Research Article

Attention-Enhanced One-Stage Algorithm for Traffic Sign Detection and Recognition

He Huang¹, Qice Liang¹, Dean Luo¹ and Dong Ha Lee²

¹*School of Geomatics and Urban Spatial Information, Beijing University of Civil Engineering and Architecture,
1 Zhanlanguan Road, Beijing 100044, China*

²*Department of Integrated Energy & Infra System, Kangwon National University, Chuncheon 24341, Republic of Korea*

Correspondence should be addressed to Qice Liang; liangqice@163.com

Received 5 October 2021; Revised 10 February 2022; Accepted 12 February 2022; Published 4 March 2022

Academic Editor: Bin Gao

Copyright © 2022 He Huang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Detection and recognition of traffic signs are the keys to creating advanced driving assistance systems. Making highly precise maps also requires the identification and extraction of such road elements as traffic signs. Traditional detection and recognition methods no longer meet today's needs, and object recognition algorithms based on deep learning have become the mainstream solution. However, current algorithms have limitations. The recognition speed of one-stage strategy algorithms is fast, but recognition accuracy is not satisfactory especially for small objects. The accuracy of two-stage algorithms is higher, but the recognition speed is extremely slow. This paper solves these problems with a proposed parallel attention convolution module, a channel attention pyramid network, and a loss function diagonal and center point IoU based on the YOLOv3 algorithm. The improved models in this paper are compared with SSD, YOLOv3, and Faster RCNN. Experimental results show that the proposed models have some improvement over the above models: the mAP of the models with PACM, CAFPN, and DCPIoU was 76.02%, compared with SSD300, SSD500, Faster RCNN, and YOLOv3, which had improvements of 9.27%, 6.93%, 2.94, and 5.3%, respectively. And the FPS of the improved model is basically the same as the original YOLOv3, without reducing the real-time performance.

1. Introduction

Autonomous driving has become a popular research topic recently. The advanced driving assistance system is emerging as the first transition technology for autonomous driving at the L4 level or above. Therefore, designing a traffic sign recognition algorithm with high accuracy and robustness is essential.

The accuracy and robustness of automatic recognition of traffic signs using deep learning significantly improve the traditional approach, which uses template matching and support vector machines. A feature expression method based on multilevel chain code histograms has been proposed in the literature [1] for template matching recognition that is less computationally intensive, is more real-time than previous methods, and has a higher ability to discriminate between the shapes of traffic signs. One model combines a multilayer perceptron with the Gabor wavelet transform

[2]. The result accurately locates traffic signs and performs real-time classification using a single fast Fourier transform. Another method first converts the image from color space to grayscale space [3], then uses shape templates to obtain the region of interest, represents the features with HOG, and then uses SVM for classification. Although traditional methods can recognize traffic signs, the process is cumbersome, poor in extracting features, and inadequate in meeting the requirements of real-time autonomous driving. Therefore, a deep learning method based on convolutional neural networks has emerged.

With its powerful feature extraction, deep learning based on a convolutional neural network has become the mainstream process for traffic sign recognition. CNN object recognition algorithms are either one-stage or two-stage. Representative one-stage algorithms are YOLOv1–v4 and SSD [4–8]; two-stage algorithms include RCNN, Fast RCNN, and Faster RCNN [9–11]. Traffic sign in photos

has a large range of sizes, with large objects up to several hundred pixels and small objects only a dozen. Moreover, small object recognition presents technical difficulties. Researchers have optimized the model algorithm in various ways to improve model recognition performance, such as backbone, feature fusion, and loss function.

The backbone plays a significant role in the feature extraction module of the network, and the ability to extract feature information from input images directly affects model prediction. Spatial pyramid pooling (SSP) is added to the backbone, based on YOLOv3 [12], and improves the model's ability to recognize forbidden signs. Another idea from the literature [13] adds mixed-depth convolution to the backbone, which improves the feature extraction ability of the backbone through multiscale convolution. Literature [14] incorporates deformable convolution into the backbone to improve the feature extraction capability of the model in complex environments. Another approach improves the one-way backbone structure of Faster RCNN by adding deconvolution in VGG16 to transfer the information of different-depth feature maps [15].

