrian images in the COCO dataset and expanded by adding noise and median filtering. The optimized algorithm is compared with the original algorithm and several other mainstream target detection algorithms on the dataset; the experimental results show that the detection accuracy and detection speed of the optimized algorithm are improved, and its detection accuracy is better than other mainstream target detection algorithms.

1. Introduction

The advancement of science and technology makes machine vision have broad application prospects in video surveillance, intelligent transportation, unmanned driving, and other projects. With the popularity of high-performance camera equipment and the surge in demand for automated analysis of video content, how to extract accurately and efficiently the target in the video has become an urgent problem to be solved, especially in the study of the pedestrian target area, and even more a hot issue in the field of machine vision research. Pedestrian detection is the basis of most pedestrian dynamic analyses. The more accurate detection result is related to whether the follow-up tracking, segmentation, estimation, and other tasks can be completed accurately and efficiently.

There are two main branches of target detection algorithms: one is the motion detection algorithm based on the difference between video sequences, and the other is the algorithm based on machine learning. The first type of method has fast calculating speed, but poor anti-interference ability. When the environment changes, the target appears dense, or the target does not move, it is easy to produce a large number of missed and wrong detections, and the robustness is poor. The common methods of this kind include frame difference method, background difference method, ViBe algorithm [1], and ViBe+ algorithm [2]. The second type of method is divided into traditional machine learning and deep learning. The deep learning method uses a multilayer neural self-learning network to repeatedly achieve excellent results in world-class target detection competitions.

Target detection based on deep learning can be divided into anchor-based and anchor-free. The most important difference between the two methods is whether the anchor box is used to extract the candidate target frame of the image during the learning process. Compared with the anchor-based method, the anchor-free method has a simpler

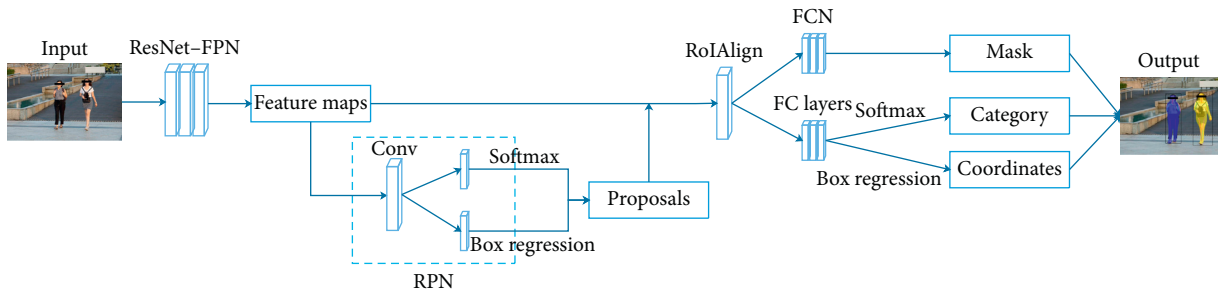

Figure 1: Mask R-CNN framework.

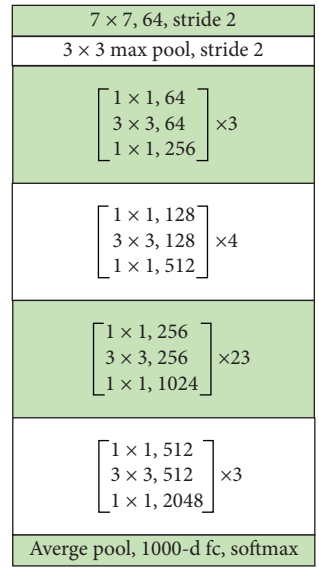

Figure 2: Network structure of ResNet101.

