Research Article

Chuanyun Xu, Yu Zheng*, Yang Zhang, Gang Li, and Ying Wang

A method for detecting objects in dense scenes

https://doi.org/10.1515/comp-2022-0231
received February 16, 2021; accepted February 1, 2022

Abstract: Recent object detectors have achieved excellent performance in accuracy and speed. Even with such impressive results, the most advanced detectors are challenging in dense scenes. In this article, we analyze and find the reasons for the decrease in detection accuracy in dense scenes. We started our work in terms of region proposal and location loss. We found that low-quality proposal regions during the training process are the main factors affecting detection accuracy. To prove our research, we established and trained a dense detection model based on Cascade R-CNN. The model achieves an accuracy of mAP 0.413 on the SKU-110K sub-dataset. Our results show that improving the quality of recommended regions can effectively improve the detection accuracy in dense scenes.

Keywords: dense detection, dense scenes, cascade R-CNN

1 Introduction

Computer vision research has achieved great success with the development of deep convolutional neural networks [1–4] and large-scale datasets [5,6,29]. Object detection is one of the most important research directions of computer vision, and with the wide application of computer vision technology, more and more appear in daily life. The recent object detection algorithms for deep learning [7–13] can quickly and reliably detect objects in real scenes.

In object detection, one-stage and two-stage detectors usually follow a general training paradigm, that is, sampling the region, and then performing recognition and positioning under the action of a multi-task objective function. Under this working paradigm, high-quality regional samples are one of the keys to a deep learning object detector.

In a crowded and dense scene, there are a large number of adjacent objects in the image, and these objects all look very similar or even the same. Similar scenarios are abundant in real life, such as retail shelves, traffic or city landscape maps. But in the existing object detection baseline, their effect is not very satisfactory.

It can be seen from Figure 1 that dense objects bring obvious problems to object detection: (1) The anchor box between the adjacent object is prone to overlap, and it is impossible to accurately determine the positive and negative samples through IoU. (2) Adjacent and similar objects make the detector unable to accurately determine the start and end positions of an object. (3) There are a large number of detection objects in the image, which will result in the existence of many small objects. (4) The imbalance between regression loss and classification loss may affect network training.

From the aforementioned problems, compared with ordinary detection tasks, the main difficulty of dense detection lies in the region proposal stage. Providing high-quality regional samples to the detector is the primary task to improve the detection accuracy in dense scenes. For one-stage object detection algorithms such as Yolo [11,12,14,39], SSD [13], and Retinanet [10], abundant regression tasks lead to the imbalance of final recognition and positioning loss, which affects the training of the entire network. At the same time, it is compared with the two-stage detection algorithm. The one-stage detection algorithm has no region proposal stage, and cannot provide relatively fine bounding boxes to participate in the regression calculation, so the final bounding box may not be accurate enough. Therefore, we believe that the two-stage detection algorithm is more suitable for dense detection tasks.

* **Corresponding author: Yu Zheng,** College of Computer Science and Engineering, Chongqing University of Technology, Chongqing, China, e-mail: 543278005@qq.com
Chuanyun Xu: College of Computer Science and Engineering, Chongqing University of Technology, Chongqing, China, e-mail: 33677670@qq.COM
Yang Zhang: College of Computer and Information Science, Chongqing Normal University, Chongqing, China, e-mail: 495461428@qq.COM
Gang Li: College of Computer Science and Engineering, Chongqing University of Technology, Chongqing, China, e-mail: 364504496@qq.COM
Ying Wang: College of Computer Science and Engineering, Chongqing University of Technology, Chongqing, China, e-mail: 1508711395@qq.COM

Figure 1: A typical image in SKU-110K. Obviously, it can be seen that the objects in the picture are very adjacent and similar. Unlike previous detections, there are dozens of objects to be detected in the picture. At the same time, due to a large number of objects in a picture, a situation where a large number of small objects are formed, a large number of small and dense objects pose a challenge to the object detector.

We propose a method that can accurately detect objects even in dense scenes. By analyzing the problems in dense detection, Cascade R-NN [25] combined with FPN is used as a backbone network. At the same time, by analyzing the impact of regression loss on dense detection, experiments were carried out using CIOU Loss and balanced L1 loss, respectively.

