2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE 2021)

An Improved Lightweight Network MobileNetv3 Based YOLOv3 for Pedestrian Detection

Xiaxia Zhang, Ning Li* and Ruixin Zhang
Nanjing University of Aeronautics and Astronautics (NUAA)
Nanjing, China
xiamu@nuaa.edu.cn, lnee@nuaa.edu.cn, 374169703@qq.com

2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE) | 978-1-7281-8319-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCECE51280.2021.9342416

***Abstract*—Recently, most object detection under videos have increasingly relied on the Unmanned Aerial Vehicle (UAV) platforms because of UAVs' timeliness, pertinence, and high flexibility in data acquisition. Convolution neural networks, especially for YOLO v3, have proved to be effective in intelligent pedestrian detection. However, two problems need to be solved in pedestrian detection of UAV images. One is more small pedestrian objects in UAV images; the other is the complex structure of Darknet53 in YOLO v3, which requires massive computation. To solve these problems, an improved lightweight network MobileNetv3 based on YOLO v3 is proposed. First, the improved MobileNetv3 takes place of the Darknet53 for feature extraction to reduce algorithm complexity and model simplify. Second, complete IoU loss by incorporating the overlap area, central point distance and aspect ratio in bounding box regression, is introduced into YOLO v3 to lead to faster convergence and better performance. Moreover, a new attention module SESAM is constructed by channel attention and spatial attention in MobileNetv3. It can effectively judge long-distance and small-volume objects. The experimental results have shown that the proposed model improves the performance of pedestrian detection of UAV images.**

Keywords-Pedestrian Detection; YOLO v3; MobileNetv3; SESAM; CIoU

I. INTRODUCTION

Pedestrian detection, as one of the most challenging research topics in computer vision, plays a critical role in various fields such as intelligent surveillance[1], traffic safety [2] and service robots [3]. Since UAVs' advantages of wide detection range and high flexibility, more and more people have applied them to object detection in recent years, especially in pedestrian detection [4-7]. At present, pedestrian detection methods based on deep learning are the most popular. It can be divided into two categories: 1) Two-stage detection algorithm based on regions, such as R-CNN [8], Fast R-CNN [9], Faster R-CNN [10], etc. These algorithms first choose a region proposal network (RPN) to obtain candidate object information, and then predict the category and location information by the detection network. Many researchers [11-14] have used these method for pedestrian detection and achieved good results. Although the two-stage detection algorithms based on regions have high precision, it is difficult to meet the real-time requirements due to the need to go through two stages of candidate objects extraction and reclassification. Moreover, the quality of the extracted candidate objects also has a significant impact on the model detection effect. 2) One-stage detection algorithm based on regression including SSD [15], YOLO series [16-18], etc. These algorithms diametrically extract features from the network to predict object classes and location [19], and have a faster detection speed.

In fact, most detection methods are difficult to simultaneously consider the real-time and accuracy of the pedestrian detection. Aiming at the problem, Zhang et al. [20] proposed to use an RPN followed by boosted forests in Faster R-CNN to detection pedestrians. On the INRIA dataset, the method achieved a miss rate of 9.6%. Fan et al. [21] made improvements based on the YOLOv3. They integrated label smoothing to reduce the degree of over-fitting and added multiple scale detection to improve the detection accuracy. The mAP reached 91.68% on the Caltech dataset. Zheng Dong et al. [22] introduced MobileNetv2 instead of VGG as the basic network to perform feature extraction in order to reduce the model size and speed up the detection process. On KITTI dataset, the detection speed increased by 13fps, mAP increased by 1.3%~1.6%. Although the above methods have achieved good detection results on public datasets, they are not suitable for pedestrian detection under UAV because the background and targets faced on UAV video images are more complex and changeable.

In this paper, we choose YOLOv3 model as the base architecture, which has a better comprehensive performance in terms of speed and accuracy, and make some improvements to address the problems of more long-distance small objects under UAV and limited computing power of UAV airborne platform:

1) Replace the Darknet53 with the lightweight network Mobilenetv3 as backbone to reduce algorithm complexity and model simplify. For the problem that long-distance and small-volume objects easily to be undetected, construct a new attention module, SESAM module.