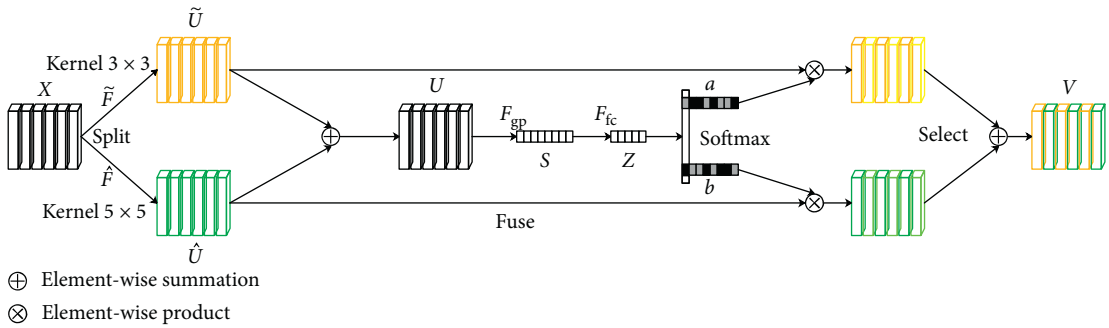

Figure 3: SKNet module.

network structure and stronger model migration ability. The anchor-free method is based on the complete feature gold tower, which has a huge amount of calculation, while the anchor-based method reduces the number of layers of the pyramid, which greatly reduces the amount of calculation, the detection speed is faster, and the detection accuracy is higher. Now, the mainstream detection algorithms such as YOLOv2 [3], YOLOv3 [4], Faster R-CNN [5], and Mask R-CNN [6] are anchor-based methods.

2. Related Work

The target detection algorithm based on deep learning has been fully developed in more than ten years. Now, there are many branches in the area of the target detection. The deep learning target detection algorithm based on region proposal represented by R-CNN [7, 8] is one of the important branches. R-CNN significantly improved the accuracy of the detection results by using the convolutional neural network in 2014. It

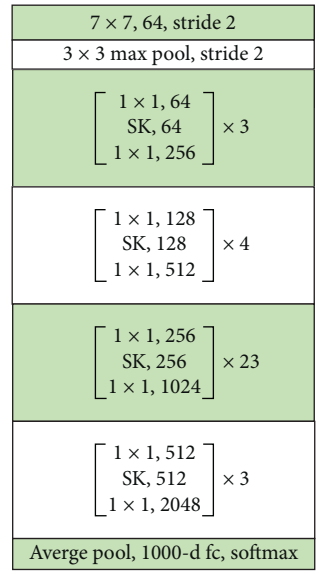

Figure 4: Network structure of SKNet-101.

was the first time to extract the features of the detection target. Subsequently, Fast R-CNN [9] optimized the serial feature extraction method of R-CNN, and only one CNN was used for each image, which greatly improved the detection speed. After that, Faster R-CNN [5] made further optimization. Instead of using selective search algorithms to generate candidate regions, the regions to be detected are extracted through a region proposal network (RPN) so that the complete target detection process is through a neural network which is used to further improve the detection accuracy and speed, and a real end-to-end target detection framework is realized. Mask R-CNN [6] is a further extension of this series of deep learning target detection algorithms. It adds a segmentation task branch based on the Faster R-CNN detection branch, and the segmentation task is performed simultaneously with the classification and regression tasks. The detection task can be better extended to other tasks, and the detection effect is also more ideal.

SENet (Squeeze-and-Excitation Network) [10] was proposed by Hu's team and won the championship in the ImageNet classification competition in 2017. It is just a lightweight network model that implements an attention mechanism on channels so that the network can adaptively select appropriate feature channels. On this basis, Li et al. proposed SKNet (Selective Kernel Network) [11], which performs an attention mechanism on the convolution kernel so that the network can adaptively select the appropriate convolution kernel in 2019.

Nowadays, target detection has made a new development. Hsu et al. [12] proposed two strategies to enable the detector to detect OOD (out-of-distribution) samples without OOD data training. Wang et al. [13] introduced the intersection of the human body and object into training to improve the detection performance. Zhang et al. [14] proposed a method to automatically select positive and negative samples based on the statistical characteristics of the object and proved that simply stacking the number of anchor boxes cannot improve the detection accuracy.

Based on the Mask R-CNN target detection algorithm, we have made some optimizations to improve the accuracy of pedestrian target detection. The main work of this article consists of the following three parts:

3. Proposed Network Framework

3.1. Mask R-CNN Algorithm. The Mask R-CNN algorithm is a melioration based on the Faster R-CNN detection algorithm which introduces a full convolutional network (FCN) to generate mask. In the real-time target detection process, the pixels of the target are classified accurately, and then the contour of the target is judged. The framework of the algorithm is shown in Figure 1.