To summarize, our novel contributions are as follows:

- Propose the difficulty in dense detection. Research from two aspects of regional proposal and regression loss.
- Established an end-to-end dense object detection model. Even in dense objects scenarios, it can still achieve industrial-grade detection accuracy.
- Verify our method on the SKU-110K [15] sub-dataset. Reach the detection accuracy of mAP 0.413.

2 Related work

Various object detection algorithms and related theories proposed in recent years are reviewed. The object detection algorithm can be divided into two steps based on deep learning. First, the deep neural network is used to extract image features, and then the object is identified and localized through the detection head. The task of object detection includes not only the accurate recognition of each object in the image, but also the acquisition of its position information [16].

Object detection: Early object detection used a sliding window method to apply a classifier to the content in each window in the entire picture [17–19]. Later, in order to improve the detection speed, the search space was first narrowed by region proposal, and then the corresponding region was classified using a complex classifier [20–22].

R-CNN [7] introduced deep learning to object detection for the first time. Nowadays, deep learning-based methods have become the mainstream method of object detection. It is mainly divided into two categories: fast R-CNN [8], SPP-Net [23], Faster R-CNN [9] as the representative of the two-stage algorithm, first stage obtains the candidate area, and the second stage is based on the candidate area to locate and classify the object. The one-stage algorithm represented by yolo and SSD [13] uses neural networks to directly locate and classify objects through the regulation of loss functions. On this basis, in order to pursue higher detection accuracy and detection speed, researchers have proposed other improved methods. For example, Cascade R-CNN [25], RetinaNet [10], Libra R-CNN [24], and yolov4 [11]. At the same time, in order to get rid of the dependence of the detector on the anchor and the problem of data imbalance in the anchor. CornerNet [26], Fcos [27], and other algorithms proposed the use of anchor free methods to detect objects.

Crowded scene: Now there are many excellent object detection datasets such as: PASCAL VOC [28], MS COCO [29], and ILSVRC [30]. However, the aforementioned dataset is shown in Figure 2, and the objects in the detection scene are relatively sparse. In recent years, some datasets of dense scenes have been proposed [31,32], but their purpose is to count rather than detect. Therefore, object detection in dense scenes remains to be studied. Recently, Goldman et al. [15]. introduced a Gaussian mixture clustering method based on the EM algorithm to solve the problem of detection overlap, and proposed the SKU-110K dense object detection dataset. In their recent work, Chu et al. [33] proposed a simple but effective method for dense object detection based on candidate boxes by using one-candidate boxes and multiple bounding boxes.

Our work in this article is to analyze the difficulty of object detection in dense scenes and provide research guidance for subsequent dense object detection. At the same time, we have established a general object detection model, which can accurately recognize in a relatively dense scene in time.

3 Dense detection method

The method in this article detects images of any size and dense scene. The network structure is shown in Figure 3,

Figure 2: The left image is the training picture in the COCO dataset. The right image is the picture in the PASCAL VOC dataset. The detection objects in the figure are relatively sparse and do not constitute a dense scene.

which is mainly composed of: (1) backbone network (feature extraction and FPN layer), (2) RPN (region proposal module), and (3) detection head. It is a complete end-to-end network model. As mentioned above, the two-stage object detector mainly involves two steps. The first step is to extract image features through the network to generate a bounding box. The second step is instance detection, which accurately detects the bounding box in the first step. In this article, we focus on the first step.

3.1 Region proposal

In the training process of the two-stage detection algorithm, the threshold of the IoU needs to be defined to determine the positive and negative samples. Training the detector with a low IoU threshold usually produces noisy boxes with multiple repeated bounding boxes around the object. However, if the IoU threshold is increased, the detection performance will tend to decrease, which will result in missed detection. The calculation method of IoU is as follows:

$$\mathrm{IoU} = \frac{A \cap B}{A \cup B}, \tag{1}$$

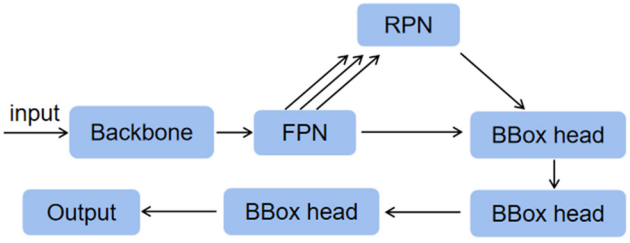

Figure 3: The main structure of the model. Improve the accuracy of dense object detection through the combination of cascade R-CNN and FPN.