2) Use the CIoU loss instead of the IoU loss in YOLOv3 to achieve faster convergence and better performance. CIoU loss incorporates the overlap area, normalized distance, aspect ratio between the predicted box and the ground-truth.

II. PROPOSED METHOD

In this section, we mainly introduce our improvements including the improved Mobilenetv3 for feature extraction and CIoU loss. The structure diagram of pedestrian detection model based on improved MobileNetv3 is shown in Fig. 1(a). First, use the improved MobileNetv3 for feature extraction. Where the SESAM module is aimed to focusing on important features and suppressing unnecessary ones in the channel and spatial dimensions. Then, make predictions using the fusion of multiple scales.

978-1-7281-8319-0/21/$31.00 ©2021 IEEE

Fig. 1(b) shows the detailed structure of each module. DBL is the basic component of the yolov3 model, which is convolution+BN+Leaky relu. Res represents the residual block. CNB is the convolution block. "bneckF" represents a linear bottleneck module without attention module. "bneckT" represents a linear bottleneck module with SESAM attention module. The number behind of res, CNB, bneckF and bneckT indicates the number of the modules. DW means depthwise separable convolution, BN means Batch Normalization.

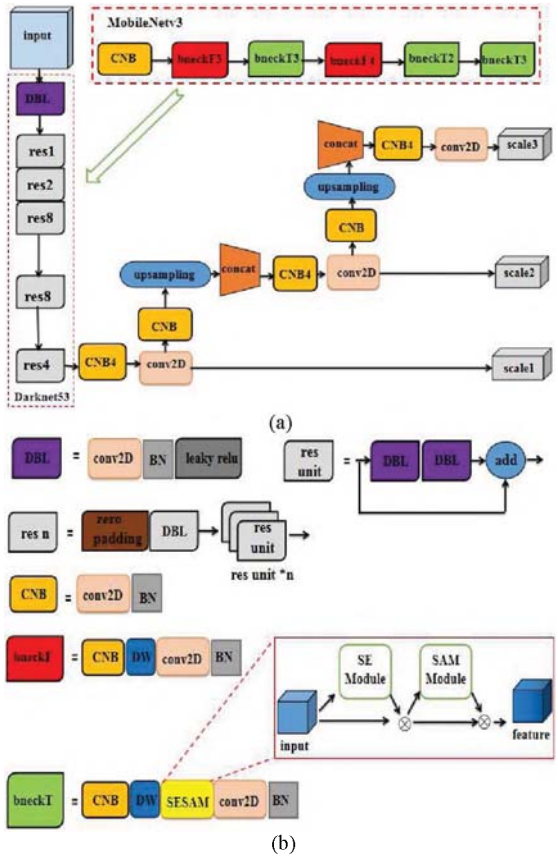


Figure 1. The network structure (a)The pedestrian detection model structure based on improved MobileNetv3 (b)The detailed structure of modules in Fig.1(a)

A. The improved MobileNetv3

While MobileNetv3 instead of Darknet53 for feature extraction reduce the complexity of model, it's easy to miss long-distance and small pedestrians. That's probably because the SE module that only learns the correlation between channels and ignores the spatial characteristics. Studies have shown that the use of spatial attention modules can effectively judge long-distance and small-volume objects [23]. The main reason why this method is effective is that each element of the feature map corresponds to an area of the original picture, and assigning different weights to each position of the output feature map is equivalent to applying different influence factors to different areas of the original picture. So, we constructed a new attention module called SESAM (Squeeze-and-Excitation and Spatial Attention Module) to substitute the SE module in MobileNetv3. SESAM is a module that combines the channel attention and the spatial attention, it can focus on important features from the channel and spatial dimensions respectively. The specific structure diagram is shown in Fig. 2.

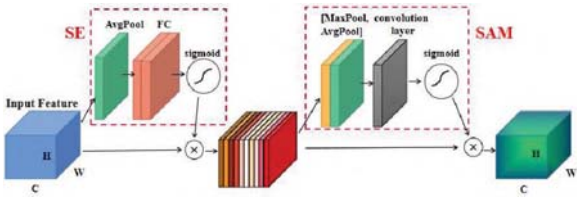


Figure 2. The basic structure of the SESAM module.