(1) In the dataset preprocessing stage, each image is added with noise and then fuzzified; the three kinds of images are used as the pretraining set so that the amount of data in the pretraining set is tripled without relabeling. Data enhancement is realized.

(2) In the RPN, the anchor box is optimized for pedestrian targets. The proportions that are more suitable for pedestrian targets are used, which makes the network training results more reasonable, higher detection accuracy, no increase in calculations, and faster detection convergence.

(3) In the ResNet, the SKNet lightweight network module is used to replace the part of the convolution module so that the model can adaptively select the best convolution kernel during the training process, increase the quality of feature representation, and improve detection accuracy.

The image is first inputted into the backbone network composed of the ResNet and the FPN. The backbone network extracts some shared feature maps that combine the coordinate information of the detected target position and the appearance texture information. Then, the RPN area

Figure 5: Dataset expansion example.

suggestion network uses a sliding window to traverse these feature maps to generate several anchor frames with a combination of fixed scale and aspect ratio. These anchor frames are candidate areas. In the proposal layer, the anchor frame that is more likely to contain the detected target is selected as the candidate area. The specific method is to exclude the anchor frame which goes beyond the image boundary, has high overlap rate, or low confidence level. Then, the nonmaximum suppression (NMS) method is used to select the anchor box with the higher score [15].

In the RoIAlign layer of the Mask R-CNN algorithm, the quantization operation in the feature aggregation process is replaced by the bilinear interpolation method, which avoids the problem of mismatch and improves the accuracy of detection and segmentation. The Mask R-CNN algorithm shares the convolutional layer with the candidate region generation network for classification and regression problems, which improves the efficiency of the algorithm. The Mask R-CNN algorithm uses the softmax function and the multitask function to obtain the classification value and the

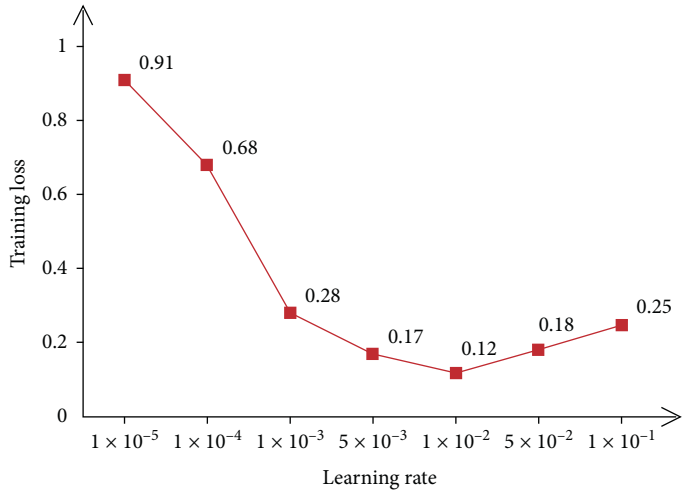

Figure 6: The change of training loss with learning rate.

Table 1: The influence of the number of training iterations on the test accuracy.

Number of training iterations/time	Test accuracy (%)
5000	63.26
10,000	80.79
15,000	87.78
20,000	84.63

Table 2: Training hyperparameters.

Hyperparameter	Value
Momentum	0.9
Weight decay	0.0001
Learning rate	0.01
Batch size	64
Iteration	15,000

Table 3: Algorithm optimization and comparison experiment results.

Network model	AP (%)	FPS
Mask R-CNN	74.63	4.99
Mask R-CNN optimized only for the dataset	80.78	4.99
Mask R-CNN optimized only for RPN	78.54	5.36
Mask R-CNN optimized only for ResNet	83.37	4.87
Mask R-CNN optimized for all three	85.09	5.27

Table 4: Comparison of the optimized algorithm and several mainstream algorithms' experimental results.

Network model	AP (%)	FPS
YOLOv2	69.73	15.23
YOLOv3	72.87	12.76
Faster R-CNN	73.58	5.42
Mask R-CNN	74.63	4.99
Optimized Mask R-CNN	85.09	5.27

regression box parameter value. In the FCN, the sigmoid function is used to output the mask value to realize pixel-level instance segmentation.