Most of the backbone structure contains multiscale output, and the interactive fusion of information from different-scale feature maps also affects the model performance. Feature fusion methods from the initial FPN to PANet [16, 17], one-way fusion increased to bidirectional fusion, and then from PANet to BiFPN [18], manual design to neural network structure search, the performance of feature fusion structures that simply change the connection method reach a bottleneck. Feature fusion combined with attention modules may be an effective direction in development. Attention mechanisms in computer vision are also a popular research topic, from spatial transformer networks in the beginning to later squeeze-and-excitation networks [19, 20]. Another contemporary approach includes fusing channel and spatial attention (convolutional block attention module or CBAM) [21]. The literature [13] modified the feature fusion of YOLOv5 to AAF (Attention Feature Fusion) [22].

The results of this study indicate that the loss function is one key to model training. A comprehensive loss function can accelerate model convergence and obtain better parameters. Taking the YOLOv3 algorithm as an example, its loss function has three parts: confidence, classification, and anchor box localization. Cross-entropy is used as the loss function for confidence and classification and achieves satisfactory results. The mean square error (MSE) is used as the loss function for localization, but MSE does not make the model fit the best state, so a more reasonable loss function for localization needs to be determined. The literature [23] proposed GIoU based on IoU, which focuses on nonoverlapping regions compared to IoU. Other literature has proposed a DIoU and CIoU based on GIoU that considers the scale relationship [24]; this achieved satisfactory results on the SSD, YOLO, and RCNN algorithms.

In summary, to further improve model performance, this paper optimizes the accuracy of small-object recognition by presenting a new feature extraction network with a parallel convolutional structure, constructing a channel attention pyramid by adding an attention mechanism design, adopt-

ing the YOLOv3 recognition strategy, and optimizing the loss function of anchor box localization. The paper also offers a new one-stage algorithm that improves small-object recognition. The main improvements offered in this paper are as follows:

- (i) A parallel attention convolution module (PACM) to improve the feature extraction of the backbone by fusing a parallel structure with an attention mechanism
- (ii) A new feature fusion structure, the fusion structure channel attention feature pyramid network (CAFPN), which realizes information sharing of channel features between differently sized feature maps and improves the quality of feature fusion
- (iii) A diagonal and center point IoU (DCPIoU) loss function, based on the optimization of the DIoU loss function, and the diagonal loss penalty term between the prediction box and the true box

2. Materials and Methods

Figure 1 shows the network structure of the proposed modified algorithm. The principal algorithm is based on the same prediction strategy used by the YOLOv3 network. However, the backbone, feature fusion, and the localization loss function of the anchor box of the original algorithm were reconstructed. The vertically connected structure on the left side of Figure 1 shows the backbone feature extraction network with the PACM. Images of 416×416 pixels are input, and 52×52 , 26×26 , and 13×13 feature maps are output after processing by four different residual blocks, six base convolution modules, and spatial pyramid pooling (SPP). The complex horizontally and vertically connected network on the right side of Figure 1 is the CAFPN feature fusion structure, which employs the bidirectional feature fusion of PANet with a channel attention pyramid structure in parallel; this ensures the bidirectional fusion of feature maps. Combining information between feature maps of different sizes makes the bidirectional fusion of channel attention weights possible for feature maps of different depths. This weights the channel dimensions of the output feature maps and identifies the useful information. The loss function of the original model was also optimized, and a DIoU-based DCPIoU loss function was improved for anchor box localization training. Compared with the original MSE loss function, the improved DCPIoU loss function not only adds the ratio of the distance between the centroids of two boxes and the diagonal of the minimum enclosing rectangle but also adds the difference between the diagonal of the minimum enclosing rectangle and the diagonal of the intersecting rectangle of the two boxes to improve the degree of model convergence.

2.1. Backbone with Parallel Attention Convolution Module.

The backbone consists of three structures: residual unit, base convolution module, and spatial pyramid pooling. The base convolution module processes the input data by 2D convolution, batch normalization, and then LeakyReLU activation.

FIGURE 2: Residual block structure diagrams: (a) residual block A; (b) residual block B.

channel weight vector is extracted from the upsampling feature map. After extraction, the upsampling feature map is fused with the feature map of size 26×26 , the weight vector is also extracted from the fused feature map, and then, the extracted vector is fused with the vector extracted from the upsampling feature map. Next, the above fusion steps are repeated. After two upsampling parallel fusions, the fused feature maps of three scales and the corresponding fused channel weight vectors have been obtained, and for the convenience of the following description, the three fused vectors are denoted as ω_1 , ω_2 , and ω_3 in the order of 52×52 , 26×26 , and 13×13 , and the three upsampled fused feature maps are denoted as P1, P2, and P3. Then, the downward fusion and output are carried out; firstly, P1 is processed by Convblock, which contains 1×1 convolution, 3×3 convolution, and spatial attention. After being processed by Convblock, it is multiplied with the corresponding weight vector ω_1 , and then, the channel alignment is realized by convolution according to the number of traffic sign categories and the aligned feature map as the output result of this size. Next, the result of multiplication with ω_1 is downsampled, and the channel weight vector is extracted from the downsampling feature map and fused with ω_2 . The fused feature map is processed by Convblock and multiplied with the fusion vector, and then, the channels are also aligned according to the number of traffic sign types as the output result of this size. The previous downward fusion step is repeated to obtain the output for 13×13 . Finally, the output results of the three sizes are sent to the detection head, and this paper use the same detection head as YOLOv3.