In the SE module, first perform global average pooling on a $H \times W \times C$ feature map to obtain a $1 \times 1 \times C$ feature vector, which reflects the global distribution of each feature response in the channel dimension. Then the feature vector is passed through two fully connected layers. The first fully connected layer has C/16 neurons, which is equivalent to reducing the dimension of C features, and the number of neurons in the second fully connected layer is C. It is equivalent to increasing the dimension to C features. The advantage of using two fully connected layers is that the complex correlation between channels can be better fitted, and the amount of parameters and calculations can be greatly reduced. After then through a sigmoid layer, a $1 \times 1 \times C$ feature map is obtained. Finally, the input feature map of $H \times W \times C$ and the feature map of $1 \times 1 \times C$ are fully multiplied to obtain the channel attention feature map. The main function of the SE structure is to selectively focus on feature channels with useful information and suppress useless feature channels.

Let the output of the SE module as the input of the SAM module. We first apply the global max pooling and global average pooling operations among the channel dimension and obtain two $H \times W \times 1$ feature maps, and concatenate the two feature maps to obtain one $H \times W \times 2$ feature map. Then the feature is further extracted through a convolution operation, and the size of the feature map is $H \times W \times 1$. After that, the spatial attention feature map is normalized by sigmoid function. The main purpose of the SAM module is to give more weight to the pixels that contain important information on a channel. Finally, multiply the spatial attention feature map with the input feature map of the SESAM module to get the final feature.

B. CIoU loss

IoU is the most popular metric in terms of evaluation metric for bounding box regression. It is commonly used to measure the similarity between two arbitrary shape objects and is scale-invariant. However, when it's used as a loss function, the following problems may occur: (1) If there is non-overlapping between the predicted box and the ground-truth, IoU=0, it cannot reflect the distance between the predicted box and the ground-truth. When the predicted box and the ground-truth are completely coincident, loss=0, there is no gradient return, and the training cannot be performed. (2) Cannot accurately reflect the intersection of predicted box and the ground-truth. As shown in Fig. 3, the IoU values of the following three cases are the same, but they intersect in different ways, with (a) having the best regression effect and (c) having the worst.

To solve the above problems, CIoU loss [24] was introduced. It incorporates the overlap area, the Euclidean distance between the center of the predicted box and the ground-truth, the aspect ratio in bounding box regression. CIoU loss is defined in the following formulas:

$$L_{CIoU} = 1 - CIoU \tag{1}$$

$$CIoU = IoU - \rho^2 c^{-2} - \alpha v \tag{2}$$

In the above equations, ρ is the Euclidean distance between the center of the predicted box and the ground-truth. c represents the diagonal distance of the the smallest box that contains both the predicted box and the ground-truth. The penalty term $\rho^2 c^{-2}$ can directly minimize the distance between the center of predicted box and the center of ground-truth. The penalty term αv makes the aspect ratio of the predicted box and the ground-truth consistent, thereby speeding up the convergence. α and v can be formulated as:

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{3}$$

$$v = \frac{4}{\pi^2}(\arctan\frac{\omega^{gt}}{h^{gt}} - \arctan\frac{\omega}{h})^2 \tag{4}$$

Where ω^{gt} and h^{gt} are the width and height of the ground-truth respectively. ω and h respectively represent the width and height of the predicted box.

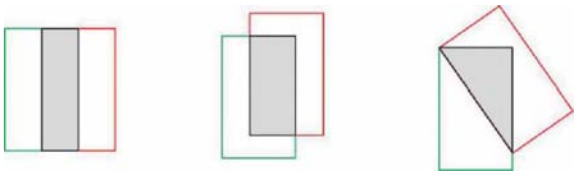


Figure 3. IoU are equal, and the predicted box and the ground-truth intersect in different ways.

III. EXPERIMENTS

A. Datasets

The dataset used in this paper was obtained by the DJI Yu PRO UAV in sunny weather with no wind. The size of the images is 3840×2160. For better comparing the performance of different detection models, the final pedestrian dataset was made into Pascal VOC format. Each image in the dataset corresponds to a label file, which indicates the name, size, path of the image, the category of the object, and the coordinates of the object's circumscribed rectangle. The dataset contained 1761 images, and the number of pedestrians is 17219. According to the ratio of 7:2:1, the dataset was divided into training set, test set and verification set.