where A represents the area of ground truth, and B represents the area of the anchor box. For dense detection, the discrimination of positive and negative samples is much more complicated than usual. In dense detection tasks, hard example is much higher than the usual detection scenarios. As shown in Figure 1, because the distance between the detected objects is too close, it is easy to appear in Figure 4(a)–(c) three wrong detection results.

Most of the negative samples in the R-CNN series of algorithms are filtered through two stages. However, in a dense object scene, a large number of detection objects will lead to a sharp increase in the number of proposal regions, and at the same time, lead to an extraordinary number of negative samples. It is still difficult to guarantee the quality of the recommended area only through the two-stage method of R-CNN.

Cascade R-CNN is extended based on Faster R-CNN. As shown in Figure 5, first obtain the feature map of the entire picture through R0, thereby generating the initial region proposal. The region proposal will cascade through three detection heads: H1, H2, and H3. Each

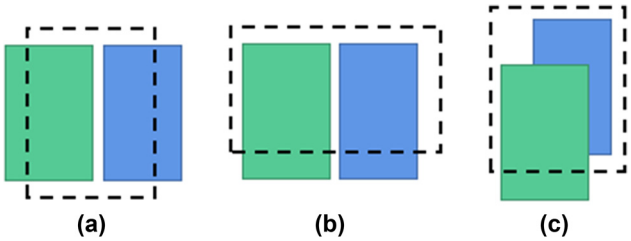

Figure 4: (a) Because the size of the anchor box is fixed, if the positive sample and the negative sample are judged according to a certain threshold, the detection box may be inaccurate due to the close distance between the objects. (b) If adjacent objects in a dense scene are very close and identical, the detector may not be able to accurately identify their positions. (c) In the case of occlusion, it also brings challenges to the accuracy of detection.

detection head will perform classification and border regression on the last incoming result.

The positive and negative samples are divided into different IoU thresholds in a cascade structure so that the detection head of each stage is focused on detecting the bounding box in a particular range. Because the closer the input IoU is to the set threshold, the learning effect of the detector will be better than other input IoU. By setting multiple thresholds, gradually increasing the size of the threshold to filter the cascade of bounding boxes. You can get a more accurate bounding box position for regression.

In dense scenarios, too many negative samples lead to the problem of inaccurate regression of the bounding box. A key issue is to define high-quality positive and negative samples. The Cascade R-CNN multi-threshold cascade mode can effectively improve the quality of the recommended area and reduce the impact of hard examples on the final detection result. On the other hand, by setting a suitable anchor, each detected object has a suitable anchor size for matching. Previously in yolov3 [14], the appropriate anchor size of the dataset was obtained by clustering to achieve the purpose of improving the detection accuracy. To improve the detection efficiency of dense scenes, Chu et al. [33] found a suitable bounding box through each detected object.

3.2 Loss function

In the dense detection task, the bounding box generated by the network during training will be much larger than the general detection task. It brings great challenges to the speed and accuracy of the bounding box regression. Therefore, there are two problems: (1) How to accurately and efficiently regress the bounding box. (2) Does a large amount of regression loss affect the overall training. The two problems raised by CIoU Loss [34] and Balanced L1 Loss verification are, respectively, passed.

The Cascade structure can generate high-quality suggestion regions, and fast and accurate regression bounding box is also an important method to improve detection efficiency. Therefore, in the process of bounding box regression, CIoU Loss is introduced as the loss function of bounding box regression:

$$\mathrm{CIoU} = 1 - \mathrm{Iou} - \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v. \tag{2}$$

where b, b^{gt} represents the center points of the bounding box and the center box, respectively, and ρ represents the calculation of the Euclidean distance between the two center points, and c^2 represents the diagonal length of the minor-closed interval that can contain both the bounding box and the ground truth box. α is a positive trade-off parameter. v is the similarity measure of the aspect ratio. CIoU loss is considered from the three critical factors of overlap area, center point distance, and aspect ratio. The bounding box regression can still be optimized when the predicted frame and the ground truth do not overlap. By directly minimizing the distance between the two bounding boxes, the convergence speed is faster than GIoU Loss. The aspect ratio is increased to make the result of the predicted frame more in line with the ground truth.