During the training process, the Mask R-CNN algorithm defines the multitask loss function for each sampled region of interest (RoI) as

$$L = L_{\text{cls}} + L_{\text{box}} + L_{\text{mask}}. \tag{1}$$

L_{cls} is the classification error, L_{box} is the detection error, and L_{mask} is the segmentation error.

L_{cls} and L_{box} in Mask R-CNN are defined as

$$L' = \frac{1}{N_{\text{cls}}}\sum_i L_{cls}\left(p_i, p_i^*\right) + \lambda\frac{1}{N_{\text{reg}}}\sum_i p_i^* L_{\text{reg}}\left(t_i, t_i^*\right). \tag{2}$$

p_i represents the predicted probability of the i-th target on the anchor point. p_i^* is determined by the sign of the anchor point sample. When the anchor point sample is positive, p_i^* is 1; otherwise, it is 0. Both t_i and t_i^* are vectors composed of four translation and scaling parameters, which, respectively, measure the degree of change of the positive sample anchor point relative to the prediction area and the label area. The weights N_{cls}, N_{reg}, and λ control the two losses to keep balance.

Classification loss and regression loss are defined as

$$L_{\text{cls}}\left(p_i, p_i^*\right) = -\log\left[p_i p_i^* + \left(1 - p_i^*\right)\left(1 - p_i\right)\right], \tag{3}$$

$$L_{\text{reg}}\left(t_i, t_i^*\right) = \text{smooth}_L\left(x\right)\left(t_i - t_i^*\right). \tag{4}$$

$\text{smooth}_L(x)$ is the robust loss, which is determined by the translation x of the corrected frame on the horizontal axis at the anchor point. It is defined as

$$\text{smooth}_L\left(x\right) = \begin{cases} 0.5x^2, & |x| < 1, \\ |x| - 0.5, & \text{otherwise.} \end{cases} \tag{5}$$

L_{mask} in Mask R-CNN is an average binary cross-entropy function that describes the loss of semantic segmentation branches. In the mask branch, the input feature map will be output into a $k \times m \times m$ format after processing, and k and m,

respectively, control the dimension and scale of the feature map. The relative entropy is obtained by the pixel-by-pixel sigmoid calculation of the output feature map, and the average entropy error is L_{mask}.

3.2. Optimized Mask R-CNN. We optimize the RPN by modifying the aspect ratio of the anchor frame. We also modify the network structure of the ResNet.

3.2.1. Optimization of RPN. In the training process of the original RPN, the anchor frame in the sliding window is composed of three kinds of areas (128^2, 256^2, and 512^2) and three aspect ratios (1 : 1, 1 : 2, and 2 : 1). There are totally 9 kinds of anchor frames [6]. However, if only pedestrian targets are detected, this setting will affect the convergence speed of training learning and reduce the detection accuracy, which is unreasonable. According to the statistical law, the average aspect ratio of the human body when standing and walking is about 0.41 [16]. Therefore, the RPN network is optimized for pedestrian targets, removing the anchor frame with an aspect ratio of 2 : 1 and replacing it with the width-height ratio. For the anchor frame with a ratio of 2 : 5, modify the types of anchor frame aspect ratios to 1 : 1, 2 : 5, and 1 : 2, and keep the original three areas unchanged, and the number of anchor frame types is still 9. For each image, the total number of anchor frames during training remains unchanged from the original Mask R-CNN algorithm.

3.2.2. Optimization of ResNet. For Mask R-CNN, the most commonly used deep residual network models are ResNet50 and ResNet101. Compared with ResNet50, ResNet101 has higher accuracy. We use the ResNet101 network model as the basis for optimization and improvement. The network structure of ResNet101 is shown in Figure 2.