FIGURE 3: The effect of prediction box and real box point positions.

FIGURE 4: Prediction box and real box for a special scene.

2.3. Diagonal and Center Point IoU. The YOLOv3 loss function is divided into anchor box localization loss, confidence loss, and classification loss. The localization loss uses MSE to calculate the offset of the center point coordinates relative

TABLE 1: Comparative experimental results.

Model	Prohibit (%)	Mandatory (%)	Warning (%)	mAP (%)
SSD300+VGG16	60.28	68.48	71.48	66.75
SSD512+VGG16	65.98	63.88	77.41	69.09
Faster RCNN	76.37	64.89	77.98	73.08
YOLOv3	74.54	61.02	76.59	70.72
YOLOv3+PACM	70.75	68.23	74.38	71.12
YOLOv3+PACM+CAFPN	73.27	67.26	83.44	74.66
YOLOv3+PACM+CAFPN+DIoU	76.54	69.33	80.24	75.37
YOLOv3+PACM+CAFPN+DCPIoU	76.33	69.62	82.11	76.02

to the grid and the translation and scaling of the anchor box; the confidence and classification both use cross-entropy as the loss function. The loss_{box} , loss_{obj} , and loss_{cls} are as specific as formulas (1) to (3). The full loss function calculation formula is shown in Equation (4):

$$\begin{aligned} \text{loss}_{\text{box}} = & \lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^B I_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i^j)^2 + (y_i - \hat{y}_i^j)^2 \right] \\ & + \lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^B I_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i^j} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i^j} \right)^2 \right], \end{aligned} \quad (1)$$

$$\begin{aligned} \text{loss}_{\text{obj}} = & - \sum_{i=0}^{s^2} \sum_{j=0}^B I_{ij}^{\text{obj}} \left[\hat{C}_i^j \log(C_i^j) + (1 - \hat{C}_i^j) \log(1 - C_i^j) \right] \\ & - \lambda_{\text{noobj}} \sum_{i=0}^{s^2} \sum_{j=0}^B I_{ij}^{\text{noobj}} \left[\hat{C}_i^j \log(C_i^j) + (1 - \hat{C}_i^j) \log(1 - C_i^j) \right], \end{aligned} \quad (2)$$

$$\text{loss}_{\text{cls}} = - \sum_{i=0}^{s^2} I_{ij}^{\text{obj}} \sum_{c \in \text{classes}} \left[\hat{P}_i^j \log(P_i^j) + (1 - \hat{P}_i^j) \log(1 - P_i^j) \right], \quad (3)$$

$$\text{Loss} = \text{loss}_{\text{box}} + \text{loss}_{\text{obj}} + \text{loss}_{\text{cls}}, \quad (4)$$

where x_i , y_i , w_i , and h_i are the predicted values of the horizontal and vertical offsets of the prediction box centroids relative to the grid and the translations and scaling of the anchor boxes, respectively; $\hat{x}$, $\hat{y}$, $\hat{w}$, and $\hat{h}$ are the true values, λ_{coord} is $2 - (\hat{w} \times \hat{h})$, I_{ij}^{obj} is the true confidence level, s^2 is the number of grids, B is the number of predefined anchor boxes, C and P are the confidence and category probabilities of grid prediction, and $\hat{C}$ and $\hat{P}$ are their true values.

The DIoU loss function has achieved satisfactory results in various models [24], and the DIoU function is calculated as shown in

$$L_{\text{DIoU}} = 1 - \text{IoU} + \frac{\rho^2(O_1, O_2)}{AB^2}, \quad (5)$$

where IoU is the intersection ratio between the prediction box and the true box, AB is the diagonal length of the smal-

TABLE 2: FPS comparison table.

Model	FPS
SSD300	38
SSD512	20
Faster RCNN	7
YOLOv3	24
YOLOv3+PACM	25
YOLOv3+PACM+CAFPN	21

lest rectangle containing the prediction box and the true box, and $\rho^2(O_1, O_2)$ is the square of the distance between the center point of the prediction box and the true box. The point location is shown in Figure 3.