At the same time, because there are a large number of bounding box regression tasks, under the guidance of multi-task loss, the loss function of object detection is defined as:

$$L = L_{\mathrm{cls}} + \lambda L_{\mathrm{loc}}, \tag{3}$$

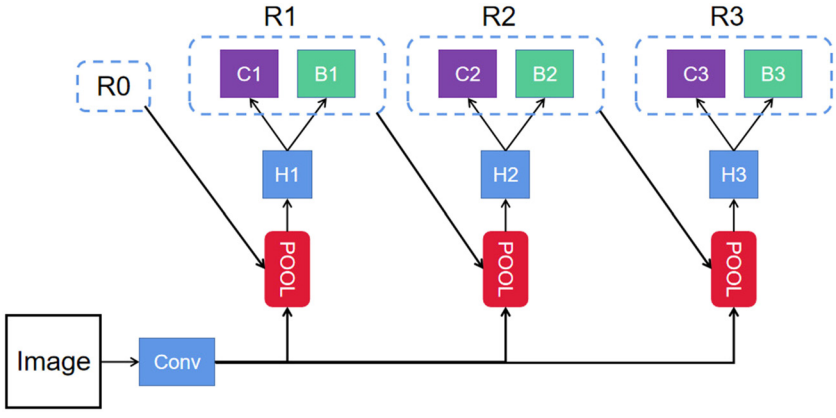

Figure 5: Cascade R-CNN model structure, through the cascaded detection head to improve the quality of regional proposal.

where L_{cls}, L_{loc} represent the loss of classification and positioning, respectively. λ represents the multitasking loss as a weight coefficient. Directly increasing the weight of the positioning loss will make the model more sensitive to outliers, which will produce excessive gradients, which is harmful to the entire training process. Considering the balance of classification loss and regression loss, promote the key regression gradient, and achieve more uniform training in classification and positioning. Take the Balanced L1 Loss in Libra R-CNN to balance and constrain the positioning loss from the loss function.

4 Experiment

4.1 SKU-110K sub-dataset

We evaluated the method in this article on the SKU-110K sub-dataset. The dataset contains 3,828 training and test pictures, with a total of 559,347 ground truth, and on average, 146 objects to be detected in each picture. In addition, some experimental results of other models on this dataset are provided. The model mainly uses the MS COCO evaluation criteria to evaluate the mean average precision (mAP). All experiments are implemented based on the PaddleDetection deep learning framework and carried out on the Tesla V100 GPU. Only one GPU is used during training and testing. The hyperparameters of the training process are set as Batch size: the size is 2; the initial learning rate is 0.01, and the learning rate of 500 iterations has warmed up the initial learning rate of 0.001. A total of six epochs are trained, and the learning rate is reduced to 0.1 times the current learning rate at the 4th and 5.5th epochs, respectively. The optimizer uses SGD + momentum. The training set uses AutoAugment [35] to enhance the picture. A multi-scale training method is used further during training to optimize the detection of multi-scale problems in pictures.

Figure 6 shows the loss of classification and positioning in the second and third detection heads in the cascade structure. After the first two detection heads, the loss of the third detection head is significantly lower than that of the first two detection heads. Still, it can also be seen that the positioning loss is higher than the classification loss. Therefore, the core task of dense detection is the regression of bounding boxes. After using the cascade structure, the loss of bounding box regression is significantly reduced. It can be seen from Table 1 (5) that the structure of cascade combined with FPN performs exceptionally on dense detection tasks, and its baseline accuracy is higher than other models in Table 2.