B. Training Details

The models in this paper are trained on the NVDIA 1080Ti GPU.

In training, we first used three image augmentation techniques to expand the dataset, including flip, translation transformation and color jittering. Then, we trained the network for 200 epochs with 0.9 momentum and 0.0005 weight decay, set every 8 samples as one processing unit, used batch normalization to normalize each weight update. We chose Adam optimizer with the initial learning rate of 0.001. To avoid over-fitting, the learning rate was adaptively adjusted. Furthermore, the IoU threshold and the confidence threshold were both 0.5.

C. Performance Evaluation

In this part, we use our person dataset and adopt YOLOv3 as the base architecture to empirically show the effectiveness of our design model. We first verified the validity of the CIoU bounding box loss function, then considered three combinations of SE and SAM attention modules, including SE-SAM, SAM-SE, SE and SAM in parallel. We mainly evaluate the performance of pedestrian detection quantitatively through precision, recall, average precision(AP) and frames per second(fps):

$$precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{5}$$

$$recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{6}$$

$$AP_{11\,point} = \frac{1}{11} * (\sum x)(x \in Max\,\Pr ecision) \tag{7}$$

Table I shows the influence of different improved methods on the model's test results. We can find that using CIoU loss as the bounding box regression loss function for pedestrian detection can improve the model performance in a certain degree. The recall is increased by 3.17%, the precision is increased by 4.89%, and the AP is 82.60%. But the detection speed is slightly lower than the original model. When replacing the Darknet53 with the lightweight network MobileNetv3 for image feature extraction, it accelerates procession of the model while ensuring the improvement of the model detection accuracy. The number of detection frames per second has been increased from 16 to 23. The size of model weight file is reduced from 235MB of Model1 to 32.5MB. Comparing the results of three models on different attention arranging methods, it can be found that introducing the SAM module into MobileNetv3 can make up for the lack of only using channel attention and SE-SAM performs slightly better than SAM-SE or SAM and SE in parallel. Therefore, we arrange the channel and spatial modules sequentially and CIoU loss as the bounding box regression loss in our final model.

TABLE 1 THE INFLUENCE OF DIFFERENT IMPROVED METHODS ON THE MODEL'S TEST RESULTS.

Model		Precision/%	Recall/%	AP/%	fps
YOLO v3(based model)		92.32	79.68	75.84	18
Our Models	YOLO v3+CIoU	97.21	82.85	82.60	16
	YOLO v3+MobileNetv3(SE)+CIoU	93.69	86.16	85.56	23
	YOLO v3+MobileNetv3(SE& SAM in parallel)+CIoU	93.47	86.29	85.98	22
	YOLO v3+MobileNetv3(SAM-SE)+CIoU	93.57	85.98	85.43	23
	YOLO v3 +MobileNetv3(SE-SAM)+CIoU	**93.43**	**87.17**	**86.62**	**23**

In order to thoroughly evaluate the effectiveness of the improved model in video images under UAV, two-stage detection algorithm (Faster R-CNN) and one-stage detection algorithm (SSD) were selected for comparative experiments. The detection results of different models on test set are shown in Table II.

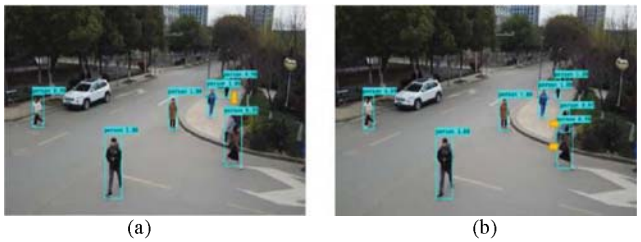
TABLE II THE DETECTION RESULTS OF DIFFERENT DETECTION MODELS ONOUR PERSON DATASET.