To further improve the positioning loss function of the anchor box, this study added the penalty term of diagonal length difference to DIoU to DCPIoU. The formula for calculating DCPIoU is shown in

$$L_{\text{DCPIoU}} = 1 - \text{IoU} + \frac{\rho^2(O_1, O_2)}{AB^2} + \sqrt{(AB - CD)^2}, \quad (6)$$

where IoU is the intersection ratio between the prediction box and the true box, AB is the diagonal length of the smallest rectangle containing the prediction box and the true box, CD is the diagonal length of the intersection rectangle between the prediction box and the true box, and $\rho^2(O_1, O_2)$ is the square of the distance between the center point of the prediction box and the true box.

The modified DCPIoU is more generalized than DIoU. In the scene shown in Figure 4, the prediction box is surrounded by the true box, and the center point is closer. Currently, the $\rho^2(O_1, O_2)/AB^2$ penalty term in DIoU is very small, and the effect on weight update is weak, so DIoU defaults to IoU. Obviously, the gap between the prediction box and the true box is still large, and the prediction box does not contain the complete object. Furthermore, the DCPIoU adds a penalty term for the diagonal length of the two boxes, which can still effectively update the weights and improve the convergence, as shown in Figure 4.

FIGURE 5: Recognition effect of algorithms: (a) YOLOv3; (b) YOLOv3+PACM+CAFPN+DIoU; (c) YOLOv3+PACM+CAFPN+DCPIoU.

FIGURE 6: Recognition effect of algorithms: (a) YOLOv3; (b) YOLOv3+PACM+CAFPN+DIoU; (c) YOLOv3+PACM+CAFPN+DCPIoU.

3. Results and Discussion

The difficulty of traffic sign recognition lies in the broad range of object areas in the image, which requires high model robustness, especially for the model's small object extraction. One-stage models, such as YOLO and SSD, have fast recognition speed but poor ability to recognize small objects. The PACM and CAFPN structures proposed in this paper enhance the attention of the model to small objects in terms of feature extraction and feature fusion, respectively, while the DCPIoU is proposed to improve model convergence. To verify the feasibility of the proposed improvements, this study constructed a hybrid dataset. The experimental dataset is selected from the open-source China traffic sign detection dataset (CCTSDb) and Tingshua100K dataset, containing 4000 images of small objects. The data are divided into three categories according to the CCTSDb labeling method: prohibitory, warning, and mandatory. Then, the dataset is randomly divided into a training set and a test set in the ratio of 9:1. The same training set and training strategy are used for all categories. The experimental host configuration has an i7-8700 Intel CPU, 16 GB of memory, and a GTX1080 graphics card. AP and mAP evaluation metrics were used to measure model performance, and IoU is set to 0.5. Our improved model is compared with SSD equipped with VGG16 and Faster RCNN, and the experimental results are shown in Table 1.

Table 1 shows that the models with different improvement modules had different degrees of improvement over SSD300, SSD512, and YOLOv3; some of them outperformed Faster RCNN. The models with only PACM

improved 4.37%, 2.03%, and 0.4% over SSD300, SSD512, and YOLOv3, respectively. The models with PACM and CAFPN improved 7.91%, 5.57%, 1.58%, and 3.94% compared to SSD300, SSD512, Faster RCNN, and YOLOv3, respectively. The model with PACM, CAFPN, and DIoU improved 8.62%, 6.28%, 2.29%, and 4.65% over SSD300, SSD500, Faster RCNN, and YOLOv3, respectively. The models with PACM, CAFPN, and DCPIoU show 9.27%, 6.93%, 2.94, and 5.3% improvements over SSD300, SSD512, Faster RCNN, and YOLOv3, respectively. The model with DCPIoU added improved 0.65% over the one with DIoU.

In order to compare the recognition speed of models, FPS of the above models are compared in this paper, and the comparison results are shown in Table 2.

Through comparison, it can be found that the algorithm in this paper can achieve recognition speed comparable to that of the original YOLOv3 model while improving its performance.

In order to represent the advantages of the model more effectively, we selected small object images from CCTSDb and Tingshua100K datasets for testing and drawing attention heat maps, which were not involved in model training.