SKNet is a lightweight embedded module that can adaptively change the size of the convolution kernel as the information scale changes, thereby controlling the receptive field of the network and better capturing the feature information of the target. As shown in Figure 3 [11], SKNet consists of three parts. In the split process, the feature maps are, respectively, passed through a convolution with a 3×3 kernel and a convolution with a 5×5 kernel to generate feature maps $\widetilde{U}$ and $\widehat{U}$. In the fuse process, $\widetilde{U}$ and $\widehat{U}$ are added to get the feature map U. U goes through an operation called global average pooling first. Then, U passes through two fully connected layers and goes through a process of first decreasing the dimension and then increasing the dimension. After that, weight matrix a and weight matrix b can be obtained. The final feature map V is obtained by weighted addition in the select process.

In this article, SKNet module was embedded into the ResNet101 network. Convolution module with 3×3 cores was replaced by the convolution module consisting of two different cores and a feature channel weight full connection layer. The new feature extraction network was named SKNet-101. The optimized ResNet can better represent the characteristics of the target, thereby further improving the detection accuracy. The optimized network structure of SKNet-101 is shown in Figure 4.

4. Experimental Results and Analysis

The program running environment is Windows 10 operating system, PyCharm 2019.3.3 platform-integrated Python 3.6 is installed, and the runtime library includes Keras 2.1.6, matplotlib 3.2.2, tensorflow 1.14.0, numpy 1.19.0, and opencv 4.2.0.

4.1. Dataset Enhancement Processing. The classic COCO 2014 dataset [17] is used as the training and testing set. The COCO dataset is a target detection dataset released by Microsoft with rich detection types. It contains 80 different types of targets and more than 200,000 labeled images. Many scholars use it for target detection training and learning. We selected 1,000 pedestrian images from the "person" category, in which scenes are under different angles, lighting, and pedestrian density as much as possible to increase the complexity of the data. This dataset is composed of 1000 pedestrian images, of which 900 are used as the training set and 100 are used as the test set. There are 892 positive sample images in the training set and 3262 pedestrian targets and 99 positive sample images in the test set and 478 pedestrian targets.

In order to achieve the purpose of data enhancement, we added salt-and-pepper noise to 900 images in the training set and then used the median filter with a kernel of 3 for each image, as shown in Figure 5. The three kinds of images were used together as the optimized training set and compared with the original training set without data enhancement. It is proved that reasonable expansion of the dataset is conducive to fully learning the characteristics of pedestrian images and improving the detection performance.

4.2. Parameter Setting. The Mask R-CNN optimized for pedestrian targets is used as a model to complete the detection training of pedestrian targets, and some hyperparameters were set before the training starts to speed up the convergence speed and prevent overfitting.

There are three important parameters in the SKNet module. Since the dual-weight model is used, the number of branches M was set to 2. In order to achieve the optimal feature representation, the number of groups G was set to 32, and the fc scaling ratio R was set to 16. As shown in Figure 6, we first recorded the changes in training loss under different learning rates (LRs) in the Mask R-CNN overall training network. It can be seen that the training loss is the smallest when LR is set to 0.01.

Under the premise of setting LR to 0.01, a comparative experiment on the influence of several training iterations on test accuracy was conducted. The experimental results are shown in Table 1.

The overall test accuracy rate rises with the increase in training iterations/time, as shown in Table 1.

The test accuracy rate reaches its peak after 15,000 iterations. There is an overfitting situation, and the test

accuracy rate drops slightly after 20,000 iterations. Therefore, we finally selected 15,000 iterations during training. The specific values of the training hyperparameters of the overall model are shown in Table 2.

4.3. Experimental Results and Analysis. We compared the learning situation of the original Mask R-CNN algorithm on the training set without data expansion, the learning situation of the original Mask R-CNN algorithm on the training set after data expansion, the learning situation of the Mask R-CNN algorithm after optimizing the RPN on the training set without data expansion, the learning situation of the Mask R-CNN algorithm after optimizing the ResNet on the training set without data expansion, and the learning situation of the Mask R-CNN algorithm after optimizing the RPN and the ResNet on the training set after data expansion. There are two main comparative experimental indicators, namely, AP (average precision) and FPS (frames per second). The specific comparative experimental results are shown in Table 3.