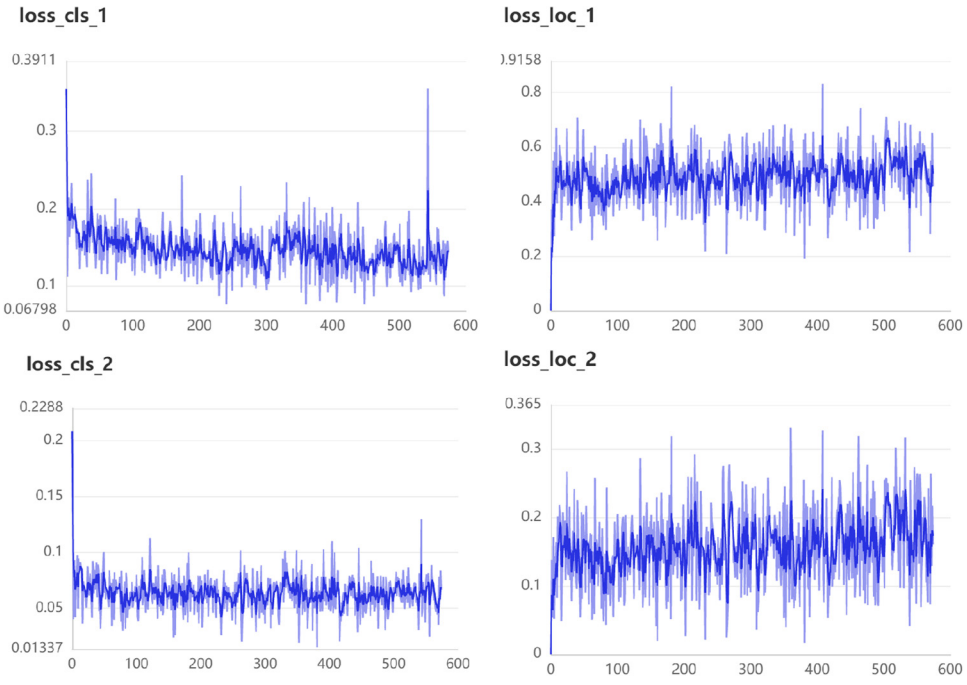

Figure 6: The first and second rows are, respectively, the second, the classification loss and location loss of the three detection heads in the training process. It can be clearly compared in the figure that the improvement of the quality of the Cascade structure to the regional recommendations is the increase in detection accuracy.

Table 1: An ablation experiment to study the influence of different components on the model detection effect: through the above implementation, it can be seen that the imbalance of positive and negative samples has a certain impact on the experimental results

Index	Method	mAP	FPS
1	Cascade Rcnn + ResNet101 + FPN + CIoU	0.411	4.252
2	Cascade Rcnn + ResNet101 + FPN + balanced L1 loss	0.413	4.203
3	Cascade Rcnn + ResNet101 + FPN	0.407	4.233
4	Cascade Rcnn + ResNet101 + BFP + balanced L1 loss	0.412	3.785
5	Cascade Rcnn + ResNet101 + FPN(baseline)	0.400	4.233

Table 2: The accuracy comparison between our model and other detection models on the SKU-110K subset

Index	Method	mAP	FPS
1	Faster Rcnn + ResNet50 + FPN(baseline)	0.332	11.273
2	Y olov3 + DarkNet53	0.302	45.571
3	RetinaNet + ResNet101 + FPN	0.373	—
4	Faster Rcnn + HrNet [36]	0.347	—
5	Ppyolo [37]	0.362	72.9
6	Faster Rcnn + Se154 [38]+FPN	0.384	5.209
7	Cascade Rcnn + ResNet101 + FPN + balanced L1 loss	**0.413**	4.233

The results show that our effect is far superior to other detection models.

Corresponding experiments are done to evaluate the influence of IoU loss and regression loss on dense detection tasks, as shown in Table 1. In Table 1 (3), it is found that the parameters such as the number of dense sampling preselection boxes are reset to the model effect. The two-stage model is used for dense detection. The number of bounding boxes is adjusted accordingly to ensure enough bounding boxes during the training process. Predicting the ground truth can improve the training effect of the model. Table 1 (1) and (2) are experiments on IoU loss and regression loss based on model (3). The results were improved by 1 and 1.5% based on model (3). It can be proved that it effectively improves the accuracy of dense object detection in terms of IoU loss and regression loss.