Model	Precision/%	Recall/%	AP/%	fps
Faster R-CNN	91.39	88.32	86.04	8
SSD	88.07	76.15	70.36	34
YOLO v3	92.32	79.68	75.84	18
Improved Model	**93.43**	**87.17**	**86.62**	**23**

From the Table II, it can be found that comparing with YOLO v3, the AP of Faster R-CNN is higher, reaching 86.04%. This is probably because Faster R-CNN use regional candidate networks to obtain information about candidate objects. But it also brought huge calculations, the procession speed is only 8fps, which is about half of the YOLO v3 model. For SSD, the procession speed reached 34fps, which is much faster than YOLO v3 model, but its AP is only 70.36%. The performance of our improved model is better, the AP reached 86.62%, which is 10.78% higher than that of YOLO v3 and 0.58% higher than that of Faster R-CNN. And it detection speed is faster than YOLOv3 and Faster R-CNN, reached 23fps. This is mainly because MobileNetv3 is based on the depthwise separable convolution. When the convolution kernel size is 3×3, the computation of depthwise separable convolution is only one-ninth of the standard convolutions, which effectively reduces the amount of model computation. The h-swish activation function in MobileNetv3 also consumed less computing resources in the case of improving the accuracy of the neural network.

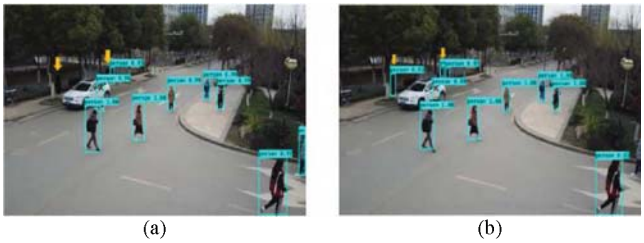
D. Experimental Results

In order to intuitively reflect the detection effect of the model, we respectively visualized the detection results of original model YOLOv3 and the improved model in three poor environmental conditions, such as occlusion between pedestrians, pedestrians in poor illumination, and smaller pedestrians. Fig. 4 is the pedestrian detection results when there is occlusion between pedestrians. YOLO v3 model can easily identify multiple people as one person, but the improved model can be detected relatively accurately. Fig. 5 and Fig. 6 are the pedestrian detection results for pedestrians with poor illumination or smaller scale. These pedestrians undetected by YOLOv3 can be effectively detected in our improved model. In summary, under the complex and changeable situation, the improved model can effectively overcome the unfavorable factors and has stronger generalization ability, and better detection effect.



(a) (b)

Figure 4. Detection results when there is occlusion between pedestrians (a)YOLOv3 (b) Improved model.



(a) (b)

Figure 5. Detection results when pedestrians are in poor illumination (a)YOLOv3 (b) Improved model.

(a) (b)

Figure 6. Detection results when pedestrians are smaller (a)YOLOv3 (b) Improved model.

IV. CONCLUSION

In this paper, we propose an improved MobileNetv3 based on YOLOv3 to improve the precision and procession speed on pedestrian detection under UAV. First, we use the lightweight network MobileNetv3 to replace the Darknet53 for feature extraction. The depthwise separable convolution and h-wish activation function used in MobileNetv3 effectively reduce the amount of model parameters. Because there are many long-distance small objects under UAV, the spatial attention module SAM is introduced into the MobileNetv3. The SE module selectively focuses on feature channels with useful information, while the SAM module focuses on which part of the input image has more effective information, it can effectively judge long-distance small objects. Second, we introduced CIoU loss to make the bounding box regression more stable. To verify the efficacy of the improved model, we conducted comparative experiments with some state-of-the-art models and confirmed that our improved model outperforms on the pedestrian detection task under UAV. In the future, we still need to improve the detection accuracy and simplify the model.

ACKNOWLEDGMENTS

This work received support from Science and Technology on Electro-optic Control Laboratory and Aviation Science Foundation Project (ASFC-20175152036) and Key Project on Artificial intelligence (1004-56XZA19008). The authors are also grateful for the support of their colleagues at the Key Laboratory of Radar Imaging and Microwave Photonics, Ministry of Education.

REFERENCE

[1] Y.F. Zou, "Study of pedestrian detection and tracking in intelligent video surveillance system," University of Science and Technology of China, 2011.