It can be seen from Table 3 that the AP of the detector can be increased by 6.15% by using data expansion, and the FPS is almost unchanged, still at 4.99. In the RPN area, it is recommended to select a suitable anchor frame at each position during the network training stage, which can increase the AP by 3.91% and the FPS slightly by 0.04. Using the SKNet-101 network structure can increase AP by 8.74% but decrease FPS slightly. Using three methods to optimize the model can increase the AP of the detector by 10.46% and FPS by 0.28 when detecting pedestrian targets. It proves that the optimization method can significantly improve the detection accuracy of pedestrian targets and slightly increase the detection speed.

We also compare the optimized detector with several mainstream target detection algorithms on the test set. The experimental results are shown in Table 4.

It can be seen in Table 4 that the AP of the optimized detector is superior to other mainstream algorithms in the pedestrian target detection, and the detection accuracy has been significantly improved.

5. Conclusion

We optimize the RPN of Mask R-CNN and generate a new network structure named SKNet-101 by introducing the SKNet module in the feature extraction stage so that the network can select adaptively the appropriate convolution kernels. We also optimize the representation of the target by modifying the scale of the anchor frame in the regional proposal stage. The training set is expanded to improve the accuracy of the algorithm when detecting pedestrian targets. However, the optimization method has certain limitations. The optimization of the RPN can only improve the detection accuracy of pedestrian targets. When detecting other targets, the detection accuracy may be reduced. Moreover, the problem of relatively slow detection speed in R-CNN series has not been solved well. In future research, the detection speed needs to be improved.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported by the Project of the Natural Science Foundation of Liaoning Province (Grant no. 2019ZD0702) and the Project of Time Series Modeling and Application Research Based on Deep Learning, Natural Science Research Project of Anhui Provincial Department of Education (Grant no. KJ2019A1203).

References

[1] O. Barnich and M. V. Droogenbroec, "Vibe: a powerful random technique to estimate the background in video sequences," in *Proceedings of the IEEE International Conference on Acoustics*, pp. 945–948, IEEE, Taipei, Taiwan, April 2009.

[2] M. V. Droogenbroeck and O. Paquot, "Background subtraction: experiments and improvements for ViBe," *Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 16, no. 7, pp. 16–21, 2012.

[3] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the 30rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7263–7271, CVPR, Honolulu, HI, USA, July 2017.

[4] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," 2018, https://arxiv.org/pdf/1804.02767.pdf,.

[5] 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.

[6] K. M. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, no. 6, pp. 386–397, 2018.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the 27rd IEEE Conference on Computer Vision And Pattern Recognition (CVPR)*, pp. 580–587, Columbus, OH, USA, November 2014.

[8] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proceedings of the Computer Vision, ECCV 2014*, pp. 834–849, Zurich, Switzerland, September 2014.

[9] R. Girshick, "Fast R-CNN," in *Proceedings of the 15rd IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, IEEE, Santiago, CL. USA, December 2015.

[10] J. Hu, L. Shen, G. Sun, and S. Albanie, "Squeeze-and-excition networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, 2017.

[11] X. Li, W. H. Wang, X. L. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Long Beach, CL, USA, November 2019.

[12] Y. C. Hsu, Y. Shen, H. X. Jin, and Z. Kira, "Generalized odin: detecting out-of-distribution image without learning from

out-of-distribution date," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Tokyo, Japan, September 2020.

[13] T. C. Wang, T. Yang, M. Danelljan, and F. Khan, "Learning human-object interaction detection using interaction points," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Tokyo, Japan, September 2020.

[14] S. F. Zhang, C. Chi, Y. Q. Yao, and Z. Lei, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Tokyo, Japan, September 2020.

[15] M. M. Karim, D. Doell, R. Lingard, and Z. Z. Yin, "A region-based deep learning algorithm for detecting and tracking objects in manufacturing plants," *Procedia Manufacturing*, vol. 39, pp. 168–177, 2019.

[16] W. Liu, S. C. Liao, W. Q. Ren, and W. D. Hu, "High-level semantic feature detection: a new perspective for pedestrian detection," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Tokyo, Japan, September 2020.

[17] T. Y. Lin, M. Marie, S. Belongie, and J. Hays, "Microsoft COCO: common objects in context," in *Proceedings of the 13th European Conference on Computer Vision-ECCV*, pp. 740–755, Zurich, Switzerland,, September 2014.