Figures 5–7 show that the improved model structure of this paper has improved the detection ability of small objects compared with the original YOLOv3. The three traffic signs in Figure 5 are of moderate size, and both YOLOv3 and the improved model of this paper can recognize them accurately, but it can be seen from Figure 6 that the original YOLOv3 model appears to miss detection when the

FIGURE 7: Attention heat map: (a) YOLOv3; (b) YOLOv3+PACM+CAFPN.

recognition objects gradually become smaller. As can be seen in Figure 7, compared with YOLOv3, the PACM and CAFPN structures proposed in this paper can improve the model's attention to the traffic sign area, and the more focused attention has a significant impact on the final recognition effect.

4. Conclusions

To solve the difficult problem of recognizing small traffic signs, this paper proposes PACM and a CAFPN for feature extraction and fusion of extracted features, respectively. DCPIoU is proposed as a new localization loss function based on DIOU. In order to verify the actual performance of the above improved algorithm, 4000 pictures matching the small object characteristics were selected from two open source datasets, CCTSDB and Tingshua100K, to construct a hybrid dataset, and the hybrid dataset was randomly divided into a training set and a test set as experimental data according to the proportion. Then, the improved model of this paper is trained on the divided training set with SSD, YOLOv3, and Faster RCNN models, and the same test set is used for testing after the training. The test results show that the improved method proposed in this paper improves the performance of the model to different degrees; compared with the original loss and DIOU, the improved DCPIoU loss function can improve the convergence of the model, and the mAP is higher with the same training set and training strategy.

Data Availability

The data used to support the findings of the study are available from the corresponding author upon request.

Conflicts of Interest

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

Authors' Contributions

He Huang did the conceptualization, investigation, and writing—original draft. Qice Liang did the conceptualization, investigation, and writing—review and editing. Dean Luo did the conceptualization and investigation. Dong Ha Lee did conceptualization and investigation.

Acknowledgments

This work was funded by the National Key Research and Development Program of China (no. 2017YFB0503702).

References

- [1] R. Qian, B. Zhang, Y. Yue, and F. Coenen, "Traffic sign detection by template matching based on multi-level chain code histogram," in *12th International Conference on Fuzzy Systems and Knowledge Discovery*, Zhangjiajie, China, August. 2015.
- [2] P. Douville, "Real-time classification of traffic signs," *Real-Time Imaging*, vol. 6, no. 3, pp. 185–193, 2000.
- [3] M. Liang, M. Yuan, X. Hu, J. Li, and H. Liu, "Traffic sign detection by ROI extraction and histogram features-based recognition," in *The 2013 international joint conference on Neural networks*, Dallas, TX, USA, August. 2013.
- [4] 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*, pp. 779–788, Honolulu, HI, USA, 2016.
- [5] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6517–6525, Honolulu, Hawaii, 2017.
- [6] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, <https://arxiv.org/abs/1804.02767>.
- [7] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: optimal speed and accuracy of object detection," 2020, <https://arxiv.org/abs/2004.10934>.
- [8] W. Liu, D. Anguelov, D. Erhan et al., "Ssd: single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Cham, 2016.

- [9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, Columbus, Ohio, 2014.
- [10] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, Santiago, Chile, 2015.
- [11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, pp. 91–99, 2015.
- [12] S. K. Tai, C. Dewi, R. C. Chen, Y. T. Liu, X. Jiang, and H. Yu, "Deep learning for traffic sign recognition based on spatial pyramid pooling with scale analysis," *Applied Sciences*, vol. 10, no. 9, p. 6997, 2020.
- [13] H. Wan, L. Gao, M. Su, Q. You, H. Qu, and Q. Sun, "A novel neural network model for traffic sign detection and recognition under extreme conditions," *Journal of Sensors*, vol. 2021, 16 pages, 2021.
- [14] Y. Cai, T. Luan, H. Gao et al., "YOLOv4-5D: an effective and efficient object detector for autonomous driving," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, no. 99, pp. 1–13, 2021.
- [15] Z. Liu, D. Li, S. S. Ge, and F. Tian, "Small traffic sign detection from large image," *Applied Intelligence*, vol. 50, no. 1, pp. 1–13, 2020.
- [16] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, Honolulu, Hawaii, 2017.
- [17] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8759–8768, Salt Lake City, Utah, 2018.
- [18] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: scalable and efficient object detection," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. , 202010781–10790, 2020.
- [19] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.
- [20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, Salt Lake City, Utah, 2018.
- [21] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: convolutional block attention module," in *Proceedings of the European conference on computer vision*, pp. 3–19, Munich, Germany, 2018.
- [22] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3560–3569, 2021.
- [23] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: a metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 658–666, Long Beach, CA, 2019.
- [24] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: faster and better learning for bounding box regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34no. 7, pp. 12993–13000, New York, USA, 2020.