[2] Z. Min, M. Ying, S. Dihua, "Tunnel pedestrian detection based on super resolution and convolutional neural network," 2019 Chinese Control And Decision Conference (CCDC). 2019.

[3] P.F. Li, K.F. Xi, "Multi-People detection and tracking for service robots," Computer Systems and Application, 2016, 25(10): 252-257.

[4] T. Giitsidis, E.G. Karakasis,A. Gasteratos, G. Ch. Sirakoulis, "Human and fire detection from high altitude UAV images," Euromicro International Conference on Parallel. 2015.

[5] H.T. Ma, Y.J. He, C.M. Wang, M.Z. Yu, J.A. Wang, "Research on human detection and tracking technology based on UAV vision," Computer Technology and Development, 2018, 28(10): 115-118.

[6] Q.Q. Guo, "Pedestrian detection in UAV scene based on convolutional neural network," Dalian University of Technology, 2019.

[7] C. Xiang, H.C. Shi, N. Li, M. Ding, "Pedestrian detection under unmanned aerial vehicle an improved single-stage detector based on retinanet," 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). 2019.

[8] R. Girshick, J. Donahue, T. Darrell,J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 27th IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580-587.

[9] R. Girshick, "Fast r-cnn," IEEE International Conference on Computer Vision, 2015, pp. 1440-1448.

[10] S. Ren, K. He, R. Girshick, J. Sun,) "Faster r-cnn: towards real-time object detection with region proposal networks," 29th Annual Conference on Neural Information Processing Systems, 2015, pp. 91-99.

[11] J.A. Li, X.D. Liang, S.M. Shen, T.F. Xu,J.S. Feng, S.H. Yan, "Scale-aware fast R-CNN for pedestrian detection," IEEE Transactions on Multimedia, 2017.

[12] Z.Q. Zhao, H.M. Bian, D.H. Hu, W.J. Cheng,H. Glotin, "Pedestrian detection based on fast R-CNN and batch normalization," International Conference on Intelligent Computing, 2017.

[13] K. Che, Z.T. Xiang, Y.F. Chen, (2018) "Research on infrared image pedestrian detection based on improved fast R-CNN," Infrared Technology, 2018, 40: 578-584.

[14] W.Y. Yao, J.P. Li, "Pedestrian detection algorithm based on improved faster R-CNN," Science Technology and Engineering, 2020, 20: 1498-1503.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, et al. "Ssd: single shot multibox detector," 14th European Conference on Computer Vision,2016, pp. 21-37.

[16] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You Only Look Once: unified, real-time object detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition. Seattle, 2016, pp. 779-788.

[17] J. Redmon, A. Farhadi, "YOLO9000: better, faster, stronger," Computer Vision and Pattern Recognition, 2017, pp. 6517-6525.

[18] J. Redmon, A. Farhadi, "YOLOv3: an incremental improvement," Computer Vision and Pattern Recognition, 2018.

[19] Y.Y. Qi, H.C. Shi, N. Li, Y.H. Li, "Vehicle detection under unmanned aerial vehicle based on improved YOLOv3," 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2019.

[20] L.L. Zhang, L. Lin, X.d. Liang, K.M. He, "Is faster R-CNN doing well for pedestrian detection?" European Conference on Computer Vision, 2016.

[21] L. Fan, B. Su, Y.H. Wang, "Improved real-time pedestrian detection algorithm based on YOLOv3," Journal of Shanxi University(Natural Science Edition), 2019, 42: 709-717.

[22] D. Zheng, X.Q. Li, X.Z .Xu, "Vehicle and pedestrian detection model based on lightweight SSD," Journal of Nanjing Normal University(Natural Science Edition), 2019, 42: 73-81.

[23] J.H. Liu, G.P. Hu, S.Y. Wang, "Fine-grained detection of railway track pedestrian based on deep learning," Computer & Digital Engineering, 2020, 48: 1367-1371.

[24] Z. Zheng, P. Wang, W. Liu, J.Z. Li, D.W. Ren, "Distance-IoU loss: faster and better learning for bounding box regression," AAAI Conference on Artificial Intelligence, 2020.