The above experiments prove that the high-quality suggestion area selected by Cascade R-CNN can significantly improve the accuracy of dense object detection. At the same time, it can be verified from the IoU Loss and localization loss experiments that efficient IoU loss and equalization regression loss are beneficial to the result of dense detection. It can be concluded that in dense object detection, high-quality suggestion areas and handling of positioning tasks in detection tasks are the keys to improving detection accuracy.

We compare our model with other models, as shown in Table 2. We can trim the Yolo series one-stage detection model to be lower than the two-stage detection model in dense scenes. Such as (1), (2) through (3), (5) at the same time, we can see that a better feature extraction network is of great help to the increase of object detection, replacing Se154 as the backbone, relative to the possible 15% improvement. The state-of-the-art one-stage detector ppyolo still has a difference in the accuracy of dense detection compared to the two-stage detector.

4.2 MS-COCO

In addition, to verify the usability of our method in general scenarios, experiments were performed on MS-COCO2017.

Table 3: Experimental results on MS-COCO2017

Index	Method	mAP	FPS
1	Faster Rcnn + ResNet50 – FPN	0.372	22.273
2	Cascade Rcnn + ResNet50 + FPN	0.408	17.507
3	Faster + ResNet101 – FPN	0.387	17.297
4	Yolo v3 + DarkNet53	0.334	50.571
5	RetinaNet + ResNet101 – FPN	0.373	—
6	PP – YOLO + ResNet50	0.453	72.9
7	Fcos + ResNet50 – FPN	0.39.6	—
8	Cascade Rcnn + ResNet101 + FPN + balanced L1 loss	0.457	4.233

The most advanced one-stage model can also show strong performance in sparse detection scenarios.

All models are trained and tested in the MS-COCO2017 dataset. Paddle provides a skeleton network pre-trained model based on ImageNet. All pre-trained models are trained on the standard Imagenet-1k dataset. It can be seen from Table 3 that the most advanced pp-yolo performed very well on MS-COCO2017, but the detection targets in MS-coco2017 are relatively sparse, so the one-stage detection model can also achieve good results. Compared with the results in Table 2, it is evident that the two-stage detection model performs better in dense scenes.

5 Discussion

Our work proved through experiments that the two-stage detector is significantly better than the one-stage detector in dense scenes. The apparent difference between the two-stage and one-stage detectors lies in the definition of positive and negative samples. The two-stage detector obtains high-quality recommended areas through one-stage screening, and the division of positive and negative samples is the key to improving the efficiency of dense scene detection. At the same time, due to the existence of many detection frames, we found through our experiments that for the detection in this scene, balancing the classification loss and regression loss can slightly improve the detection efficiency. Our method is not suitable for scenarios with high real-time requirements. Due to the limitations of the cascade structure, the overall detection speed is slow. The definition of positive and negative samples during training has always been a core issue in target detection. Our follow-up work will consider improving the detection speed of dense detection from the definition of positive and negative samples in the one-stage detector.

6 Conclusion

To improve the performer of object detection in dense scenes, this article researches from two aspects: region proposal and location loss. The Cascade structure is used to improve the quality of regional proposals. CIoU loss and balanced L1 loss are, respectively, tested on two aspects of location loss. At the same time, the FPN multi-scale detection algorithm is used to improve the detection efficiency of small objects. In dense object detection, high-quality suggested regions are the key to improving detection accuracy. Second, the regulation of IoU loss has a particular effect on the final result. The experimental results show that the detection scheme in this article can achieve high detection accuracy, with an mAP of 0.413, which exceeds the accuracy requirements of industrial production. In the later work of detection in dense scenes, research can be carried out to improve the quality of the suggested region.

Conflict of interest: Authors state no conflict of interest.

Data availability statement: Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, 2017, pp 84–90.

[2] C. Szegedy, W. Liu, Y. Angqing Jia, P. Sermanet, S. Reed, D. Anguelov, et al., *Going deeper with convolutions*, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. pp. 1–9.

[3] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2014, arXiv:http://arXiv.org/abs/arXiv:1409.1556.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[5] J. Deng, W. Dong, R. Socher, L-J. Li, K. Li, and L. Fei-Fei. "Imagenet: A large-scale hierarchical image database," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, p. 1.

[6] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, 2017.

[7] 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*, 2014, pp. 580–587.

[8] R. Girshick, "Fast r-cnn," In *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[9] S. Ren, K. He, R. Girshick, and J. Sun "Faster r-cnn: Towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process Syst.*, 2015, vol. 28, pp. 91–99.

[10] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[11] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao *YOLOv4: Optimal Speed and Accuracy of Object Detection*, 2020, arXiv: http://arXiv.org/abs/arXiv:2004.10934.

[12] J. Redmon, S. K. Divvala, R. B. Girshick, and Ali Farhadi, "You only look once: unified, real-time object detection,"

In *Proceedings of Conference on Computer Vision Pattern Recognition*, 2016.
[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, et al., "SSD: Single shot multibox detector," In *European Conference on Computer Vision*, 2016.
[14] A. Farhadi, and J. Redmon. "Yolov3: An incremental improvement," In *Computer Vision and Pattern Recognition*, Berlin/Heidelberg, Germany: Springer, 2018, p. 1804.02767.
[15] E. Goldman, R. Herzig, A. Eisenschtat, J. Goldberger, and T. Hassner, "Precise detection in densely packed scenes," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5227–5236.
[16] Z. Q. Zhao, P. Zheng, S. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Networks Learning Syst.*, vol. 30, no. 11, pp. 3212–3232, 2019.
[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR'05*), vol. 1, IEEE, 2005, pp. 886–893.
[18] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, 2010, pp. 2241–2248.
[19] 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*, 2014, pp. 580–587.
[20] CL Zitnick and P. Dollár, "Edge boxes: locating object proposals from edges," In *European Conference on Computer Vision*, Cham: Springer, 2014, pp. 391–405.
[21] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vision*, vol. 104, no. 2, pp. 154–171, 2013.
[22] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 34, no. 11, pp. 2189–2202, 2012.
[23] P. Purkait, C. Zhao, and C. Zach, "SPP-Net: Deep absolute pose regression with synthetic views," 2017, arXiv:http://arXiv.org/abs/arXiv:1712.03452.
[24] E. Goldman, R. Herzig, A. Eisenschtat, J. Goldberger, and T. Hassner, "Precise detection in densely packed scenes," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5227–5236.
[25] 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*, 2017, pp. 2117–2125.
[26] H Rezatofighi, N Tsoi, JY Gwak, A. Sadeghian, I. Reid, S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 658–666.
[27] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," In *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9627–9636.
[28] M. Everingham, L. Van Gool, C. K. I. Williams, and J. Winn, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vision*, vol. 88, no. 2, 303–338, 2010.
[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al., "Microsoft coco: Common objects in context," In *ECCV*, 2014. p. 1.
[30] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2009, pp. 248–255.
[31] C. Arteta, V. Lempitsky, and A. Zisserman, "Counting in the wild," In *European Conference on Computer Vision*, Cham: Springer, 2016, pp. 483–498.
[32] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," *European Conference on Computer Vision*. Cham: Springer, 2016, pp. 615–629.
[33] X. Chu, A. Zheng, X. Zhang, and J. Sun, "Detection in crowded scenes: one proposal multiple predictions," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12214–12223.
[34] 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. 34, no. 7, pp. 12993–13000, 2020.
[35] B. Zoph, E. D. Cubuk, G. Ghiasi, T. Y. Lin, J. Shlens, and Q. V. Le, "Learning data augmentation strategies for object detection," In *European Conference on Computer Vision*, Cham: Springer, 2020, pp. 566–583.
[36] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.
[37] X. Long, K. Deng, G. Wang, Y. Zhang, Q. Dang, and Y. Gao, "PP-YOLO: An effective and efficient implementation of object detector," 2020, arXiv:http://arXiv.org/abs/arXiv:2007.12099.
[38] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
[39] Z. Huang, J. Wang, X. Fu, T. Yu, Y. Guo, and R. Wang, "DC-SPP-YOLO: Dense connection and spatial pyramid pooling based YOLO for object detection," *Inform. Sci.*, vol. 522, pp. 241–258, 2020.